Ming-Hui Chen
Dipak K. Dey · Peter Müller
Dongchu Sun · Keying Ye
Editors

Frontiers of Statistical Decision Making and Bayesian Analysis

In Honor of James O. Berger

Springer

Frontiers of Statistical Decision Making and Bayesian Analysis

Ming-Hui Chen · Dipak K. Dey · Peter Müller ·
Dongchu Sun · Keying Ye
Editors

Frontiers of Statistical Decision Making and Bayesian Analysis

In Honor of James O. Berger

Springer

Editors
Prof. Ming-Hui Chen
Department of Statistics
University of Connecticut
215 Glenbrook Road, U-4120
Storrs, CT 06269
USA
mhchen@stat.uconn.edu

Prof. Dipak K. Dey
Department of Statistics
University of Connecticut
215 Glenbrook Road, U-4120
Storrs, CT 06269
USA
dipak.dey@uconn.edu

Prof. Peter Müller
Department of Biostatistics
The University of Texas
M. D. Anderson Cancer Center
1515 Holcombe Boulevard, Box 447
Houston, TX 77030
USA
pmueller@mdanderson.org

Prof. Dongchu Sun
Department of Statistics
University of Missouri-Columbia
146 Middlebush Hall
Columbia, MO 65211
USA
sund@missouri.edu

Prof. Keying Ye
Department of Management Science
and Statistics, College of Business
University of Texas at San Antonio
San Antonio, TX 78249
USA
Keying.Ye@utsa.edu

ISBN 978-1-4419-6943-9 e-ISBN 978-1-4419-6944-6
DOI 10.1007/978-1-4419-6944-6
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2010931142

Dedicated to Jim and Ann Berger for their encouragement, support and love all through our academic life and beyond

Preface

This book surveys the current research frontiers in Bayesian statistics. Over the last decades we have seen an unprecedented explosion of exciting Bayesian work. The explosion is in sheer number of researchers and publications, as well as in the diversity of application areas and research directions. Over the past few years several excellent new introductory texts on Bayesian inference have appeared, as well as specialized monographs that focus on specific aspects of Bayesian inference, including dynamic models, multilevel data, non-parametric Bayes, bioinformatics and many others. Thus this is a natural time for a book that can pull all these diverse areas together and present a snapshot of current research frontiers in Bayesian inference and decision making. The intention of this volume is to provide such a snapshot.

Many of the research frontiers that are discussed in this volume have a close connection to the life and career of Jim Berger, thus the subtitle of this book. Not coincidentally, many authors are former students, collaborators, and friends of Jim Berger. Like few others Jim has been instrumental in shaping the recent expansion of Bayesian research. Jim's early work in admissibility, shrinkage estimation, and on conditioning established some of the foundations that current work builds on. Working on admissibility naturally leads to later work on robustness, an issue that is again becoming a paramount concern as we move to increasingly more complex models. Work on robustness and admissibility also reflects Jim's lifelong interest in the Bayesian-frequentist interface. This interest led him to study Bayesian p-values and objective Bayes methodology, two other important research directions in Jim's work. Reflecting a general trend in Bayesian statistics research, Jim also became increasingly involved with substantial applications and practical aspects of Bayesian inference, starting with work on fuel efficiency in the 90's, and continuing with work on astronomy, computer experiments, and more.

Besides his own research, Jim's close association with Bayesian research and statistics research in general arises from his long record of service in the profession, including serving as president of IMS, ISBA and ASA/SBSS, as co-editor of the *Annals of Statistics* and as organizer of countless conferences and as a member of the US National Academy of Science as well as the Spanish Real Academia de Ciencias. Over the last eight years Jim has shaped statistics research also by his

leadership as the founding director for the new Statistical and Applied Mathematical Sciences Institute (SAMSI). Many researchers who participated in the exciting programs at SAMSI over the last 10 years greatly appreciate Jim's hard work to create and maintain the unique research environment at SAMSI. Most importantly to us and many authors of the chapters in this volume, Jim has substantially contributed to the development of statistics research as an outstanding advisor, mentor and colleague. Jim has been advisor to over 30 Ph.D. students. We truly appreciate the privilege of Jim's guidance and help, which often went way beyond our years as graduate students. Here we need to also acknowledge Ann Berger as the super lurking variable behind Jim's extraordinary career and the success of many of his students. Thanks!

The chapters in this book were chosen to provide a broad survey of current research frontiers in Bayesian analysis, ranging from foundations, to methodology issues, to computational themes and applications. It is an amazing feature of Jim's life and research that the chapters in this book happen to simultaneously also almost be an inventory of his many research interests.

March 2010

Ming-Hui Chen, Dipak K. Dey, Peter Müller, Dongchu Sun, and Keying Ye

Contents

List of Contributors

Edoardo M. Airoldi
Department of Statistics
Harvard University
Science Center
1 Oxford Street
Cambridge, MA 02138-2901, USA
e-mail: airoldi@fas.harvard.edu

James H. Albert
Department of Mathematics
and Statistics
Bowling Green State University
Bowling Green, Ohio 43403, USA
e-mail: albert@math.bgsu.edu
and e-mail: albertcb1@gmail.com

Sudipto Banerjee
Division of Biostatistics
School of Public Health
University of Minnesota
420 Delaware Street SE
A460 Mayo Bldg. MMC 303
Minneapolis, MN 55455, USA
e-mail: baner009@umn.edu

Suhrid Balakrishnan
AT&T Labs Research
180 Park Avenue
Florham Park, NJ 07932, USA
e-mail: suhrid@research.att.com

Maria M. Barbieri
Dipartimento di Economia
Università Roma Tre
via Silvio D'Amico, 77
00145 Roma, ITALY
e-mail: barbieri@uniroma3.it

Susie M.J. Bayarri
Department of Statistics and O.R.
University of Valencia
Dr. Moliner 50
46100 Burjassot, Valencia, Spain
e-mail: Susie.Bayarri@uv.es

José M. Bernardo
Department of Statistics and O.R.
University of Valencia
Dr. Moliner 50
46100 Burjassot, Valencia, Spain
e-mail: jose.m.bernardo@uv.es

K. Sham Bhat
Department of Statistics
Pennsylvania State University
326 Thomas Building
University Park, PA 16802, USA
e-mail: kgb130@psu.edu

Erik W. Bloomquist
Mathematical Biosciences Institute
The Ohio State University
381 Jennings Hall

1735 Neil Avenue
Columbus, OH 43210, USA
e-mail: bloomquist.2@osu.edu

Arijit Chakrabarti
Bayesian and Interdisciplinary
Research Unit
Indian Statistical Institute
Kolkata - 700108, West Bengal, India
e-mail: arc@isical.ac.in

Ming-Hui Chen
Department of Statistics
University of Connecticut
215 Glenbrook Road, U-4120
Storrs, CT 06269, USA
e-mail: mhchen@stat.uconn.edu

Hyunsoon Cho
National Cancer Institute
6116 Executive Boulevard Suite 504
Bethesda, MD 20892, USA
e-mail: choh3@mail.nih.gov

Bertrand Clarke
Department of Medicine
Center for Computational Sciences
and Department of Epidemiology
and Public Health
University of Miami
1120 NW 14th Street
CRB 611 (C-213)
Miami, FL, 33136, USA
e-mail: bclarke2@med.miami.edu

Luyan Dai
Boehringer Ingelheim Pharm Inc.
CC G2225A
900 Ridgebury Road
Ridgefield, CT 06877, USA
e-mail: luyan.dai@boehringer-
ingelheim.com

Gauri S. Datta
Department of Statistics
University of Georgia
Athens, GA 30602-1952, USA
e-mail: gaurisdatta@gmail.com

Victor De Oliveira
Department of Management Science
and Statistics
The University of Texas at San Antonio
One UTSA Circle
San Antonio, TX 78249, USA
e-mail: Victor.DeOliveira@utsa.edu

Dipak K. Dey
Department of Statistics
University of Connecticut
215 Glenbrook Road, U-4120
Storrs, CT 06269, USA
e-mail: dipak.dey@uconn.edu

Soma Dhavala
Statistics Department
Texas A&M University
College Station, TX 77843-3143 USA
e-mail: soma@stat.tamu.edu

Stephen E. Fienberg
132G Baker Hall
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213-3890, USA
e-mail: fienberg@stat.cmu.edu

Duncan K.H. Fong
Department of Marketing
Smeal College of Business
456 Business Building
Penn State University
University Park, PA 16802, USA
e-mail: i2v@psu.edu

Alan E. Gelfand
Department of Statistical Science
Duke University
Durham, NC 27708-0251, USA
e-mail: alan@stat.duke.edu

Andrew Gelman
Department of Statistics
1255 Amsterdam Ave, room 1016
Department of Political Science
International Affairs Bldg, room 731
Columbia University

New York, N.Y. 10027, USA
e-mail: gelman@stat.columbia.edu

Edward I. George
Department of Statistics
The Wharton School
University of Pennsylvania
Philadelphia, PA 19104-6302, USA
e-mail: edgeorge@wharton.upenn.edu

Jayanta K. Ghosh
Department of Statistics
Purdue University
250 N. University Street
West Lafayette, IN 47907-2066, USA
e-mail: ghosh@stat.purdue.edu

Malay Ghosh
Department of Statistics
University of Florida
102 Griffin-Floyd Hall
P.O. Box 118545
Gainesville, Florida 32611, USA
e-mail: ghoshm@stat.ufl.edu

Marlos Goes
ASA - South America
Rua Fidalga, 711
São Paulo, Brazil 05432-070
e-mail: mpg14@psu.edu

Murali Haran
Department of Statistics
Pennsylvania State University
326 Thomas Building
University Park, PA 16802, USA
e-mail: mharan@stat.psu.edu

Matthew J. Heaton
Department of Statistical Science
Duke University
Durham, NC 27708-0251, USA
e-mail: matt@stat.duke.edu

Satadru Hore
Federal Reserve Bank of Boston
600 Atlantic Avenue
Boston, MA 02210, USA
e-mail: satadru.hore@gmail.com

Joseph G. Ibrahim
Department of Biostatistics
University of North Carolina
McGavran-Greenberg Hall
Chapel Hill, NC 27599, USA
e-mail: ibrahim@bios.unc.edu

Eric Jacquier
3000 chemin de la Côte-Sainte-Catherine
Finance Department
CIRANO and HEC. Montréal
Montréal, QC H3T 2A7, Canada
e-mail: eric.jacquier@hec.ca

Michael Johannes
Graduate School of Business
Columbia University
3022 Broadway, Uris Hall 424
New York, NY 10027, USA
e-mail: mj335@columbia.edu

Valen E. Johnson
Department of Biostatistics
The University of Texas
M. D. Anderson Cancer Center
1515 Holcombe Boulevard, Box 447
Houston, TX 77030, USA
e-mail: vejohnson@mdanderson.org

Michael I. Jordan
427 Evans Hall
Department of Statistics
and Department of EECS
University of California
Berkeley, CA 94720-3860, USA
e-mail: jordan@eecs.berkeley.edu

Cyrille J. Joutard
Departement de Mathematiques
Universite Montpellier 2
34095 Montpellier Cedex, France
e-mail: cjoutard@math.univ-montp2.fr

Sungduk Kim
Division of Epidemiology,
Statistics and Prevention Research
Eunice Kennedy Shriver National

Institute of Child Health
and Human Development
NIH, Rockville, MD 20852, USA
e-mail: kims2@mail.nih.gov

Lynn Kuo
Department of Statistics
University of Connecticut
215 Glenbrook Road, U-4120
Storrs, CT 06269-4120, USA
e-mail: lynn@stat.uconn.edu

Purushottam W. Laud
Division of Biostatistics
Medical College of Wisconsin
8701 Watertown Plank Road
Milwaukee, WI 53226, USA
e-mail: laud@mcw.edu

Jaeyong Lee
Department of Statistics
Seoul National University
Seoul 151-742, Korea
e-mail: jylc@snu.ac.kr

Hongzhe Li
Department of Biostatistics
and Epidemiology
School of Medicine
University of Pennsylvania
920 Blockley Hall
423 Guardian Drive
Philadelphia, PA 19104-6021, USA
e-mail: hongzhe@mail.med.upenn.edu

John C. Liechty
Department of Marketing
Smeal College of Business
and Department of Statistics
Pennsylvania State University
University Park, PA 16803, USA
e-mail: jcl12@psu.edu

Merrill W. Liechty
Department of Decision Sciences
LeBow College of Business
224 Academic Building
Drexel University
Philadelphia, PA 19104, USA
e-mail: merrill@drexel.edu

Ilya Lipkovich
Eli Lilly and Company
Indianapolis, IN 46285, USA
e-mail: Lipkovich_ilya_a@lilly.com

Brunero Liseo
Dipartimento di studi geoeconomici
Sapienza Università di Roma
Viale del Castro Laurenziano, 9
00161 Rome, Italy
e-mail: brunero.liseo@uniroma1.it
and e-mail: brunero.liseo@gmail.com

Hedibert Lopes
Booth School of Business
University of Chicago
5807 South Woodlawn Avenue
Chicago, IL 60637, USA
e-mail: hlopes@ChicagoBooth.edu

Tanzy M. Love
Department of Biostatistics and
Computational Biology
University of Rochester Medical Center
601 Elmwood Avenue, Box 630
Rochester, NY 14642, USA
e-mail:
Tanzy_Love@urmc.rochester.edu

David Madigan
Department of Statistics
Columbia University
1255 Amsterdam Ave.
New York, NY 10027, USA
e-mail: madigan@stat.columbia.edu

Bani K. Mallick
Statistics Department
Texas A&M University
College Station, TX 77843-3143 USA
e-mail: bmallick@stat.tamu.edu

Jean-Michel Marin
Institut de Mathématiques et
Modélisation de Montpellier
Université Montpellier 2,

Case Courrier 51
34095 Montpellier cedex 5, France
e-mail:
Jean-Michel.Marin@univ-montp2.fr

Robert E. McCulloch
McCombs School of Business
University of Texas at Austin
1 University Station, B6500
Austin, TX 78712-0212, USA
e-mail: robert.mcculloch1@gmail.com

Xiao-Li Meng
Department of Statistics
Harvard University
Science Center
1 Oxford Street
Cambridge, MA 02138-2901, USA
e-mail: meng@stat.harvard.edu

Stefano Monni
Department of Public Health
Weill Cornell Medical College
402 East 67th Street
New York, NY 10065-6304, USA
e-mail: stm2013@med.cornell.edu

Bhramar Mukherjee
Department of Biostatistics
University of Michigan
Ann Arbor, MI 48109, USA
e-mail: bhramar@umich.edu

Peter Müller
Department of Biostatistics
The University of Texas
M. D. Anderson Cancer Center
1515 Holcombe Boulevard, Box 447
Houston, TX 77030, USA
e-mail: pmueller@mdanderson.org

Luis Raúl Pericchi Guerra
Department of Mathematics
University of Puerto Rico at Rio
Piedras Campus, P.O. Box 23355
San Juan 00931-3355, Puerto Rico
e-mail: lrpericchi@uprrp.edu and
e-mail: luarpr@gmail.com

Nicholas G. Polson
Booth School of Business
University of Chicago
5807 South Woodlawn Avenue
Chicago, IL 60637, USA
e-mail: ngp@ChicagoBooth.edu

Adrian Raftery
Department of Statistics
University of Washington
Box 354322
Seattle, WA 98195-4322, USA
e-mail: raftery@u.washington.edu

J.N.K. Rao
School of Mathematics and Statistics
Carleton University
Ottawa ON K1S 5B6, Canada
e-mail: jrao@math.carleton.ca

Shubhankar Ray
Biometrics Research, RY 33-300
Merck & Company, P.O. Box 2000
Rahway, NJ 07065, USA
e-mail: shubhankar_ray@merck.com

Christian P. Robert
CEREMADE
Université Paris Dauphine
75775 Paris, and
CREST, INSEE, Paris, France
e-mail: xian@ceremade.dauphine.fr

James G. Scott
Department of Information, Risk,
and Operations Management
University of Texas at Austin
1 University Station, B6500
Austin, TX 78712, USA
e-mail:
james.scott@mccombs.utexas.edu

Haige Shen
Novartis Oncology, Biometrics
180 Park Ave, 104-2K11, Florham Park,
NJ 07932, USA
e-mail: haigeshen@yahoo.com

Siva Sivaganesan
Department of Mathematical Sciences
University of Cincinnati
811-C Old Chemistry
PO Box 210025
Cincinnati, OH 45221-0025, USA
e-mail: siva.sivaganesan@uc.edu

Eric P. Smith
Department of Statistics
Virginia Tech
Blacksburg, VA 24061, USA
e-mail: epsmith@vt.edu

Russell J. Steele
Department of Mathematics and Statistics
McGill University
805 Sherbrooke Ouest
Montreal, QC, Canada H3A 2K6
e-mail: steele@math.mcgill.ca

William E. Strawderman
Department of Statistics
561 Hill Center, Busch Campus
Rutgers University
Piscataway, NJ 08854-8019, USA
e-mail: straw@stat.rutgers.edu

Marc A. Suchard
Departments of Biomathematics, Biostatistics and Human Genetics
David Geffen School of Medicine at UCLA, 6558 Gonda Building
695 Charles E. Young Drive, South
Los Angeles, CA 90095-1766, USA
e-mail: msuchard@ucla.edu

Dongchu Sun
Department of Statistics
University of Missouri-Columbia
146 Middlebush Hall
Columbia, MO 65211, USA
e-mail: sund@missouri.edu

Andrea Tancredi
Dipartimento di studi geoeconomici
Sapienza Università di Roma
Viale del Castro Laurenziano, 9
00161 Rome, Italy
e-mail: andrea.tancredi@uniroma1.it

Niansheng Tang
Department of Statistics
Yunnan University
Kunming 650091, Yunnan
P. R. of China
e-mail: nstang@ynu.edu.cn

Peter F. Thall
Department of Biostatistics
The University of Texas
M. D. Anderson Cancer Center
1515 Holcombe Boulevard, Box 447
Houston, TX 77030, USA
e-mail: rex@mdanderson.org

Surya T. Tokdar
Department of Statistical Science
Duke University
Durham, NC 27708-0251, USA
e-mail: st118@stat.duke.edu

Vera Tomazella
Departamento de Estatística
Universidade Federal de São Carlos
Rodovia Washington Luiz, Km 235
Monjolinho, 13565-905 - São Carlos
SP - Brasil - Caixa-Postal: 676
e-mail: vera@power.ufscar.br

Xia Wang
National Institute of Statistical Sciences
19 T.W. Alexander Drive
P. O. Box 14006
Research Triangle Park, NC 27709, USA
e-mail: xiawang.z@gmail.com

Yan Wang
Center for Drug Evaluation and Research
U.S. Food and Drug Administration
10903 New Hampshire Avenue
Silver Spring, MD 20993, USA
e-mail: yanwang20@yahoo.com

J. Kyle Wathen
Department of Biostatistics
The University of Texas
M. D. Anderson Cancer Center
1515 Holcombe Boulevard, Box 447
Houston, TX 77030, USA
e-mail: jkwathen@mdanderson.org

Robert E. Weiss
Department of Biostatistics
UCLA School of Public Health
Los Angeles, CA 90095, USA
e-mail: robweiss@ucla.edu

Mike West
Department of Statistical Science
Duke University
Durham, NC 27708-0251, USA
e-mail: mw@stat.duke.edu

Wangang Xie
Abbott Lab
100 Abbott Park, R436/AP9A-2
Abbott Park, IL 60064, USA
e-mail: Wangang.Xie@abbott.com

Xinyi Xu
Department of Statistics
The Ohio State University
1958 Neil Avenue
Columbus, OH 43210-1247, USA
e-mail: xinyi@stat.osu.edu

Keying Ye
Department of Management Science
and Statistics, College of Business
University of Texas at San Antonio
San Antonio, TX 78249, USA
e-mail: Keying.Ye@utsa.edu

Ao Yuan
Statistical Genetics and
Bioinformatics Unit
National Human Genome Center
Howard University
2216 Sixth Street, N.W., Suite 206
Washington, DC 20059, USA
e-mail: ayuan@howard.edu

Hongtu Zhu
Department of Biostatistics
University of North Carolina
McGavran-Greenberg Hall
Chapel Hill, NC 27599, USA
e-mail: hzhu@bios.unc.edu

Chapter 1
Introduction

In the years since the 1985 publication of *Statistical Decision Theory and Bayesian Analysis* by James Berger, there has been an enormous increase in the use of Bayesian analysis and decision theory in statistics and science. The rapid expansion in the use of Bayesian methods is due in part to substantial advances in computational and modeling techniques, and Bayesian methods are now central in many branches of science. The aim of this book is to review current research frontiers in Bayesian analysis and decision theory. It is impossible to provide an exhaustive discussion of all current research in Bayesian statistics, so the book instead summarizes current research frontiers by providing representative examples of research challenges chosen from a wide variety of areas.

In this first chapter, Section 1.1 is a short biography of James Berger. One of the many important aspects of James Berger's contributions to statistics and applied mathematics is his leadership of the Statistical and Applied Mathematical Science Institute (SAMSI). The exciting research themes that defined SAMSI programs over the last eight years are reviewed in Section 1.2. The last section of this chapter contains an overview of the book.

1.1 Biography of James O. Berger

James Berger was born on April 6, 1950 in Minneapolis, Minnesota, to Orvis and Thelma Berger. He received his PhD in Mathematics from Cornell University in 1974, five years after he graduated from high school. He then joined the department of statistics at Purdue University, receiving tenure two years later, and was promoted to full professor in 1980. In 1985, he became Richard M. Brumfield Distinguish Professor of Statistics at Purdue University. Since 1997, he has been Arts and Science Distinguished Professor of Statistics at Duke University. Currently he is also the director of the Statistical and Applied Mathematical Science Institute (SAMSI), located in Research Triangle Park, North Carolina, USA.

M.-H. Chen et al. (eds.), *Frontiers of Statistical Decision Making and Bayesian Analysis*, DOI 10.1007/978-1-4419-6944-6_1,

Berger was president of the Institute of Mathematical Statistics from 1995 to 1996, chair of the Section on Bayesian Statistical Science of the American Statistical Association in 1995, and president of the International Society for Bayesian Analysis during 2004. He has been involved with numerous editorial activities, including co-editorship of the *Annals of Statistics* (the landmark journal in the statistical society) during the period 1998–2000, and has organized or participated in the organization of over 39 conferences, including the Purdue Symposiums on Statistical Decision Theory and Related Topics and the Valencia International Meetings on Bayesian Statistics. He has also served on numerous statistical administrative and program committees, including the Committee on Applied and Theoretical Statistics of the National Research Council. Berger has also been on various university and NSF site visit teams and panels, and NSF advisory committees.

Among the awards and honors Berger has received are Guggenheim and Sloan Fellowships, the 1985 COPSS President's Award (an award given to a researcher no more than 40, in recognition of outstanding contributions to the statistics profession), the Sigma Xi Research Award at Purdue University for contribution of the year to science in 1993, the Fisher Lectureship in 2001, election as foreign member of the Spanish Real Academia de Ciencias in 2002, election to the U.S. National Academy of Sciences in 2003, award of an honorary Doctor of Science degree from Purdue University in 2004, and the Wald Lectureship in 2007.

Berger's research has primarily been in Bayesian statistics, foundations of statistics, statistical decision theory, simulation, model selection, and various interdisciplinary areas of science and industry, especially astronomy and the interface between computer modeling and statistics. He has supervised over 30 PhD dissertations, published over 170 articles, and has written or edited 14 books or special volumes.

Berger married Ann Louise Duer (whom he first met when they were in the seventh grade together) in 1970, and they have two children, Jill Berger who is married to Sascha Hallstein and works as optical scientist in Silicon Valley, and Julie Gish who is married to Ryan Gish and works as a consultant in Chicago. The Berger's have three grandchildren, Charles and Alexander Gish and Sophia Hallstein.

1.2 The Frontiers of Research at SAMSI

James Berger was the founding Director of the Statistical and Applied Mathematical Sciences Institute (SAMSI, http://www.samsi.info) in 2002. Created as part of the Mathematical Sciences Institutes program at the National Science Foundation, SAMSI's vision is to forge a new synthesis of the statistical sciences and the applied mathematical sciences with disciplinary science to confront the very hardest and most important data- and model-driven scientific challenges. Each year, more than 100 researchers have participated in SAMSI research working groups through extended visits, and SAMSI workshops have drawn over 1000 national and international participants annually.

In over eight years of stewardship of SAMSI, Berger has overseen the development and implementation of 24 major scientific programs (see http://www.samsi.info/programs/index.shtml), with others still under development. These programs provide a snapshot of the views of Berger and other leaders in the statistical community concerning which topics are of central importance to statistics and its interaction with other disciplines. We can thus summarize this view of the *frontiers of statistics* by briefly reviewing the past, current, and future programs of SAMSI. Details of each of the programs and references can be found at the SAMSI website.

1.2.1 Research Topics from Past SAMSI Programs

1.2.1.1 Stochastic Computation, 2002–2003

Stochastic computation explored the use of methods of stochastic computation in the following key areas of statistical modeling:

Stochastic computation in problems of model and variable selection: (i) computing hard likelihoods and Bayes' factors for model comparison and selection; and (ii) synthesis of existing stochastic computational approaches to model uncertainty and selection.

Stochastic computation in inference and imputation in contingency tables analysis: (i) perfect simulation approaches to "missing data" problems; (ii) synthesis of existing Markov chain simulation methods — "local" and "global" move approaches; (iii) experiments with approaches in large, sparse tables; and (iv) applications in genetics and other areas.

Stochastic computation in analysis of large-scale graphical models: (i) Monte Carlo and related stochastic search methods for sparse, large-scale graphical models; (ii) model definition and specification for sparse models; and (iii) applications in genomics (large-scale gene expression studies).

Stochastic computation in financial mathematics, especially in options pricing models: (i) Monte Carlo methods in specific options pricing models; and (ii) sequential Monte Carlo methods and particle filtering in stochastic volatility models.

The core research focuses included studies of the performance characteristics of current stochastic computational methods, refinements and extensions of existing approaches, and development of innovative new approaches. In two of the areas, in particular, there was an explicit focus on the development of interactions between statisticians (coming from methodological and applied perspectives) and theoretical probabilists and mathematicians working on related problems. One key example was the component on imputation in contingency tables, with an interest in connecting Markov chain methods from statistics with perfect sampling from probability and algebraic approaches from mathematics. Another example was in modeling

in financial pricing studies where statistical and mathematical "schools" have had limited interactions.

1.2.1.2 Inverse Problem Methodology in Complex Stochastic Models, 2002–2003

In a diverse range of fields, including engineering, biology, physical sciences, material sciences, and medicine, there is heightening interest in the use of complex dynamical systems to characterize underlying features of physical and biological processes. For example, a critical problem in the study of HIV disease is elucidation of mechanisms governing the evolution of patient resistance to antiretroviral therapy, and there is growing consensus that progress may be made by representing the interaction between virus and host immune system by nonlinear dynamical systems whose parameters describe these mechanisms. Similarly, recovery of molecular information for polymers via light beam interrogation may be characterized by a dynamical system approach. Risk assessment is increasingly focused on deriving insights from physiologically based pharmacokinetic dynamical systems models describing underlying processes following exposure to potential carcinogens.

In all of these applications, a main objective is to use complex systems to uncover such mechanisms based on experimental findings; that is, in applied mathematical parlance, to solve the relevant inverse problem based on observations of the system (data), or, in statistical terminology, to make inference on underlying parameters and model components that characterize the mechanisms from the data. In many settings, there is a realization that dynamical systems should incorporate stochastic components to take account of heterogeneity in underlying mechanisms, e.g., inter-cell variation in viral production delays in within-host HIV dynamics. In addition, heterogeneity may arise from data structure in which observations are collected from several individuals or samples with the broader focus on understanding not just individual-level dynamics but variation in mechanistic behavior across the population. In both cases, it is natural to treat unknown, unobservable system parameters as random quantities whose distribution is to be estimated. There is a large inverse problem literature for systems without such stochastic components; even here, forward solutions (i.e., solutions of the dynamics when parameters are specified) for complex systems often necessitate sophisticated techniques, so that inverse problem methodologies pose considerable challenges. Similarly, there is a vast statistical literature devoted to estimation and accounting for uncertainty in highly nonlinear models with random components, involving hierarchical specifications and complex computational issues.

With the potential overlap and emerging challenges, there has been interaction between applied mathematicians and statisticians to develop relevant inverse problem/statistical inferential methodologies. When combined, the computational and theoretical hurdles posed by both mathematical and statistical issues are substantial, and their resolution requires an integrated effort. The research from *Inverse Prob-*

lem Methodology in Complex Stochastic Models entailed facilitating the essential cooperative effort required to catalyze collaborative research in this direction.

1.2.1.3 Large-scale Computer Models for Environmental Systems, 2002–2003

Modeling of complex environmental systems is a major area of involvement of statisticians and applied mathematicians with disciplinary scientists. Yet it is also a prime example of the differing emphases of the two groups, applied mathematicians focusing on deterministic modeling of the systems and statisticians on utilizing the (often extensive) data for development and analysis of more generic stochastic models. There are two major areas of particular interest and scientific importance.

Large-scale atmospheric models. Much contemporary work in atmospheric science revolves around large-scale models such as NCAR's Community Climate System Model (which includes atmosphere, ocean, land, and ice), mesoscale models such as the Penn-State/NCAR MM5 model, and the multi-scale air quality model (Models-3) developed by the EPA. Such models are generally deterministic, but their formulation and use involve a variety of sources of uncertainty, such as unknown initial and boundary conditions; "model parameterizations" (the treatment of relevant physical phenomena varying at scales smaller than the grid size of the model, e.g., clouds); and numerical stability issues. Some climatologists have begun work on "stochastic parameterizations." Other basic issues central to these models include the need for an improved scientific understanding of unresolvable, subgrid-scale phenomena, and large-scale (chaotic) non-determinism.

Flows in porous media. Porous medium dynamics is an active and increasingly interdisciplinary research area with applications from a diverse set of fields including applied sciences (such as environmental studies, geology, hydrology, petroleum engineering, civil engineering, and soil physics), and basic sciences (such as physics, chemistry, and mathematics). However, there were few collaborative efforts that have equal footing in mathematics, disciplinary science, and statistics. The SAMSI program was built around four major themes: model formulation, parameter estimation, numerical methods, and design and optimization.

1.2.1.4 Data Mining and Machine Learning (DMML), 2003–2004

Data mining and machine learning — the discovery of patterns, information, and knowledge in what are almost always large, complex (and, often, unstructured) data sets — have seen a proliferation of techniques over the past decade. Yet, there remains incomplete understanding of fundamental statistical and computational issues in data mining, machine learning and large (sample size or dimension) data sets.

The goals of the DMML program were to advance significantly the understanding of fundamental statistical and computational issues in data mining, machine learning, and large data sets, to articulate future research needs for DMML, espe-

cially from the perspective of the statistical sciences, and to catalyze the formation of collaborations among statistical, mathematical, and computer scientists to pursue the research agenda.

By almost every measure, the program was a strong success. The high points are (i) a deeper understanding of the points of connection between data mining and statistical theory and methodology; (ii) effective analyses of a large, extremely complex testbed database provided by General Motors (GM), an affiliate of NISS and SAMSI, thus also strengthening SAMSI's industrial connections; (iii) a strong and continuing collaboration in the area of metabolomics, involving chemists, computer scientists, and statistical scientists; and (iv) a range of specific progress on issues ranging from false discovery rates to overcompleteness to support vector machines.

1.2.1.5 Network Modeling for the Internet, 2003–2004

Because of the size and complexity of the internet, and the nature of the protocols, internet traffic has proved to be very challenging to model effectively. Yet modeling is critical to improving quality of service and efficiency. The main research goal of this program was to address these issues by bringing together researchers from three communities: (a) applied probabilists studying heavy traffic queueing theory and fluid flow models; (b) mainstream internet traffic measurers/modelers and hardware/software architects; and (c) statisticians.

The timing was right for simultaneous interaction among all three communities, because of the trend away from dealing with quality of service issues through overprovisioning of equipment. This trend suggested that heavy traffic models would be ideally situated to play a leading role in future modeling of internet traffic, and in attaining deeper understanding of the complex drivers behind quality of service. The following topics were of particular interest.

Changepoints and extremes: To explore approaches in which the transitions between bursty and non-bursty internet traffic are considered changepoints, with models and methods from extreme value theory used to characterize the bursty periods.

Formulation of suite of models: To develop useful statistical models for internet traffic flow that are simple to analyze and simulate, and can capture the characteristics of actual traffic data that are important to electronics engineers and computer scientists.

Multifractional Brownian and stable motion: To capture two major characteristic features of the network traffic: time scale invariance (statistical self-similarity) and long-range dependence.

Structural breaks: To explore structural breaks in the context of internet traffic modeling where evidence of long-range dependence is also ubiquitous.

1.2.1.6 Multiscale Model Development and Control Design, 2003–2004

Multiscale analysis is ubiquitous in the modeling, design, and control of high performance systems utilizing novel material architectures. In applications including quantum computing, nanopositioning, granular flows, artificial muscle design, flow control, liquid crystal polymers, and actuator implants to stimulate tissue and bone growth, it is necessary to develop multiscale modeling hierarchies ranging from quantum to system levels for time scales ranging from nanoseconds to hours. Control techniques must be designed in concert with the models to guarantee the symbiosis required to achieve the novel design specifications. A crucial component of multiscale analysis is the development of homogenization techniques to bridge disparate temporal and spatial scales. This is necessitated by the fact that even with projected computing capabilities, monoscale models are prohibitively large to permit feasible system design or control implementation.

A number of these issues can be illustrated in the context of two prototypical materials, piezoceramics and ionic polymers, which are being considered for applications ranging from quantum storage to artificial muscle design. In both cases, the unique transducer properties provided by the compounds are inherently coupled to highly nonlinear dynamics which must be accommodated in material characterization, numerical approximation, device design, and control implementation.

The real-time approximation of comprehensive material models for device design and model-based control implementation is a significant challenge which must be addressed before novel material constructs and device architectures can achieve their full potential. The realization of these goals requires the development of reduced-order models which retain fundamental physics but are sufficiently low-order to permit real-time implementation.

Other components include control design, a crucial aspect of which is robustness with regard to disturbance and unmodeled dynamics. Deterministic robust control designs often provide uncertainty bounds which are overly conservative and hence provide limited control authority. Alternatively, one can provide statistical bounds on uncertainties.

1.2.1.7 Computational Biology of Infectious Diseases, 2004–2005

Infectious disease remains a major cause of suffering and mortality among people in the developing world, and a constant threat worldwide. The advent of genome science and the continuing rapid growth of computational resources together heralded an opportunity for the mathematical and statistical sciences to play a key role in the elucidation of pathogenesis and immunity and in the development of the next generation of therapies and global strategies. The program encompassed both genomic and population-level studies, including microbial and immunological genomics, vaccine design and proteomics, drug target identification, gene expression modeling and analysis, molecular evolution of host-microbe systems, drug resis-

tance, epidemiology and public health, systems immunology, and microbial ecology.

1.2.1.8 Latent Variable Models in the Social Sciences, 2004–2005

Latent variables are widespread in the social sciences. Whether it is intelligence or socioeconomic status, many variables cannot be directly measured. Factor analysis, latent class analysis, structural equation models, error-in-variable models, and item response theory illustrate models that incorporate latent variables. Issues of causality, multilevel models, longitudinal data, measurement error and categorical variables in latent variable models were examined.

Categorical variables. Many categorical observed variables in the social sciences are imperfect measures of underlying latent variables. Statistical issues emerge when categorical observed variables are part of a model with latent variables or with measurement error. There were mainly two approaches to models with categorical outcomes: James Hardin's recent advances in models that correct for measurement error in nonlinear models within the GLM framework, and Ken Bollen's two-stage least squares approach to latent variable models.

Complex surveys. A range of issues arise from survey data. Among these, two issues were of particular interest: (i) latent class analysis (LCA) of measurement error in surveys; and (ii) weighting and estimation for complex sample designs.

1.2.1.9 Data Assimilation for Geophysical Systems, 2004–2005

Data assimilation aims at accurate re-analysis, estimation and prediction of an unknown, true state by merging observed information into a model. This issue arises in all scientific areas that enjoy a profusion of data.

The problem of assimilating data into a geophysical system, such as the one related to the atmosphere or oceans, is both fundamental in that it aims at the estimation and prediction of an unknown true state, and challenging as it does not naturally afford a clean solution. It has two equally important elements: observations and computational models. Observations measured by instruments provide direct information of the true state. Such observations are heterogeneous, inhomogeneous in space, irregular in time, and subject to differing accuracies. In contrast, computational models use knowledge of the underlying physics and dynamics to provide a description of state evolution in time. Models are also far from perfect: due to model error, uncertainty in the initial conditions and computational limitations, model evolution cannot accurately generate the true state.

In its broadest sense, data assimilation (henceforth referred to as DA) arises in all scientific areas that enjoy a profusion of data. By its very nature, DA is a complex interdisciplinary subject that involves statistics, applied mathematics, and the relevant domain science. Driven by operational demands for numerical weather pre-

diction, however, the development of DA so far has been predominantly led by the geophysical community. In order to further DA beyond the current state-of-art, the development of effective methods must now be viewed as one of the fundamental challenges in scientific prediction.

1.2.1.10 Financial Mathematics, Statistics and Econometrics, 2005–2006

The goal of the study on Financial Mathematics, Statistics and Econometrics (FMSE) was to identify short and long term research directions deemed necessary to achieve both fundamental and practical advances in this rapidly growing field and to initiate collaborative research programs — among mathematicians, statisticians, and economists — focused on the multi-disciplinary and overlapping set of fields which involves disciplines such as: Applied Mathematics, Economics and Finance, Econometrics, and Statistics. A prominent theme throughout both the workshop and program was the necessity of exploiting the natural synergy between areas of financial mathematics, statistics, and econometrics. The goal of the SAMSI program in Financial Mathematics and Econometrics was to bring together these disciplines and initiate a discussion regarding what is really important and what is missing in three essential tasks.

Modeling. Model development was considered in domains ranging from financial and energy derivatives to real options.

Data. The size of financial data can be considerable when looking at high frequency data for large numbers of stocks for example.

Computation. Once a model has been written and calibrated to data, it remains necessary to compute quantities of interest. These three key themes transpired through the entire program and all its activities, including workshops, courses, and the diversity of visitors and participants.

1.2.1.11 National Defense and Homeland Security, 2005–2006

For several years, groups of researchers have been seeking to define appropriate roles for the statistical sciences, applied mathematical sciences, and decision sciences in problems of National Defense and Homeland Security (NDHS). Many efforts have focused on short-term applicability of existing methods and tools, rather than articulating or initiating a longer-term research agenda. Moreover, none of them has really spanned the statistical sciences, the applied mathematical sciences, and the decision sciences. Perhaps most important point is that, despite progress, these efforts have not "jelled" to produce a self-sustaining research momentum in the statistical sciences, applied mathematical sciences, and decision sciences on problems of NDHS. The NDHS program was meant fill this gap, in part by providing proof of concept that the necessary collaborations are feasible.

Research in the NDHS program was cross-disciplinary, and many of the efforts it catalyzed were addressable only by multi-institution teams of researchers. Some of the research needed to address problems in NDHS is technology-oriented; for example, development of sensors or biometric identification devices, but it is clear that data (statistical sciences), models (applied mathematics), and decisions (decision sciences) are essential components of the effort. Theory and methodology foci of the program included:

Biointelligence. It intersects planned development of a CDC Biointelligence Center. Such a center would need to analyze data from a variety of sources, which have differing characteristics in terms of temporal and spatial resolution, seasonal and regional variation, accuracy, completeness, and complexity.

Real-time inference, also known as data streams. Clearly many of these problems present deep questions of estimation and control, and so naturally require collaboration of statistical and applied mathematical scientists. These are also decision problems, leading naturally to engagement of the decision sciences and operations research.

Anomaly detection. The particular attention is paid to multivariate (possibly very high-dimensional) data, extremely rare events and false positives.

Data integration. It focuses on attendant problems of privacy, confidentiality and "new forms" of data such as images or biometric identifications.

Dynamics of massive databases. It focuses in part on a fundamental issue of data quality. How long is it before an accumulating database (e.g., one containing facial images) becomes hopelessly contaminated? Is the contamination global or local? What strategies can retard or reverse the process?

1.2.1.12 Astrostatistics, 2006

Historically, astronomy has served as fertile ground for stimulating the growth of new statistical and mathematical methodologies. Conversely, coping with the current and future needs of astronomy missions requires concerted efforts by cross-disciplinary collaborations involving astronomers, computer scientists, mathematicians, and statisticians. The following areas were of particular interest in the program.

Exoplanets: Including work on individual systems, populations studies, and optimal scheduling of observations; and in particular, statistical inference with small samples and nonlinear models, MCMC algorithms for nonlinear/multimodel problems, hierarchical/empirical Bayesian methods, and experimental design.

Surveys and population studies: Including work on small or moderate scale surveys (exoplanets, Kuiper belt objects, GRBs), large-scale surveys (galaxy surveys, AGN surveys), modeling "Number-Size" or "Size-Frequency" distributions (power-law models, nonparametric models, comparison with large-scale numerical sim-

ulations), selection effects and source uncertainties (Malmquist-Eddington/Lutz-Kelker biases, handling upper limits, etc.), and coincidences between surveys. Statistics areas include survey sampling, regression and measurement error models/EVM, nonparametric modeling, survival analysis, and multiple testing.

Gravitational lensing: Including work on the determination of basic statistics for random number of micro-images due to stars across a smooth dark matter background with external shear, computation of magnification probability distributions due to dark matter substructures, development of a statistical model of substructure population on galaxy-cluster scales, and exploration of the solution space for mass reconstruction.

Source detection and feature detection: Including work on detecting point sources and extended sources in images, classification of detections, and anomaly detection.

1.2.1.13 High Dimensional Inference and Random Matrices, 2006–2007

Random matrix theory lies at the confluence of several areas of mathematics, especially number theory, combinatorics, dynamical systems, diffusion processes, probability, and statistics. At the same time random matrix theory may hold the key to solving critical problems for a broad range of complex systems from biophysics to quantum chaos to signals and communication theory to machine learning to finance to geoscience modeling.

The following areas were of particular interest in the program.

Direct problems: Understanding the spectral properties of random matrices under various models and assumptions. The follows topics are included.

Extreme eigenvalues of random covariance matrices: Asymptotic and non-asymptotic distributions, in the Gaussian setting; Robustness of results to Gaussian assumption; Non-diagonal covariance matrices.

Dynamic behavior of eigenvalues of matrix processes: Matrices whose elements undergo diffusion (Dyson Processes); Stochastic differential equations for their eigenvalues; Scaling limits (Airy processes) and descriptions via partial differential equations.

Limiting theory for eigenvectors: Using free probability techniques.

Spectral properties of other random matrices arising in multivariate statistics: Techniques such as canonical correlation analysis and related random matrix problems; and Covariance matrices with covariance in space and time.

Inverse problems: Recovering from an observed random matrix information about the process from which it was generated. The following were of interest.

Estimation of large covariance matrices: Regularization techniques by banding, filtering, and using L1 penalties; and their theoretical and practical properties.

Consistency and estimability problems: Consistency of sample eigenvectors in covariance estimation problems; and Estimability issues for eigenvectors and eigenvalues.

Applications: Use of newly gained understanding of random matrices to advance research in a wide array of scientific disciplines. It includes the following applications.

Climatology: Empirical orthogonal functions and related techniques; and Summarization of evolving geophysical fields via spectral techniques.

Dynamical systems: Design of snapshots, i.e., finding a minimal number of functions, tailored to a specific problem, that accurately represent the problem's dynamics.

Data assimilation: Combination of numerical models and observations; Ensemble Kalman filtering: impact of sample information on propagation of covariance, effect of ensemble size, and tapering and inflation methods.

Graphical Gaussian models: Consistent estimation of the model structure; and Study of the rates of convergence.

Computation and connection: Computation of moments of large random Wishart matrices; Connection to graph theory; and Connection to methods in free probability.

General statistical inference: Consistency of regression functions dependent upon the behavior of certain large random matrices; and Problems arising in estimation, testing and model selection.

1.2.1.14 Development, Assessment and Utilization of Complex Computer Models, 2006–2007

Mathematical models intended for computational simulation of complex real-world processes are a crucial ingredient in virtually every field of science, engineering, medicine, and business, and in everyday life as well. Cellular telephones attempt to meet a caller's needs by optimizing a network model that adapts to local data, and people threatened by hurricanes decide whether to stay or flee depending on the predictions of a continuously updated computational model.

Two related but independent phenomena have led to the near-ubiquity of models: the remarkable growth in computing power and the matching gains in algorithmic speed and accuracy. Together, these factors have vastly increased the applicability and reliability of simulation—not only by drastically reducing simulation time, thus permitting solution of larger and larger problems, but also by allowing simulation of previously intractable problems.

The intellectual content of computational modeling comes from a variety of disciplines, including statistics and probability, applied mathematics, operations research, and computer science, and the application areas are remarkably diverse. Despite this diversity of methodology and application, there are a variety of common

challenges, detailed below, in developing, evaluating, and using complex computer models of processes.

Engineering subprogram. The engineering subprogram studied three frequently occurring problem areas in finite-element and other engineering models. These problems are those of validation, calibration, and combining data from physical experiments and computer experiments. The emphasis was on applications where the computer models require substantial running times and the physical models are difficult or expensive, so that, in some cases, physical experiments can be conducted for only subcomponents of the desired system or a physical simulator may only be possible for the desired system. Issues of combining codes from system components to produce valid codes for the entire system can then arise.

Two types of biological modeling. (a) To explore the impact of drug therapy and resistance on acute viral infections, these models are based on a multi-scale approach, integrating within-host models (i.e., ones that describe infection within a given individual) with between-host (epidemiological) models that describe the spread of infection at the population level. (b) The system biological models range from small biochemical networks corresponding to sets of coupled ODEs to large spatio-temporal models requiring advanced numerical methods.

Methodology subprogram. The subprogram engaged in an in-depth treatment of methodological issues that arise in the design, analysis, and utilization of computer models across many fields of application. The subprogram also evolved in close collaboration with the four disciplinary subprograms, engaging them in an overall research umbrella.

1.2.1.15 Multiplicity and Reproducibility in Scientific Studies, 2006

Concerns over multiplicities in statistical analysis and reproducibility of scientific experiments are becoming increasingly prominent in almost every scientific discipline, as experimental and computational capabilities have vastly increased in recent years. The following key issues were of interest.

Reproducibility. Scientists use statistical methods to help them judge if something has happened beyond chance. They expect that if others replicate their work, that a similar finding will happen. To clear a drug the FDA requires two studies, each significant at 0.05. Ioannidis (2005) showed startling and disconcerting lack of reproducibility of influential statistical studies published in major medical journals. It found that about 30% of randomized, double-blinded medical trials failed to replicate and that 5 out of 6 non-randomized studies failed to replicate — about an 80% failure rate. It is necessary to explore and clarify the causes of failures to reproduce and, more broadly, to identify commonalities that lead to these problems, and attempting to estimate its prevalence. Multiplicities (both obvious and hidden) need be considered, along with selection biases and regression to the mean.

Subgroup analysis. Large, complex data sets are becoming more commonplace and people want to know which subgroups are responding differently to one another and why. The overall sample is often quite large, but subgroups may be very small and there are often many questions. Genetic data are being collected on clinical trials. Which patients will respond better to a drug and which will have more severe side effects? Disease, drug, or side effects can result from different mechanisms. Identification of subgroups of people where there is a common mechanism is useful for diagnosis and prescribing of treatment. Large educational surveys involve groups with different demographics, different educational resources and subject to different educational practices. What groups are different and how are differences related to resources and practices? What really works and why? Is the finding the result of chance? There is a need for effective statistical methods for finding subgroups that are responding differently. There is a need to be able to identify complex patterns of response and not be fooled by false positive results that come about from multiple testing.

Massive multiple testing. The routine use of massively multiple comparisons in inference for large-scale genomic data has generated a controversy and discussion about appropriate ways to adjust for multiplicities. There are different approaches to formally describe and address the multiplicity problem, including the control of various error rates, decision theoretic approaches, hierarchical modeling, probability models on the space of multiplicities, and model selection techniques. Besides applications in inference for genomic data, similar problems occured in clinical trial design and analysis, record matching problems, classification in spatial inference, anomaly discovery, and syndrome surveillance. It is interesting to identify the relative merits and limitations of the competing approaches for diverse applications, and to understand which features of reproducibility are addressed.

1.2.1.16 Risk Analysis, Extreme Events, and Decision Theory, 2007–2008

Over the past several years, there has been a wealth of scientific progress on risk analysis. As the set of underlying problems has become increasingly diverse, drawing from areas ranging from national defense and homeland security to genetically modified organisms to animal disease epidemics and public health to critical infrastructure, much research has become narrowly focused on a single area. It has also become clear, however, that the need is urgent and compelling for research on risk analysis, extreme events (such as major hurricanes), and decision theory in a broader context. Availability of past information, expert opinion, complex system models, and financial or other cost implications as well as the space of possible decisions may be used to characterize the risks in different settings.

Risk analysis and extreme events also carry a significant public policy component, which is driven in part by the increasing stakes and the multiplicity of stakeholders. In particular, policy concerns direct attention not only to the dramatic risks for huge numbers of people associated, for example, with events of the magnitude of Hurricane Katrina or bioterrorism, but also to “small-scale” risks such as drug in-

teractions driven by rare combinations of genetic factors. The following key issues were of interest in the program.

Extreme values: Theory for multidimensional extremes. Probability theory and statistical methodology for one-dimensional distributions of extreme values have been developed over the past half-century, but these do not extend easily to higher dimensions.

Financial risk: Risk assessment and risk management for critical resources, infrastructure, and energy markets. Prediction of financial consequences of extreme events is formulated differently in actuarial science, in operations research, in financial mathematics, and in risk management and decision sciences.

Experts and decisions: Prior elicitation and modeling with expert opinion. Elicitation is the process of formulating a person's knowledge and beliefs about one or more unknowns (parameters) into a (joint) probability distribution for those unknowns in a decision-making setting.

Adversarial risk: Formalizing analysis of risk from intelligent opponents. Traditional methodology is based on the canonical risk equation: (Likelihood of Attack) $\times$ (Consequence) $\times$ (1-System Effectiveness) $=$ Risk. However, new business and government scenarios require considering risk analysis that takes into account opponents' intelligence, possible willingness to cooperate, and transfer of risk following decisions and actions.

Environmental risk analysis: Ecological risk assessment and risks associated with extreme climatic events. The occurrence of extreme environmental disruptions depends jointly on the highly unstable occurrence of these events and the locally, at least, highly variable consequences.

Industrial risk: Pharmaceutical and health risk. Statistical issues include inference from non-randomized clinical trials, multiplicity and asymmetries between null and alternative hypotheses, analysis of rare events, and competing risks.

1.2.1.17 Random Media, 2007–2008

The field of random media is a classical one which is presently receiving widespread attention as new theory, approximation techniques, and computational capabilities are applied to emerging applications. Due to the breadth of the field, the inherent deterministic, stochastic, and applied components have typically been investigated in isolation. However, it is increasingly recognized that these components are inexorably coupled and that synergistic investigations are necessary to provide significant fundamental and technological advances in the field. The following key issues were of interest.

Time reversal. The component on time reversal built on recent analysis and experimental observations that time reversal of waves propagating in disordered media permit refocusing. This somewhat unexpected property has profound ramifications

in domains such as wireless communications, medical imaging, nondestructive evaluation, and underwater acoustics. Whereas the behavior of one-dimensional acoustic waves is mathematically and statistically understood, questions regarding multidimensional media remain wide open with the exception of the baraxial wave equation.

Interface problems. Interface problems arise in a diverse range of applications, including multiphase flows and phase transitions in fluid mechanics, thin film and crystal growth simulations in material science, and mathematical biology problems modeled by partial differential equations involving moving fronts.

Imaging problems. Imaging problems in random media arise in a number of applications, including biomedical imaging and seismic analysis. In the latter category, a detailed knowledge of earth medium heterogeneities is necessary for oil and gas recovery, earthquake and volcanic predictions, and environmental analysis.

Scattering theory. Whereas mathematical scattering theory for one-dimensional regimes is fairly mature, it was of interest to extend to multidimensional media with the exception of the baraxial wave equation.

Porous media. It includes topics pertaining to stochastic transport processes and physics associated with porous media.

1.2.1.18 Environmental Sensor Networks, 2007–2008

Environmental sensor networks have the capability of capturing local and broadly dispersed information simultaneously; they also have the capacity to respond to sudden change in one location by triggering observations selectively across the network while simultaneously updating the underlying complex system model and/or reconfiguring the network. Data gathered by wireless sensor networks, either fixed or mobile, pose unique challenges for environmental modeling: a complex system is being observed by a dynamical network. Technical challenges in statistics (sampling design to prediction and prediction uncertainty), in mathematics (computational geometry to data fusion to robotics), and in computers science (self-organizing networks to algorithm analysis) combine with the technical challenges of the models themselves and the sciences that underlie them. The following key issues were of interest.

Sampling from wireless networks. Cost of spatio-temporal data in terms of both energy and delay: Each sample has a footprint in power in space and time, some value to one or more process models (e.g., importance of parameters in space and time, sensitivity of estimates to the observation), and some cost (e.g., data transmission). One wants to know if it is possible to derive frameworks such that the utility of each sample exceeds its cost.

Environmental modeling from sensor networks. Model complexity and adequacy: In the trade-off of dimensionality and predictive accuracy, what are the diagnostics for excessive vs. insufficient parametrization? Can models be developed so

that reduced forms (e.g., deleting submodels, subsets of parameters, or reducing resolution of observations and/or parameter specification) still function simultaneously with near-optimality at several scales?

Networks, forests, and global change. Process level understanding of how forested ecosystems respond to global change is critical for anticipating consequences of human impacts on landscapes. To be successful an approach will entail integrated models of a complex system and will involve heterogeneous data and analyses that directly address uncertainty and model selection issues.

1.2.1.19 Challenges in Dynamic Treatment Regimes and Multistage Decision Making, Summer 2007

The management of chronic disorders, such as mental illness, substance dependence, cancer, and HIV infection, presents considerable challenges. In particular the heterogeneity in response, the potential for relapse, burdensome treatments, and problems with adherence demand that treatment of these disorders involve a series of clinical decisions made over time. Decisions need to be made about when to change treatment dose or type and regarding which treatment should be used next. Indeed, clinicians routinely and freely tailor treatment to the characteristics of the individual patient with a goal of maximizing favorable outcomes for that patient. To a large extent the tailoring of sequences of treatments is based on clinical judgment and instinct rather than a formal, evidence-based process.

These realities have led to great interest in the development of so-called "dynamic treatment regimes" or "adaptive treatment strategies." A dynamic treatment regime is an explicit, operationalized series of decision rules specifying how treatment level and type should vary over time. The rule at each stage uses time-varying measurements of response, adherence, and other patient characteristics up to that point to determine the next treatment level and type to be administered, thereby tailoring treatment decisions to the patient. The objective in developing such multistage decision-making strategies is to improve patient outcomes over time.

Methodology for designing dynamic treatment regimes is an emerging area that presents challenges in two areas. First, experimental designs for collecting suitable data that can be used efficiently to develop dynamic regimes are required. Second, techniques for using these and other data to deduce the decision-making rules involved in a dynamic regime must be developed. In both areas, input from researchers in a variety of disciplines and collaborations among them will be critical.

1.2.1.20 Geometry and Statistics of Shape Spaces, Summer 2007

Shapes are prevalent in the outside world and in science. They manifest themselves in live animals, plants, landscapes, or in man-made materials, like cars, planes, building, bridges, and they are designed from aesthetic as well as efficacy considerations. Internal organs of humans or other animals also have a commonly accepted,

well-defined shape, and their study is an old science called anatomy. For the human mind, there is an intuitive notion of what shapes are, why they differ or look alike, or when they present abnormalities with respect to ordinary observations. Sculpture is the art of rendering existing shapes, or creating new ones, and the fact that artists are still able to provide unambiguous instances of subjects through distorted or schematic representations is a strong indication of the robustness of the human shape recognition engine.

However, an analytical description of a shape is much less obvious, and humans are much less efficient for this task, as if the understanding and recognition of forms works without an accurate extraction of their constituting components, which is probably the case. We can recognize a squash from an eggplant or a pepper using a simple outline, and even provide a series of discriminative features we can distinguish, but it is much harder to instantiate a verbal description of any of them, accurate enough, say for a painter to reproduce it. It is therefore not surprising that, for mathematics, shape description remains mostly a challenge. The last fifty years of research in computer vision has shown an amazingly large variety of points of view and techniques designed for this purpose: 2D or 3D sets they delineate (via either volume or boundary), moment-based features, medial axes or surfaces, null sets of polynomials, and configurations of points of interest (landmarks), to name but a few.

But beyond the shape characterization issue, the more ambitious program which has interested a large group of researchers during the last two decades, starting with the seminal work of David Kendall, is the study of shapes spaces and their statistics. Here shapes are not only considered individually, but they are seen as variables, belonging to some generally infinite dimensional space which possesses a specific geometry. The theoretical study of such spaces, the definition of computationally feasible algorithmic and statistical procedures has been the subject of a still-growing line of work. For example, Kendall's original contribution focused on collections of landmarks modulo the action of rotation and scale. It has since been extended to the actions of other groups and to plane curves instead of points. Other examples build shape spaces using the medial axis representation. The last few years have seen the emergence and the development of several new techniques, building infinite dimensional Riemannian metrics on curves and other shape representations, involving several groups over the world. Within applied mathematics, the analysis of shape spaces arises at a nodal point in which geometry, statistics, and numerical analysis each have a fundamental contribution.

1.2.1.21 Meta-analysis: Synthesis and Appraisal of Multiple Sources of Empirical Evidence, Summer 2008

Seldom is there only a single empirical research study relevant to a question of scientific interest. However, both experimental and observational studies have traditionally been analyzed in isolation, without regard for previous similar or other closely related studies. A new research area has arisen to address the location, ap-

praisal, reconstruction, quantification, contrast, and possible combination of similar sources of evidence. Variously called meta-analysis, systematic reviewing, research synthesis or evidence synthesis, this new field is gaining popularity in diverse fields including medicine, psychology, epidemiology, education, genetics, ecology, and criminology.

The combination of results from similar studies is often known simply as meta-analysis. Common examples are combining results of randomized controlled trials of the same intervention in evidence-based medicine; similarly across studies in social science; or of odds ratios measuring association between an exposure and an outcome in epidemiology. More complex syntheses of multiple sources of evidence have developed recently, including combined analyses of clinical trials of different interventions, and combined analysis of data from multiple microarray experiments (sometimes called cross-study analysis). For straightforward meta-analyses, general least-squares methods may be used, but for complex meta-analyses, the technical statistical approach is not so obvious. Often likelihood and Bayesian approaches provide very different perspectives; and in practice the possible benefits of more complex approaches may be hard to discern as many meta-analyses are compromised by limited or biased availability of data from studies as well as by varying methodological limitations of the studies themselves.

The presence of multiple sources of evidence has long been a recognized challenge in the development and appraisal of statistical methods, from Laplace and Gauss to Fisher and Lindley. In the 1980s Richard Peto argued that a combined analysis would be more important than the individual analyses, a view taken still further by Greenland, who has suggested that that individual study publications should not attempt to draw conclusions at all, but should instead only describe and report results, so that a later meta-analysis can more appropriately assess the study's evidence fully informed by other study designs and results. Will combined analyses actually replace individual analyses (or at least decrease their impact)? If so, it is time to re-examine the perennial problems of statistical inference in this context.

There are three challenges: (i) to substantiate and clarify how existing statistical methodology can effectively combine multiple sources of evidence, given perfect conduct and reporting of all studies; (ii) to identify statistical areas in need of development or improvement, both in theory and application, for the practical situations of studies having methodological limitations and studies providing biased or incomplete data; and (iii) To identify and develop material and pedagogy for undergraduate and graduate programs in statistics, to allow future statisticians to deal effectively with multiple sources of evidence, and to motivate further development of new methodology.

1.2.1.22 Sequential Monte Carlo Methods, 2008–2009

Monte Carlo (MC) methods are central to modern numerical modeling and computation in complex systems. Markov chain Monte Carlo (MCMC) methods provide enormous scope for realistic statistical modeling and have attracted much attention

from disciplinary scientists as well as research statisticians. Many scientific problems are not, however, naturally posed in a form accessible to evaluation via MCMC, and many are inaccessible to such methods in any practical sense. For example, for real-time, fast data processing problems that inherently involve sequential analysis, MCMC methods are often not obviously appropriate at all due to their inherent "batch" nature. The recent emergence of sequential MC concepts and techniques has led to a swift uptake of basic forms of sequential methods across several areas, including communications engineering and signal processing, robotics, computer vision, and financial time series. This adoption by practitioners reflects the need for new methods and the early successes and attractiveness of SMC methods. In such, probability distributions of interest are approximated by large clouds of random samples that evolve as data are processed using a combination of sequential importance sampling and resampling ideas. Variants of particle filtering, sequential importance sampling, sequential and adaptive Metropolis MC and stochastic search, and others have emerged and are becoming popular for solving variants of "filtering" problems; i.e. sequentially revising sequences of probability distributions for complex state-space models.

Many problems and existing simulation methods can be formulated for analysis via SMC: sequential and batch Bayesian inference, computation of p-values, inference in contingency tables, rare event probabilities, optimization, counting the number of objects with a certain property for combinatorial structures, computation of eigenvalues and eigenmeasures of positive operators, PDEs admitting a Feynman-Kac representation, and so on. This research area is poised to explode, as witnessed by this major growth in adoption of the methods. The following key issues were of interest.

Continuous time modeling and parameter estimation. Modeling and parameter estimation for continuous time stochastic processes includes exact simulation methods for inference in partially observed diffusions, jump diffusions and Levy processes. Both batch-based and on-line strategies were studied, as were both parameter estimation and state estimation.

Tracking and large-scale dynamical systems. There is much interest in tracking and inference for large groups of objects, with applications in medical imaging, dynamic object tracking in robotic control in industrial, commercial, and military areas, and tracking in media applications. Particular focus areas were drawn from representations of many interacting objects using random fields, graphical models, and automated inference about group structures, types of interaction, intentionalities, etc.

Decision making, econometrics, and finance. SMC methods are under-explored and appear to have a great deal of potential in problems of numerical solution of decision problems under uncertainty. Research areas included: (i) applications in policy-oriented macro-economic modeling; and (ii) state and parameter estimation leading into prediction financial time series models, together with numerical approaches to the coupled portfolio decision problem.

Population Monte Carlo. MCMC can easily get stuck for high-dimensional multimodal distributions. This has led, in part, to the development of adaptive and population Monte Carlo algorithms, which can provide promising alternatives for these problems. Some of these methods are inherently SMC methods, and there was interest in developing new adaptive methods to sample from high-dimensional distributions.

1.2.1.23 Algebraic Methods in Systems Biology and Statistics, 2008–2009

In recent years, methods from algebra, algebraic geometry, and discrete mathematics have found new and unexpected applications in systems biology as well as in statistics, leading to the emerging new fields of "algebraic biology" and "algebraic statistics." Furthermore, there are emerging applications of algebraic statistics to problems in biology. There has been development and maturation of these two areas of research as well as their interconnections. The common mathematical tool set as well as the increasingly close interaction between biology and statistics allows researchers working in algebra, algebraic geometry, discrete mathematics, and mathematical logic to interact with statisticians and biologists and make fundamental advances in the development and application of algebraic methods to systems biology and statistics. The following key issues were of interest.

Systems biology. The development of revolutionary new technologies for high-throughput data generation in molecular biology in the last decades has made it possible for the first time to obtain a system-level view of the molecular networks that govern cellular and organismal function. Whole genome sequencing is now commonplace, gene transcription can be observed at the system level, and large-scale protein and metabolite measurements are maturing into a quantitative methodology. The field of systems biology has evolved to take advantage of this new type of data for the construction of large-scale mathematical models. System-level approaches to biochemical network analysis and modeling promise to have a major impact on biomedicine, in particular drug discovery.

Statistics. It has long been recognized that the geometry of the parameter spaces of statistical models determines in fundamental ways the behavior of procedures for statistical inference. This connection has in particular been the object of study in the field of information geometry, where differential geometric techniques are applied to obtain an improved understanding of inference procedures in smooth models. Many statistical models, however, have parameter spaces that are not smooth but have singularities. Typical examples include hidden variables models such as the phylogenetic tree models and the hidden Markov models that are ubiquitous in the analysis of biological data. Algebraic geometry provides the necessary mathematical tools to study non-smooth models and is likely to be an influential ingredient in a general statistical theory for non-smooth models.

Algebraic methods. Algebraic biology is emerging as a new approach to modeling and analysis of biological systems using tools from algebra, algebraic geometry,

discrete mathematics, and mathematical logic. Application areas cover a wide range of molecular biology, from the analysis of DNA and protein sequence data to the study of secondary RNA structures, assembly of viruses, modeling of cellular biochemical networks, and algebraic model checking for metabolic networks, to name a few.

1.2.1.24 Psychometrics, Summer 2009

Much of current psychometric research involves the development of novel statistical methodology to model educational and psychological processes, and a wide variety of new psychometric models have appeared over the last quarter century. Such models include (but are not limited to) extensions of item response theory (IRT) models, cognitive diagnosis models, and generalized linear latent and mixed models. The development of several of these models has been spearheaded by quantitative psychologists, a group of researchers who find their academic homes primarily in psychology and education departments. During the same period, very similar models and methodologies were developed, often independently, by academic statisticians residing in mathematics and statistics departments. The interaction between these two groups has made a substantial effort to develop methodology crucial to both fields.

The following statistical models are of interest: IRT models, cognitive diagnostic models, and variations of generalized linear latent and mixed models. Practical applications of methodology include (i) the analysis of patient reported outcomes (PROs), (ii) journal and grant peer review, and (iii) cognitive diagnostic models.

1.2.2 Research Topics from Current SAMSI Programs

1.2.2.1 Stochastic Dynamics, 2009–2010

The broad topic of stochastic dynamics includes analysis, computational methods, and applications of systems governed by stochastic differential equations. Two application areas are being emphasized: problems in biological sciences and dynamics of networks.

Stochastic analysis and numerical methods. In recent years it has become increasingly clear that to effectively understand complex stochastic systems, a combination of modern numerical analysis, estimation and sampling techniques, and rigorous analysis of stochastic dynamics is required. Whether one speaks of path sampling techniques, estimation in complex non-linear dynamics, or simulation of rare events it is important to bring both sophisticated analytic tools and an understanding of what one can compute efficiently.

Multi-scale and multi-physics computing. The classical continuum equations arising in fluid flow, elasticity, or electromagnetic propagation in materials require constitutive laws to derive a closed-form system. In such cases, Gaussian statistics are well verified at the microscopic level and the moments of Gaussian distributions can be computed analytically. However, many physical processes exhibit significant localized departure from Gaussian statistics. For example, when a solid breaks, the motion of the atoms in the crystalline lattice along the crack propagation path is no longer governed by a Maxwell-Boltzmann distribution. An interesting characteristic is that macroscopic features impose the departure from local thermodynamic equilibrium and macroscopic quantities of are practical interest. the solid deforms before it breaks. New statistical methods are needed for the complicated situations.

Stochastic modeling and computation in biology. The explosion of interest in mathematical and statistical modeling and computation in the biological and medical sciences, where stochasticity is present at nearly every scale, has been one of the most exciting trends in the biosciences in the last ten years. Math biology has grown from a niche area to a major research group in many US math departments, and graduate programs in mathematics are scrambling to cope with a wave of students seeking to do graduate work in interdisciplinary research areas. Programs in biostatistics, bio-informatics, and bio-medical engineering are also seeing increased growth.

Dynamics of social networks. Social network data are distinguished by the inherent dependencies among units. These dependencies, usually represented by binary or more general links, are a primary focus of many analyses. To date, however, both models of, and techniques for inference for, social networks, have focused on static networks. Virtually no extant models address the appearance or disappearance of nodes, the evolution of link existence or strength, or the characteristics of nodes. One SAMSI-generated exception is Banks et al. (2008), which uses stochastic differential equations to model joint evolution of the edge set and node characteristics, with a focus on characterizing the role of stochastic variability. Indeed, at an NSF workshop in October 2007 on "Discovery in Complex or Massive Data: Common Statistical Themes," there was consensus about urgent need for models of the dynamics of networks and associated tools for inference.

1.2.2.2 Space-Time Analysis for Environmental Mapping, Epidemiology, and Climate Change, 2009–2010

It is challenging to deal with problems encountered in dealing with random space-time fields, both those that arise in nature and those that are used as statistical representations of other processes. The sub-themes of environmental mapping, spatial epidemiology, and climate change are interrelated both in terms of key issues in underlying science and in the statistical and mathematical methodologies needed to address the science. Researchers from statistics, applied mathematics, environmental sciences, epidemiology, and meteorology are involved, and the program is

promoting the opportunity for interdisciplinary, methodological, and theoretical research. The following key issues are of interest.

Environmental mapping. Spatial or spatial-temporal statistical analysis in environmetrics often entails the prediction of unobserved random fields over a dense grid of sites in a geographical domain, based on observational data from a limited number of sites and possibly simulated data generated by deterministic physical models. In important special cases, spatial prediction requires statisticians to estimate spatial covariance functions and generalized regression tools (also called geostatistical methods). Methods for spherical data, especially appropriate for climate research, are currently being developed, but they need to address complications similar to those that occur for multivariate random fields.

Spatial epidemiology. Many studies during the past two decades have demonstrated a statistical association between exposure to air pollutants (principally, particulate matter and ozone) with various (mostly acute) human health outcomes, including mortality, hospital admissions, and incidences of specific diseases such as asthma. While a number of different study designs have been used, two dominate. The first, the time series studies, relate variations in daily counts of these adverse health outcomes with variations in ambient air pollution concentrations through multiple regression models that include air pollution concentrations while removing the effects of long-term trends, day of week effects, as well as possible confounders such as meteorology. However, the relative health risks of air pollution are small say compared to smoking. Thus some studies have through Bayesian hierarchical modeling combined the estimated air pollution coefficients for various urban areas to borrow strength.

Climate change. Much of the case for climate change and the estimation of its deleterious effects has relied on deterministic climate models that embrace physical and chemical modeling. The GCM (General Climate (or Circulation) Model) yields simulated climate data at fairly coarse spatial scales that serves as input to the RGCM (regional GCM) that runs at finer spatial scales.

The results of climate models are extremely multi-dimensional. It is very difficult to present all of this information concisely in a manner that can be understood by decision makers. Dimension reduction and data presentation techniques are needed for contrasting spatial data, explaining what is being presented, and determining how to describe the confidence of projections from non-random samples.

1.2.3 Research Topics in Future Programs

1.2.3.1 Semiparametric Bayesian Inference: Applications in Pharmacokinetics and Pharmacodynamics, Summer 2010

Pharmacokinetics (PK) is the study of the time course of drug concentration resulting from a particular dosing regimen. PK is often studied in conjunction with

pharmacodynamics (PD). PD explores what the drug does to the body, i.e., the relationship of drug concentrations and a resulting pharmacological effect. Pharmacogenetics (PGx) studies the genetic variation that determines differing response to drugs. Understanding the PK, PD, and PGx of a drug is important for evaluating efficacy and determining how best to use such agents clinically.

Hierarchical models have allowed great progress in statistical inference in many application areas. Hierarchical models for PK and PD data that allow borrowing of strength across a patient population are known as population PK/PD models. These models have allowed investigators to learn about important sources of variation in drug absorption, disposition, metabolism, and excretion, allowing the researchers to begin to tailor drug therapy to individuals. Newer Bayesian non-parametric population models and semi-parametric models offer the promise of individualizing therapy and discovering subgroups among patients even further, by freeing modelers from restrictive assumptions about underlying distributions of key parameters across the population. The purpose of this program is to bring together a mix of experts in PK and PD modeling, non-parametric Bayesian inference, and computation.

The aims of the study are (i) to identify the critical new developments of inference methods for PK and PD data; (ii) to determine open challenges; and (iii) to establish inference for PK and PD as an important motivating application area of non-parametric Bayes. Among these, (iii) is particularly important for new and promising researchers.

1.2.3.2 Complex Networks, 2010–2011

Network science is highly interdisciplinary field and is characterized by novel interactions in the mathematical sciences occurring at the interface of applied mathematics, statistics, computer science, and statistical physics, as well as those areas with network-oriented thrusts in biology, computer networks, engineering, and the social sciences. Four research areas are as follows.

Network modeling and inference. The analysis of network data has become a major endeavor across the sciences, and network modeling plays a key role. Frequently, there is an inferential component to the process of network modeling, i.e., inference of network model parameters, of network summary measures, or of the network topology itself. For most standard types of data (e.g., independent and identically distributed, time series, spatial, etc.), there is a well-developed mathematical infrastructure guiding modeling and inference in practice.

Flows on networks. In their simplest form, network flows are defined on directed graphs. Each edge receives a flow in an amount that cannot exceed the capacity of the edge. Many transport applications correspond to network flows: hydraulics and pipeline flows, rivers, sewer and water systems, traffics and roads, supply chains and cardiovascular systems, to name but a few. Several by now classical problems for network flows such as maximum flow have been solved for static flow. These results only partially carry over to dynamic flows (time extended networks) and much re-

mains to be done. Some applications such as communication systems typically split data into packages. There are obvious technical limitations regarding the fineness of such decompositions that have to be taken into account when seeking (quasi-) optimal solutions. Several of the relevant open questions fall under the umbrella of combinatorial optimization.

Network models for disease transmission. Network models provide a natural way to model many infectious diseases. Many diseases, such as sexually transmitted infections (STIs), have long been studied in terms of networks, but in recent years the approach has been adopted in a wider range of disease settings, including acute rapidly spreading infections. Disease transmission networks are highly dependent on the infection of interest: the sexual partnership network across which an STI spreads has a quite different structure from the social network on which a respiratory infection (such as influenza) would spread. Even in the same population, different diseases "see" different networks.

Dynamics of networks. The changing structure of networks over time is inherent in the study of a broad array of phenomena. Examples for which a static transmission network is inadequate abound: from disease transmission to communications networks with changing landscape of connections to political networks where associations and voting similarities vary from one legislative session to the next. While the nature of the underlying processes differs, the flow of generalized information, for all three examples, depends in a nontrivial way on the changes in the node roles, in the structure of communities and in other coarse structural units. The importance of dynamics in networks has been long recognized. The increasing accessibility of network data has led to renewed interest in this area; examples of data include longitudinal data waves and financial correlations with strengths of connection defined over moving windows in time.

1.2.3.3 Analysis of Object Data, 2010–2011

Analysis of Object Data extends the very active research area of functional data analysis and generalizes the fundamental FDA concept of curves as data points, to the more general concept of objects as data points. Examples include images, shapes of objects in 3D, points on a manifold, tree structured objects, and various types of movies. Specific AOOD contexts can be grouped in a number of interesting ways.

Euclidean, i.e., (constant length) vectors of real numbers. There are two interesting areas. One area is on Functional Data Analysis (FDA), viewing curves as data. These curves are commonly either simply digitized, or else decomposed by a basis expansion, which gives a vector that represents each data curve. A second area is Time Dynamics Data, with an emphasis on differential equations and dynamic systems as drivers of fully or incompletely observed samples of stochastic processes. Various applications include engineering, biological modeling of growth or cell kinetics in control, the analysis of auction dynamics in e-commerce, repeated

events such as child births of a woman in the social sciences, the dynamics of HIV infections, and the dynamics of gene expression in medical studies.

Mildly non-Euclidean, i.e., points on a manifold and shapes. Shape Analysis and Manifold Data, where for example 2 or 3 dimensional locations of a set of common landmarks are collected into vectors that represent shapes. While these vectors are just standard multivariate data, they frequently violate standard multivariate assumptions, such as the sample size being (usually much) larger than the dimension. Research in the direction of High Dimension Low Sample Size (HDLSS) issues will be a major emphasis of the proposed SAMSI program. In addition, the landmarks may be invariant to certain transformations such as location, rotation, and scale, and Kendall's shape analysis of such objects leads to non-Euclidean distances being the most natural. Further recent examples include analysis of shapes of unlabeled points, especially on curves, surfaces, and images. The closely related manifold data also are based on non-Euclidean distances.

Strongly non-Euclidean, i.e., tree or graph structured objects. The data space admits no tangent plane approximation. Thus, there is no apparent approach to adapting even approximate Euclidean methodologies, and statistical analysis must be invented from the ground up.

1.3 Overview of the Book

The book is organized as follows. There are a total of 14 chapters. Each chapter consists of 2 to 4 sections reviewing research challenges related to a common theme. All cited references, author indices, and subject indices are placed at the end of the book.

Chapters 2 through 4 deal with the basic framework of Bayesian inference, including a discussion of prior choices in Chapter 2, some aspects of posterior inference in Chapter 3 and model comparison and testing in Chapter 4. Objective priors play a crucial role, both in classical and Bayesian inference. Chapter 2 discusses the constructions, the properties, and applications of reference priors. In Section 2.1, Bernardo and Tomazella use the classic hypothesis of the Hardy-Weinberg equilibrium to introduce objective Bayesian testing. In Section 2.2, Liseo, Tancredi, and Barbieri discuss and illustrate a novel approach for deriving reference priors when the statistical model can be expressed via a stochastic representation or a latent structure. In Section 2.3, Clarke and Yuan derive asymptotic expansions for prior-to-posterior distance under an empirical likelihood and use the results to propose a corresponding reference prior.

Chapter 3 reviews shrinkage estimation in Bayesian analysis and demonstrates that this long-standing research area still remains an active frontier with many open problems and new applications. In Section 3.1, Strawderman gives a decision theoretic account of Bayesian shrinkage estimation, focusing on inference in multivariate normal location models under quadratic loss. He emphasizes Stein-

type shrinkage and minimaxity in higher dimensions. In Section 3.2, George and Xu describe a variety of recent results that use a decision theoretic framework based on expected Kullback-Leibler loss to evaluate the long run performance of Bayesian predictive estimators. In particular, they focus on high dimensional prediction under a multivariate normal model and extensions to normal linear regression. In Section 3.3, Meng provides a spirited discussion of a recently published paper on shrinkage estimation for a study in gene-environment interaction. Please see `http://www.stat.uconn.edu/bergerbook.html` for a rejoinder from the authors of the original paper.

Chapter 4 reviews current research frontiers in model comparison and Bayesian testing. In Section 4.1, Steele and Raftery discuss Bayesian model selection for Gaussian mixture models. They report a simulation study to compare six methods for choosing the number of components in a mixture model. The design of the simulation study is based on a survey of literature on the estimation of mixture model parameters. In Section 4.2, Pericchi discusses alternative approaches for choosing an optimal training sample size. In Section 4.3, Johnson demonstrates that misspecification of the prior distribution used to define the alternative hypothesis always results in a decrease in the expected weight of evidence collected against the null hypothesis. In Section 4.4, Lipkovich, Ye, and Smith use simulation to evaluate the performance of Bayesian model averaging in linear regression models. They compare performance over different subsets of models and evaluate the importance of correlation among predictors on efficiency.

Chapters 5 and 6 introduce extensions of basic parametric Bayesian inference that give rise to challenging current research areas. Chapter 5 reviews Bayesian inference for large, complex and highly structured computer models, introducing the notions of simulation models and more efficient emulation models. In Section 5.1, Bayarri provides a review of Bayesian inference for deterministic computer models. She highlights the unavoidable non-identifiability issue and its consequences and discusses a partial solution to it. The calibration of computer models involves combining information from simulations of a complex computer model with physical observations of the process being simulated by the model. In Section 5.2, Bhat, Haran, and Goes study an approach for computer model calibration with multivariate spatial data. They demonstrate the application of this approach to the problem of inferring parameters in a climate model.

Chapter 6 deals with non-parametric and semi-parametric Bayesian inference. This is one of the currently fastest growing research areas in Bayesian inference. In Section 6.1, Tokdar, Chakrabarti, and Ghosh survey the literature on Bayes nonparametric testing. The discussion concentrates on Bayesian testing of goodness of fit for a parametric null with nonparametric alternatives. In Section 6.2, Lee reviews the species sampling model and issues that arise in statistical inference with species sampling model. In Section 6.3, Jordan discusses hierarchical and nested modeling concepts within the framework of Bayesian nonparametrics.

Chapter 7 reports on research challenges related to taking a critical second look at Bayesian inference, by questioning and critiquing the influence of the various elements of a Bayesian inference model. The investigation includes the evaluation

and consideration of frequentist summaries. In Section 7.1, Zhu, Ibrahim, Cho, and Tang provide a comparative review of three primary classes of Bayesian influence methods including Bayesian case influence measures, Bayesian global robustness, and Bayesian local robustness. They elaborate on the advantages and disadvantages of each class of Bayesian influence methods. In Section 7.2. Datta and Rao review the choice of hyperpriors in hierarchical models. They choose an approach based on frequentist validation, i.e., considering frequentist summaries of the implied inference to determine a prior choice. In Section 7.3, Dai and Sun discuss the objective inference of parameters in a multivariate normal model based on a class of objective priors on normal means and variance. They derive the exact frequentist matching priors for multivariate normal parameters and their related functions.

The next few chapters discuss research frontiers in Bayesian analysis that arise from important application areas. Chapter 8 discusses the rapidly expanding research area of Bayesian inference in clinical trial design. Many current problems are related to adaptive clinical trial design and hierarchical models to borrow strength across related subpopulations and studies. In Section 8.1, Wathen and Thall review a recently developed Bayesian sequential design and illustrate it with an application to a trial with non small cell lung cancer patients. In Section 8.2, Weiss and Wang develop Bayesian methodology to choose the sample size and experimental design for normal longitudinal models and apply the methodology to designing a follow-up longitudinal study with a predictive prior based on an earlier study. In Section 8.3, Müller, Sivaganesan, and Laud discuss a Bayesian approach to subgroup analysis. The underlying research challenge is the correct adjustment for multiplicities when considering results in a clinical study specific to many possible subpopulations.

Chapter 9 reports on the rapidly expanding area of Bayesian inference for high throughput genomic data and related challenges. In Section 9.1, Shen and West present a Bayesian approach for inference on multiple gene expression signatures. Their discussion includes a Monte Carlo variational method for estimating marginal likelihoods for model comparisons. In Section 9.2, Monni and Li present a Bayesian variable selection procedure when covariates are measured on a graph. The motivating application is the incorporation of known molecular pathways in inference for genomic data. One of the strengths of Bayesian analysis that gives rise to many research opportunities is the natural and principled approach to simultaneously modeling multiple related processes. In Section 9.3, Bloomquist and Suchard review how this feature is exploited in Bayesian inference for phylogenetics.

Bayesian data mining and machine learning involves many challenging research problems related to the exploration of redundant and high-dimensional data. This is the focus of Chapter 10. In Section 10.1, Mallick, Ray, and Dhavala address automatic model selection of the ideal number of principal components using exact inference with reversible jump MCMC. In Section 10.2, Balakrishnan and Madigan explore the use of proper priors for variance parameters of certain sparse Bayesian regression models that can be defined as generalizations of the popular Bayesian Lasso. In Section 10.3, Airoldi, Fienberg, Joutard, and Love utilize hierarchical Bayesian mixed-membership models and present several examples of model specifi-

cation and variations, both parametric and nonparametric, in the context of learning the number of latent groups and associated patterns for clustering units.

Chapter 11 presents current research challenges and applications of Bayesian analysis in political science, finance, and marketing research. In Section 11.1, Gelman discusses the role of prior constructions in political science applications. The importance of subjective priors in political science examples contrasts the semi-automatic approaches implemented in machine learning methods discussed in the previous chapter. In Section 11.2, Hore, Johannes, Lopes, McCulloch, and Polson describe computationally challenging inference problems in Finance. In Section 11.3, Jacquier and Polson consider a simulation-based approach to optimal portfolio selection. In Section 11.4, Fong reviews recent development of Bayesian multidimensional scaling models for multiple choice data in marketing research.

Chapter 12 discusses research challenges and recent development of Bayesian inference for binary and categorical data. In Section 12.1, Albert presents a Good (1967) Bayesian approach to inference for sparse contingency tables. In Section 12.2, Ghosh and Mukherjee consider the analysis of data with matched pairs, mainly in the context of case-control studies. In Section 12.3, Chen, Kim, Kuo, and Xie discuss critical issues involved in modeling binary response data and present a recently developed Stepping-Stone method for computing marginal likelihoods.

Chapter 13 focuses on Bayesian modeling and inference for spatial and/or time series data. In Section 13.1, Banerjee and Gelfand develop inference for spatial gradients. The method is illustrated through an analysis of urban land value gradients using a portion of Olcott's classic Chicago land value data. In Section 13.2, Wang, Dey, and Banerjee apply a new flexible skewed link function in the binomial model for the point-level spatial data based on the generalized extreme value distribution. In Section 13.3, De Oliveira provides a review of the main results obtained in the last decade on objective (default) Bayesian methods for the analysis of spatial data using Gaussian random fields.

A review of frontiers in Bayesian analysis would be incomplete without a discussion of ever-evolving computational methods, including posterior simulation, variable selection and model comparison. Chapter 14 discusses related research problems. In Section 14.1, Marin and Robert review alternative approaches to evaluate Bayes factors for model comparison, including brute-force evaluation of marginal probabilities, importance sampling methods, bridge sampling, and the use of candidates formula. In Section 14.2, Heaton and Scott discuss an important special case of model comparison. They summarize several recently proposed methods for variable selection in linear models. Heaton and Scott convincingly argue that this venerable topic is still and again a very active research area. Finally, in Section 14.3, Liechty, Liechty, and Müller focus on another specific research challenge related to posterior simulation. They consider problems that arise from the use of constrained parameter and sample spaces.

Chapter 2
Objective Bayesian Inference with Applications

It is natural to start a review of research frontiers in Bayesian analysis with a discussion of research challenges related to prior choices. In particular, in this chapter we discuss the definition of reference priors in some non-standard settings as well as the use of reference priors to define objective Bayesian testing.

2.1 Bayesian Reference Analysis of the Hardy-Weinberg Equilibrium

José M. Bernardo and Vera Tomazella

An important problem in genetics, testing whether or not a trinomial population is in Hardy-Weinberg equilibrium, is analyzed from an objective Bayesian perspective. The corresponding precise hypothesis testing problem is considered from a decision-theoretical viewpoint, where the null hypothesis is rejected if the null model is expected to be too far from the true model in the logarithmic divergence (Kullback-Leibler) sense. The quantity of interest in this problem is the divergence of the null model from the true model; as a consequence, the analysis is made using the reference prior for the trinomial model which corresponds to that divergence being the parameter of interest. The results are illustrated using examples both with simulated data and with data previously analyzed in the relevant literature.

M.-H. Chen et al. (eds.), *Frontiers of Statistical Decision Making and Bayesian Analysis*, DOI 10.1007/978-1-4419-6944-6_2,

2.1.1 Problem Statement

2.1.1.1 The Hardy-Weinberg (HW) Equilibrium in Genetics

At a single autosomal locus with two alleles, a diploid individual has three possible genotypes, typically denoted $\{AA, aa, Aa\}$, with (unknown) population frequencies $\{\alpha_1, \alpha_2, \alpha_3\}$, where $0 < \alpha_i < 1$ and $\sum_{i=1}^{3} \alpha_i = 1$.

The population is said to be in HW equilibrium if there exists a probability $p = P(A)$, $0 < p < 1$, such that $\{\alpha_1, \alpha_2, \alpha_3\} = \{p^2, (1-p)^2, 2p(1-p)\}$. To determine whether or not a population is in HW equilibrium, which is often the case when random mating takes place, is an important problem in biology.

Given a random sample of size n from the population, and observed $\{n_1, n_2, n_3\}$ individuals (with $n = n_1 + n_2 + n_3$) from each of the three possible genotypes $\{AA, aa, Aa\}$, the question is whether or not these data support the hypothesis of HW equilibrium.

This is an important example of *precise* hypothesis in the sciences, for HW equilibrium corresponds to a zero measure set within the original parameter space.

2.1.1.2 Statistical Formulation

Since $\sum_{i=1}^{3} \alpha_i = 1$, there are only two independent parameters. In terms of the population frequencies α_1 and α_2 of the two pure genotypes AA and aa, the relevant statistical model is the trinomial

$$\text{Tri}(n_1, n_2 | n, \alpha_1, \alpha_2) = \frac{n!}{n_1!\, n_2!\, (n - n_1 - n_2)!}\, \alpha_1^{n_1}\, \alpha_2^{n_2}\, (1 - \alpha_1 - \alpha_2)^{n - n_1 - n_2}$$

with $0 < \alpha_1 < 1$, $0 < \alpha_2 < 1$, and $0 < \alpha_1 + \alpha_2 < 1$ and, in conventional language, it is required to test the null hypothesis

$$H_0 = \{(\alpha_1, \alpha_2);\ \alpha_1 = p^2, \alpha_2 = (1-p)^2,\ 0 < p < 1\}.$$

This is the parametric form of the equation of the line $\sqrt{\alpha_1} + \sqrt{\alpha_2} = 1$, represented with a solid line in Figure 2.1, and it is a set of zero measure within the parameter space, the simplex $\mathscr{A} = \{(\alpha_1, \alpha_2);\ 0 < \alpha_1 < 1,\ 0 < \alpha_2 < 1,\ 0 < \alpha_1 + \alpha_2 < 1\}$.

Testing a trinomial population for HW equilibrium is a problem that has received a fair amount of attention in the statistical literature. Main pointers include the frequentist analysis of Haldane (1954), an "exact" test based on the distribution $p(n_1, n_2 | H_0, n_1 - n_2, n)$, and the Bayesian analysis of Lindley (1988) who reparametrizes to

$$\psi(\alpha_1, \alpha_2) = \frac{1}{2} \log \frac{4\, \alpha_1\, \alpha_2}{(1 - \alpha_1 - \alpha_2)^2},$$

so that $\psi = 0$ when H_0 is true, and then obtains approximations to the posterior density of ψ, $\pi(\psi | n_1, n_2, n_3)$ for a range of different prior choices.

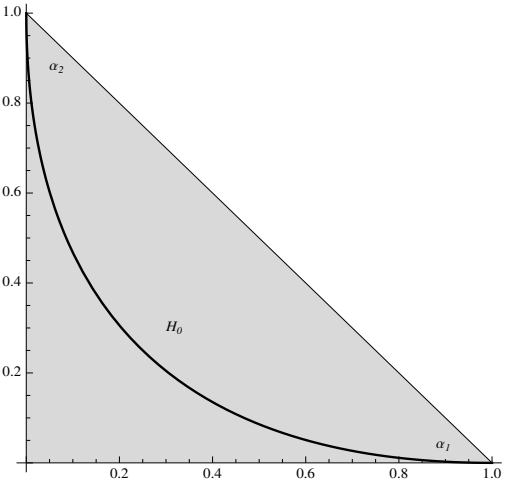


FIGURE 2.1. Precise null (solid line) within the parameter space (shaded region).

2.1.2 Objective Precise Bayesian Testing

2.1.2.1 The Decision Problem and the Intrinsic Loss Function

If data $\mathbf{z}$ are assumed to have been generated from the probability model $\mathscr{M} \equiv \{p_{\mathbf{z}}(\cdot|\phi,\omega),\ \mathbf{z}\in\mathscr{Z},\ \phi\in\Phi,\ \omega\in\Omega\}$, then testing whether of not the observed data $\mathbf{z}$ are compatible with the precise hypothesis $H_0 = \{\phi = \phi_0\}$ may be seen as a simple decision problem with only two alternatives:

1. a_0: To accept H_0, and work *as if* data were generated from the *reduced* model $\mathscr{M}_0 \equiv \{p_{\mathbf{z}}(\cdot|\phi_0,\omega),\ \mathbf{z}\in\mathscr{Z},\ \omega\in\Omega\}$; and
2. a_1: To reject H_0, and keep working with the *assumed* model $\mathscr{M}$.

Foundations then dictate (see, e.g., Bernardo and Smith, 1994, Chapter 2 and references therein) that one must

1. Specify a loss function $\ell\{a_i,(\phi,\omega)\}$, $i = 0,1$.
2. Specify a prior function $p(\phi,\omega)$, on $\Phi\times\Omega$, and use Bayes to obtain

$$p(\phi,\omega|\mathbf{z}) \propto p(\mathbf{z}|\phi,\omega)\, p(\phi,\omega).$$

3. Reject H_0 if, and only if, $l(a_0|\mathbf{z}) > l(a_1|\mathbf{z})$, where

$$l(a_i|\mathbf{z}) = \int_{\Phi}\int_{\Omega} \ell\{a_i,(\phi,\omega)\}p(\phi,\omega|\mathbf{z})d\phi d\omega.$$

One should then reject H_0 if, and only if, $l(a_0|\mathbf{z}) > l(a_1|\mathbf{z})$, hence if, and only if,

$$\int_{\Phi}\int_{\Omega} \left[\,\ell\{a_0,(\phi,\omega)\} - \ell\{a_1,(\phi,\omega)\}\,\right] p(\phi,\omega|\mathbf{z})d\phi d\omega > 0,$$

which only depends on the loss increase from rejecting H_0, given by

$$\Delta(\phi,\omega) = \ell\{a_0,(\phi,\omega)\} - \ell\{a_1,(\phi,\omega)\}.$$

Without loss of generality, the loss increase $\Delta(\phi,\omega)$ may be written in the form $\delta\{\phi_0,(\phi,\omega)\} - d_0$, where

1. $\delta\{\phi_0,(\phi,\omega)\}$ is the non-negative terminal loss to be suffered by accepting $\phi = \phi_0$ as a function of (ϕ,ω); and
2. d_0 is the strictly positive utility of accepting H_0 when it is true.

With this notation, one should reject the null if, and only if,

$$\int_{\Phi}\int_{\Omega} \delta\{\phi_0,(\phi,\omega)\}\, p(\phi,\omega|\mathbf{z})\, d\phi d\omega > d_0,$$

that is, if (and only if) the null model is expected to be too divergent from the true model.

For any one-to-one function $\psi = \psi(\phi)$ the conditions to reject $\phi = \phi_0$ should certainly be *precisely the same* as the conditions to reject $\psi = \psi(\phi_0)$ (a property unfortunately *not* satisfied by many published hypothesis testing procedures). This requires the use of an *invariant* loss function.

Model-based loss functions are loss functions defined in terms of the discrepancy measures between probability models. Within a family $\mathscr{F} \equiv \{p_{\mathbf{z}}(\cdot|\psi), \psi \in \Psi\}$, the loss suffered from using an estimate $\tilde{\psi}$ is of the form

$$\ell(\tilde{\psi},\psi) = \delta\{p_{\mathbf{z}}(\cdot|\tilde{\psi}), p_{\mathbf{z}}(\cdot|\psi)\},$$

defined in terms of the discrepancy of $p_{\mathbf{z}}(\cdot|\tilde{\psi})$ from $p_{\mathbf{z}}(\cdot|\psi)$, rather than on the discrepancy of $\tilde{\psi}$ from ψ. Model-based loss functions are obviously invariant under one-to-one reparametrization.

A model-based loss function with unique additive properties and built in calibration, is the *intrinsic loss function*, defined as the minimum expected log-likelihood ratio against the null:

$$\delta\{\phi_0,(\phi,\omega)\} = \inf_{\omega_0\in\Omega} \int_{\mathscr{Z}} p(\mathbf{z}|\phi,\omega) \log \frac{p(\mathbf{z}|\phi,\omega)}{p(\mathbf{z}|\phi_0,\omega_0)}\, d\mathbf{z}.$$

This may be also be described as the minimum (Kullback-Leibler) logarithmic divergence of $\mathscr{M}_0$ from the assumed model.

2.1.2.2 Reference Analysis and Precise Hypothesis Testing

Given a model $\mathscr{M} \equiv \{p_{\mathbf{z}}(\cdot|\phi,\omega),\ \mathbf{z}\in\mathscr{Z},\ \phi\in\Phi,\ \omega\in\Omega\}$, the θ-reference prior function $\pi_\theta(\phi,\omega)$ (see Bernardo, 2005, and references therein) is that which maximizes the missing information about $\theta = \theta(\phi,\omega)$. The corresponding marginal

reference posterior $\pi(\theta|\mathbf{z})$ summarizes inferential statements about a quantity of interest θ which only depend on the model assumed and the data obtained.

The *Bayesian Reference Criterion* (BRC) to test $H_0 \equiv \{\phi = \phi_0\}$ is the solution to the hypothesis testing decision problem corresponding to the *intrinsic loss* and the relevant *reference prior*. It only requires computing the *intrinsic test statistic*, defined as the reference posterior expectation,

$$d(H_0|\mathbf{z}) = \int_0^\infty \delta\, \pi(\delta|\mathbf{z}) d\delta,$$

of the intrinsic discrepancy loss $\delta(\phi,\omega) = \delta\{\phi_0, (\phi,\omega)\}$, which is in this case of the quantity of interest.

The intrinsic test statistic is a direct *measure of evidence against* H_0, in a *log-likelihood ratio scale*, which is independent of the sample size, the dimensionality of the problem, and the parametrization used. For further details and many examples, see Bernardo (2005) and references therein.

2.1.3 Testing for Hardy-Weinberg Equilibrium

2.1.3.1 The Quantity of Interest

Within the trinomial model,

$$\mathrm{Tri}\{n_1,n_2|n,\alpha_1,\alpha_2\} = \frac{n!}{n_1!\,n_2!\,(n-n_1-n_2)!}\,\alpha_1^{n_1}\alpha_2^{n_2}\,(1-\alpha_1-\alpha_2)^{n-n_1-n_2},$$

the logarithmic divergence of a member $\mathrm{Tri}\{n_1,n_2|n,p_0^2,(1-p_0)^2\}$ of the null

$$H_0 = \{(\alpha_1,\alpha_2);\ \alpha_1 = p^2, \alpha_2 = (1-p)^2, \quad 0 < p < 1\}$$

from the assumed model $\mathrm{Tri}\{n_1,n_2|n,\alpha_1,\alpha_2\}$ is

$$k\{p_0|\alpha_1,\alpha_2\} = E_{(n_1,n_2|\alpha_1,\alpha_2)}\left[\log\frac{\mathrm{Tri}\{n_1,n_2|n,\alpha_1,\alpha_2\}}{\mathrm{Tri}\{n_1,n_2|n,p_0^2,(1-p_0)^2\}}\right]$$

which, after some algebra, reduces to

$$n[(\alpha_2-\alpha_1-1)\log(p_0)+(\alpha_1-\alpha_2-1)\log(1-p_0)-(1-\alpha_1-\alpha_2)\log(2)-\mathrm{H}\{\alpha\}],$$

where $\mathrm{H}\{\alpha\} = -\alpha_1\log\alpha_1 - \alpha_2\log\alpha_2 - (1-\alpha_1-\alpha_2)\log(1-\alpha_1-\alpha_2)$ is the entropy of $\alpha = \{\alpha_1,\alpha_2,1-\alpha_1-\alpha_2\}$. The last expression is minimized, for $0 < p_0 < 1$, when $p_0 = (1+\alpha_1-\alpha_2)/2$, and substitution yields the intrinsic loss function,

$$\delta\{H_0,(\alpha_1,\alpha_2)\} = \inf_{0<p_0<1} k\{p_0|\alpha_1,\alpha_2\} = n\,\theta(\alpha_1,\alpha_2),$$

where

$$\theta(\alpha_1,\alpha_2) = 2\,\mathrm{H}\{\omega, 1-\omega\} - \mathrm{H}\{\alpha_1,\alpha_2,1-\alpha_1-\alpha_2\} - (1-\alpha_1-\alpha_2)\log(2),$$

and $\omega = \omega(\alpha_1,\alpha_2) = (1+\alpha_1-\alpha_2)/2$ is the value of p for a trinomial population $\mathrm{Tri}\{n_1,n_2|n,p^2,(1-p)^2\}$ in HW equilibrium which is closest, in the logarithmic divergence sense, to the trinomial population $\mathrm{Tri}\{n_1,n_2|n,\alpha_1,\alpha_2\}$. The function $\delta\{H_0,(\alpha_1,\alpha_2)\}$ measures the discrepancy of the null from the trinomial model $\mathrm{Tri}\{\cdot|n,\alpha_1,\alpha_2\}$.

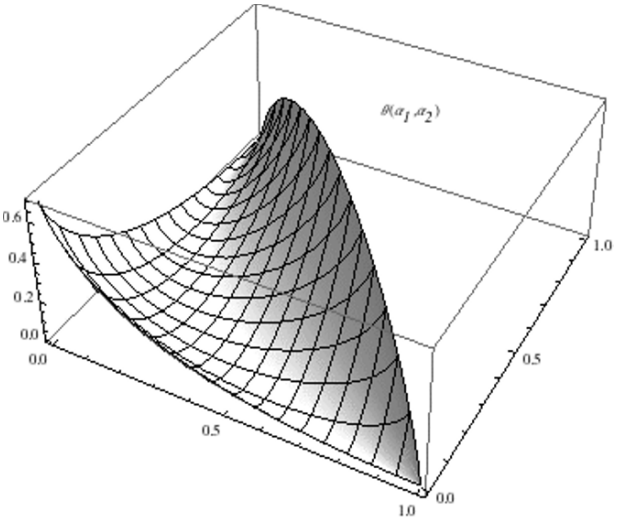


FIGURE 2.2. The quantity of interest, $\theta = \theta(\alpha_1,\alpha_2)$.

The quantity of interest in this problem is clearly the function $\theta = \theta(\alpha_1,\alpha_2)$ since $\delta\{H_0,(\alpha_1,\alpha_2)\} = n\,\theta(\alpha_1,\alpha_2)$ precisely measures how far the null H_0 is from the assumed model. In particular, the population is in HW equilibrium if, and only if, $\theta = 0$, in which case, $\sqrt{\alpha_1}+\sqrt{\alpha_2} = 1$ or $\alpha_2 = (1-\sqrt{\alpha_1})^2$. Figure 2.2 provides a 3D plot of the surface $\theta(\alpha_1,\alpha_2)$. It is zero for all HW equilibrium values and achieves its maximum value, $\log(2)$, at both $(0,0)$ and $(1/2,1/2)$. Hence, in this problem, the intrinsic loss is a bounded function.

2.1.3.2 The Reference Prior

To obtain the joint reference prior $\pi_\theta(\alpha_1,\alpha_2)$ when $\theta = \theta(\alpha_1,\alpha_2)$ is the quantity of interest, a complementary parameter $\omega = \omega(\alpha_1,\alpha_2)$ must be chosen, so that (θ,ω) is a one-to-one transformation of (α_1,α_2). A convenient choice is the function $\omega(\alpha_1,\alpha_2) = (1+\alpha_1-\alpha_2)/2$, which occurs in the expression of $\delta\{H_0,(\alpha_1,\alpha_2)\}$ obtained above. The reference prior in this parametrization when θ is the param-

eter of interest is then obtained as $\pi_\theta(\theta,\omega) = \pi(\omega|\theta)\,\pi(\theta)$. Finally, the required reference prior in the original parametrization is obtained as

$$\pi_\theta(\alpha_1,\alpha_2) = |\mathrm{J}(\alpha_1,\alpha_2)|\,\pi_\theta(\theta(\alpha_1,\alpha_2),\omega(\alpha_1,\alpha_2)),$$

where $\mathrm{J}(\alpha_1,\alpha_2) = \left(\frac{\partial\theta}{\partial\alpha_1}\,\frac{\partial\omega}{\partial\alpha_2}\right)$ is the corresponding Jacobian matrix.

The required transformation, represented in Figure 2.6, is delicate. Indeed, the Jacobian determinant $|\mathrm{J}(\alpha_1,\alpha_2)| = \log(1-\alpha_1-\alpha_2) - \frac{1}{2}\log(4\alpha_1\,\alpha_2)$ is null at the HW line, positive below, negative above, and diverges at the simplex borders. A one-to-one transformation is only obtained in each of the two separate regions defined by the equilibrium line. Thus a one-to-one transformation is $\{\alpha_1,\alpha_2\} \iff \{\theta,\omega,\lambda\}$ where $\lambda \in \{1,2\}$ indicates region, with $\lambda = 1$ when $\sqrt{\alpha_1}+\sqrt{\alpha_2} < 1$, and $\lambda = 2$ when $\sqrt{\alpha_1}+\sqrt{\alpha_2} > 1$. Formally,

$$\pi_\theta(\alpha_1,\alpha_2) = \pi_\theta(\alpha_1,\alpha_2|\lambda=1) + \pi_\theta(\alpha_1,\alpha_2|\lambda=2).$$

The joint reference priors in each of the two regions must be be separately computed.

This model is regular. Hence, the reference prior $\pi(\omega|\theta)\,\pi(\theta)$ may be found in terms of the relevant Fisher information matrix. In the original parametrization, the inverse of Fisher matrix F_1 is

$$\mathrm{F}_1^{-1}(\alpha_1,\alpha_2) = \begin{pmatrix} \alpha_1(1-\alpha_1) & -\alpha_1\,\alpha_2 \\ -\alpha_1\,\alpha_2 & \alpha_2(1-\alpha_2) \end{pmatrix},$$

so that, in the new parametrization, Fisher matrix is F_2 such that

$$\mathrm{F}_2^{-1}(\theta,\omega) = \mathrm{J}(\alpha_1,\alpha_2)\cdot\mathrm{F}_1^{-1}(\alpha_1,\alpha_2)\cdot\mathrm{J}^t(\alpha_1,\alpha_2),$$

evaluated with the inverse functions $\alpha_1(\theta,\omega)$ and $\alpha_2(\theta,\omega)$. Fisher matrix F_2 has a complex, but analytical expression, in terms of α_1 and α_2, but the inverse functions $\alpha_i(\theta,\omega)$ must be numerically computed.

The reference prior $\pi(\omega|\theta)\,\pi(\theta)$ may be found in terms of $\mathrm{H} = \mathrm{F}_2$ and $\mathrm{V} = \mathrm{F}_2^{-1}$ (Berger and Bernardo, 1992a), from

$$\pi(\omega|\theta) \propto h_{22}^{1/2}(\theta,\omega)$$

and

$$\pi(\theta) \propto \exp\left[\int_{\Omega(\theta)} \pi(\omega|\theta)\log\{v_{11}^{-1/2}(\theta,\omega)\}\,d\omega\right].$$

Lower region: $R_1 = \{(\alpha_1,\alpha_1);\ \sqrt{\alpha_1}+\sqrt{\alpha_2} \le 1\}$**.** The reference conditional priors are numerically found to be approximate the Beta densities (see Figure 2.3)

$$\pi_1(\omega|\theta) \approx \frac{1}{\omega_1(\theta)-\omega_0(\theta)}\,\mathrm{Be}\left(\frac{\omega-\omega_0(\theta)}{\omega_1(\theta)-\omega_0(\theta)}\middle|\frac{1}{2},\frac{1}{2}\right), \quad \omega_0(\theta) < \omega < \omega_1(\theta),$$

where $\omega_0(\theta)$ and $\omega_1(\theta)$ are respectively the inverse functions of

$$\theta_1(\omega) = (2\omega - 1)\log(2\omega - 1) - 2\omega\log(\omega), \quad 1/2 < \omega < 1,$$

$$\theta_0(\omega) = (1 - 2\omega)\log(1 - 2\omega) - 2(1 - \omega)\log(1 - \omega), \quad 0 < \omega < 1/2.$$

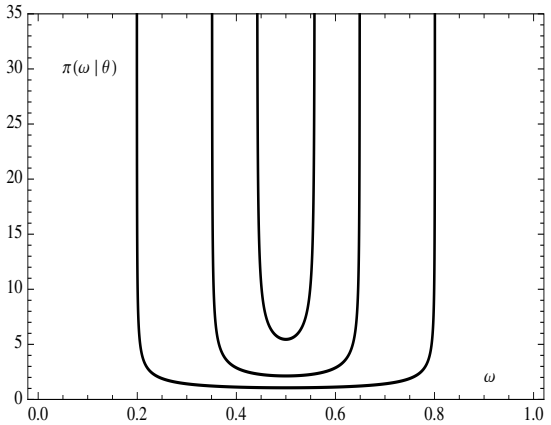


FIGURE 2.3. Conditional reference priors of $\omega \in (\omega_0(\theta), \omega_1(\theta))$, in the lower region of the parameter space, for $\theta = 0.05, 0.20$ and 0.40.

Using the analytical approximation for the conditional reference priors, the marginal reference prior for the quantity of interest results

$$\pi_1(\theta) \approx \frac{1}{\log(2)} \, \text{Be}\left(\frac{\theta}{\log(2)} \middle| \frac{1}{2}, \frac{1}{2}\right), \quad 0 < \theta < \log(2).$$

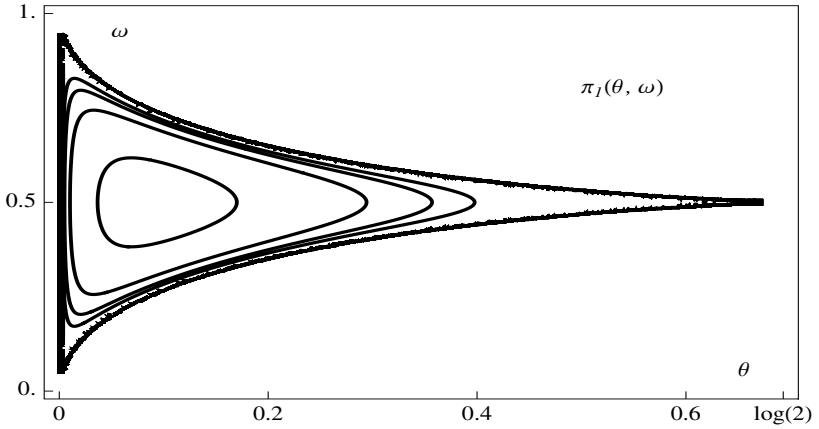


FIGURE 2.4. Contour plot of the joint reference prior $\pi_1(\theta, \omega)$ in the lower region.

The joint reference prior for this region is then $\pi_1(\theta,\omega) = \pi_1(\omega|\theta)\,\pi_1(\theta)$. The contour plot of this joint refernce prior is shown in Figure 2.4. Notice that these reference priors are all proper.

Upper region: $R_2 = \{(\alpha_1,\alpha_1);\ \sqrt{\alpha_1}+\sqrt{\alpha_2} \leq 1\}$**.** Similarly, in the region over the HW equilibrium line, the reference conditional priors are numerically found to be

$$\pi_2(\omega|\theta) \approx \frac{1}{\omega_1(\theta)-\omega_0(\theta)}\ \mathrm{Be}\left(\frac{\omega-\omega_0(\theta)}{\omega_1(\theta)-\omega_0(\theta)}\middle|\frac{1}{2},\frac{1}{2}\right), \quad \omega_0(\theta) < \omega < \omega_1(\theta),$$

where $\omega_1(\theta)$ and $\omega_0(\theta)$ are respectively the inverse functions of

$$\theta_1(\omega) = -\omega\log(\omega) - (1-\omega)\log(1-\omega), \quad 1/2 < \omega < 1$$

$$\theta_0(\omega) = -\omega\log(\omega) - (1-\omega)\log(1-\omega), \quad 0 < \omega < 1/2.$$

The marginal reference prior for θ in the upper region is

$$\pi_2(\theta) \approx \frac{1}{\log(2)}\ \mathrm{Be}\Big(\frac{\theta}{\log(2)}\Big|\frac{1}{2},\frac{1}{2}\Big), \quad 0 < \theta < \log(2).$$

The joint reference prior for the upper region is then $\pi_2(\theta,\omega) = \pi_2(\omega|\theta)\,\pi_2(\theta)$. Again, all these reference priors are all proper.

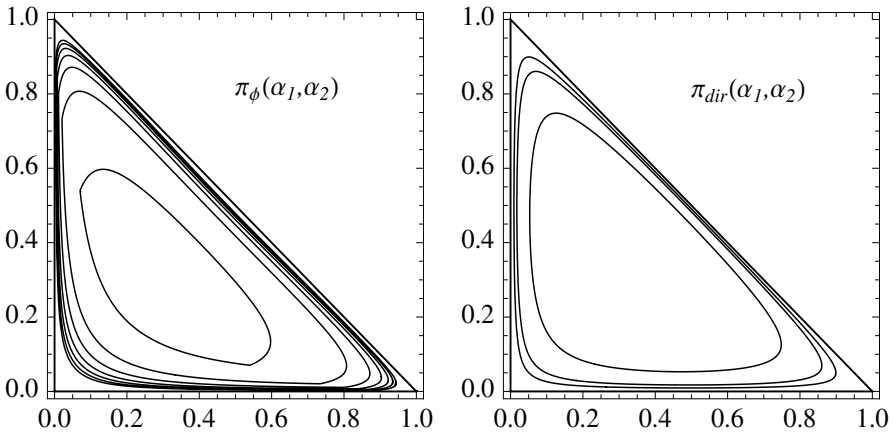


FIGURE 2.5. Contour plots of the joint reference prior in the original parametrization and a Dirichlet density with parameter $(1/3,1/3,1/3)$.

Joint reference prior in the original parametrization. Returning to the original parametrization and combining the results from the two regions produces $\pi_\theta(\alpha_1,\alpha_2)$, whose contour plot is represented in the left panel of Figure 2.5. For comparison, the right panel represents the contour plot of a Dirichlet density with

parameter vector $(1/3, 1/3, 1/3)$. This could be used as an approximation if exact computation is not needed.

2.1.3.3 Posterior Inference: Estimation and Testing

Joint reference posterior. For any data set, $\{n_1, n_2, n_3\}$, where n_1 and n_2 are respectively the number of observed pure genotypes AA and aa, and n_3 is the number of observed mixed genotypes Aa, the joint reference posterior is

$$\pi_\theta(\alpha_1, \alpha_1 | n_1, n_2, n_3) = c^{-1}(n_1, n_2 | n)\, \mathrm{Tri}\{n_1, n_2 | n, \alpha_1, \alpha_2\}\, \pi_\theta(\alpha_1, \alpha_2),$$

where $n = n_1 + n_2 + n_3$ and

$$c(n_1, n_2 | n) = \int_0^1 \left\{ \int_0^{1-\alpha_1} \mathrm{Tri}\{n_1, n_2 | n, \alpha_1, \alpha_2\}\, \pi_\theta(\alpha_1, \alpha_2)\, d\alpha_2 \right\} d\alpha_1,$$

a delicate numerical integral given the prior shape.

The posterior probabilities of the two non-equilibrium regions are

$$P[R_1 | n_1, n_2, n_3] = \int_0^1 \left\{ \int_0^{(1-\sqrt{\alpha_1})^2} \pi_\theta(\alpha_1, \alpha_1 | n_1, n_2, n_3)\, d\alpha_2 \right\} d\alpha_1,$$

and $P[R_2 | n_1, n_2, n_3] = 1 - P[R_1 | n_1, n_2, n_3]$.

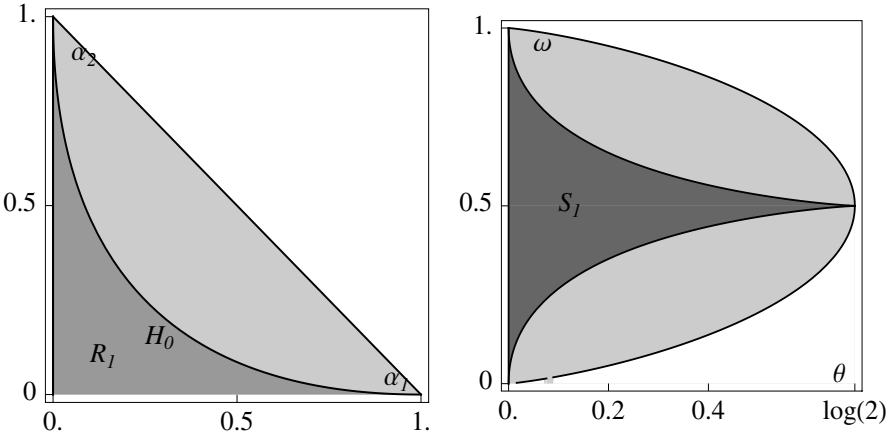


FIGURE 2.6. Original and transformed parameter spaces.

Since the transformation between (α_1, α_2) and (θ, ω) is not one-to-one, computing the joint posterior density in terms of the (θ, ω) requires identification of the two possible inverse values $\alpha_1(\theta, \omega)$ and $\alpha_2(\theta, \omega)$. This is done in terms of

$S_1 = \text{Image}(R_1)$, where R_1 is the region below H_0, and $S_2 = \text{Image}(R_2)$, where R_2 is the region above H_0. Thus, if $(\theta, \omega) \in S_1$, which is contained in S_2), then there are two diffferent pairs of (α_1, α_2) values which map into (θ, ω) (see Figure 2.6).

It follows that, for any data $\mathbf{z} = \{n_1, n_2, n_3\}$,

$$\pi(\theta, \omega|\mathbf{z}) = \pi(\theta, \omega|\mathbf{z}, S_1)P(R_1|\mathbf{z}) + \pi(\theta, \omega|\mathbf{z}, S_2)P(R_2|\mathbf{z})$$

$$\pi(\theta, \omega|\mathbf{z}, S_i) = \frac{\pi(\alpha_1, \alpha_2|\mathbf{z}, R_i)}{|\mathbf{J}(\alpha_1, \alpha_2)|}, \quad \alpha_j \to \alpha_{ji}(\theta, \omega), \quad i = 1, 2,$$

where $\{\alpha_{1i}, \alpha_{2i}\}$ is the inverse function mapping S_i into R_i.

The required marginal reference posterior for the quantity of interest θ is then

$$\pi(\theta|\mathbf{z}) = \int_{\Omega(\theta)} \pi(\theta, \omega|\mathbf{z})\, d\omega.$$

This will concentrate on its extreme value $\theta = 0$ if, and only if, the population is in approximate HW equilibrium.

Intrinsic test statistic. As described in Section 2.1.2, the intrinsic test statistic $d(H_0|\mathbf{z})$ is the reference posterior expectation of $\delta\{H_0, (\alpha_1, \alpha_2)\}$, defined as the minimum logarithmic divergence of the null model from the true model. Since $\delta\{H_0, (\alpha_1, \alpha_2)\} = n\,\theta(\alpha_1, \alpha_2)$, the intrinsic statistic is simply

$$d(H_0|\mathbf{z}) = n \int_0^{\log(2)} \theta\, \pi(\theta|\mathbf{z})\, d\theta = n\, E[\theta|\mathbf{z}],$$

the reference posterior expectation of the quantity of interest times the sample size. This is precisely the reference posterior expectation of the log-likelihood ratio against the null and, therefore, $d(H_0|\mathbf{z})$ has an immediate meaning as an objective measure of the evidence against the null provided by the data.

2.1.4 Examples

2.1.4.1 Simulations

Data simulated under HW equilibrium. A trinomial sample of size $n = 30$ from a population in HW equilibrium was simulated with $P[A] = p = 0.3$, so that $\{\alpha_1, \alpha_2\} = \{p^2, (1-p)^2\} = \{0.09, 0.49\}$, $\omega = p = 0.3$, and $\theta = 0$. The simulation yielded $\{n_1, n_2, n_3\} = \{2, 15, 13\}$.

Figure 2.7 represents the marginal reference posterior of $\delta = n\,\theta$ which, as expected, concentrates around the null value $\delta = 0$, with $d(H_0|\mathbf{z}) = n$, $E[\theta|\mathbf{z}] = 0.321 = \log(1.38)$, so that the likelihood ratio against the null is expected to be only about 1.38, and the null is accepted: one may safely proceed as if the population where in HW equilibrium, suggesting random mating.

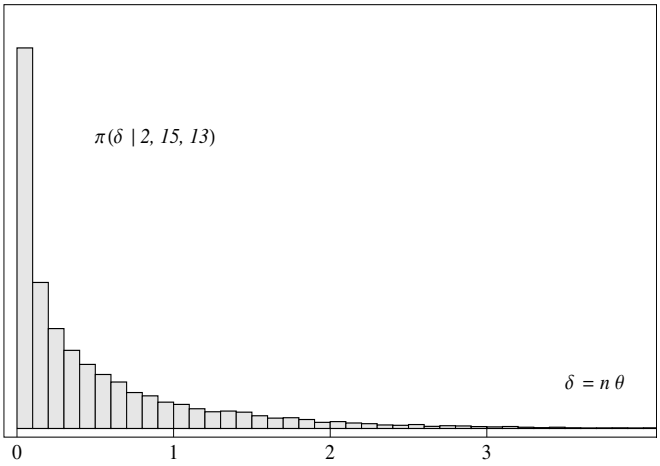


FIGURE 2.7. Reference posterior distribution of $\delta = n\theta$ with data simulated from a population in HW equilibrium.

Data simulated under non-HW equilibrium. A trinomial sample of size $n = 30$ was simulated with $\{\alpha_1, \alpha_2\} = \{0.45, 0.40\}$, so that $\omega = 0.525$, $\theta = 0.269$, and the population is *not* in HW equilibrium. The simulation then yielded $\{n_1, n_2, n_3\} = \{12, 12, 6\}$.

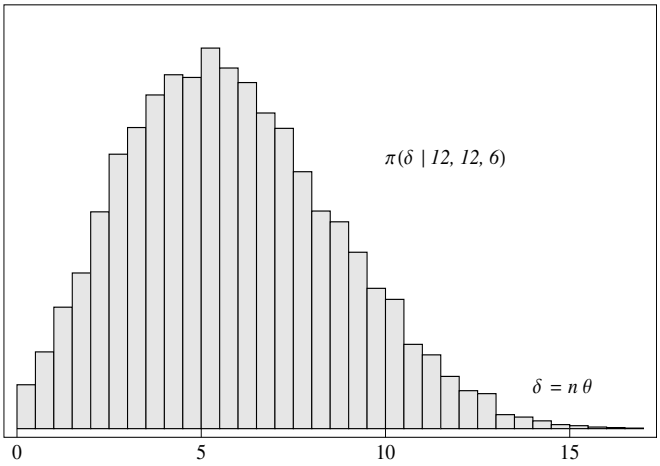


FIGURE 2.8. Reference posterior distribution of $\delta = n\theta$ with data simulated from a population not in HW equilibrium.

As Figure 2.8 illustrates, the marginal reference posterior of $\delta = n\theta$ has an interior mode, $d(H_0|\mathbf{z}) = n$, and $E[\theta|\mathbf{z}] = 5.84 \approx \log(344)$, so that the likelihood ratio

against the null is expected to be about 344. Thus, the null should certainly be *rejected*, and one should work under the assumption that the population is *not* in HW equilibrium, thus suggesting non random mating.

2.1.4.2 An Example from the Literature

Lindley data. Lindley (1988) analyzed the data $\mathbf{z} = \{0, 90, 10\}$ from a Bayesian viewpoint, noting that asymptotic results are scarcely satisfactory in this case, and performing an analysis of the clear dependence of the results on the prior chosen. This could be expected, for these data are somewhat extreme due to the fact that there are no observations from the pure *AA* genotype. Conclusions from extreme data are often very sensitive to the prior, and they cannot be usually be well approximated with asymptotic arguments.

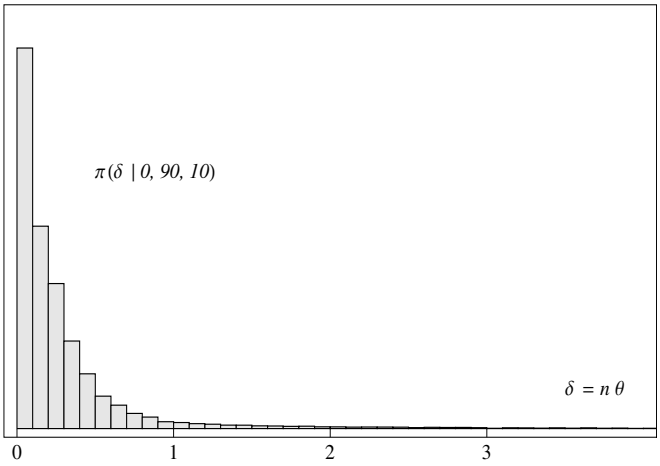


FIGURE 2.9. Marginal reference posterior distribution of $\delta = n\,\theta$ for Lindley (1988) data.

Reference analysis has been known to perform fine in many other problems with extreme data. This provides yet another example. As Figure 2.9 illustrates, it is found that $\pi(\delta|\mathbf{z})$, the reference posterior density of the expected discrepancy from the null is again very concentrated around the null value $\delta = 0$. Indeed, the test statistic is $d(H_0|\mathbf{z}) = nE[\theta|\mathbf{z}] = 0.51 = \log(1.66)$ and, hence, the likelihood ratio against the null may the expected to be just about 1.66.

We must therefore conclude that the HW equilibrium hypothesis is compatible with these data.

2.2 Approximate Reference Priors in the Presence of Latent Structure

Brunero Liseo, Andrea Tancredi, and Maria M. Barbieri

Objective priors play a crucial role both in classical and Bayesian inference. For a subjective Bayesian they represent, at the very least, the formalization of the vague idea of absence of prior information and they are a particularly useful tool in applications; from a frequentist perspective, objective priors can be seen as a black box to produce statistical procedures with excellent repeated sampling properties; for a remarkable example of this phenomenon see Berger and Sun (2008). The Jeffreys' and the reference prior approaches are nowadays the most popular in applications and both of them rely upon the explicit calculation of the expected Fisher information matrix $\mathbf{I}(\theta)$ (but see Berger, Bernardo, and Sun (2009) for a different and more general route to derive reference priors).

It may happen that the direct calculation of $\mathbf{I}(\theta)$ is too difficult or sometimes impossible to obtain in closed form. In some of these cases one can try to express the model into a more general way, based on the introduction of a vector of latent quantities, say $\mathbf{z}$. The idea of model completion has been already well exploited in Bayesian computation (Robert, 2001), since it is at the core of the data augmentation strategies (Tanner and Wong, 1987) or for extensions of the use of Gibbs sampling in non-conjugate settings (Damien, Wakefield, and Walker, 1999).

Suppose that the generic sampling distribution of our model, say $p(\cdot;\theta)$, can be written as

$$p(\cdot;\theta) = \int_{\mathbf{Z}} g(\cdot;\mathbf{z},\theta)h(\mathbf{z};\theta)d\mathbf{z}, \tag{2.2.1}$$

where $\mathbf{z}$ is a suitable vector of latent or auxiliary variables. One can pretend, then, that the working model is actually

$$g(\cdot;\mathbf{z},\theta),$$

where θ is the vector of unknown parameters and $\mathbf{z}$ can be considered either an additional parameter to be estimated or an additional vector of observations. In both cases the extended Fisher information matrix can be calculated for the new model; in many practical cases (see Section 2.2.2 below) it turns out that the extended Fisher matrix will have a simpler expression.

Sun and Berger (1998) propose a method of deriving reference prior in the presence of partial information. In particular they consider two cases of interest for our development: assume that the parameter vector θ can be split into (θ_1, θ_2), for inferential reasons. Then one can be interested in deriving

(A) the reference conditional prior for θ_1 given θ_2 when the marginal prior for θ_2 is available from subjective information sources.

(B) the reference marginal prior for θ_1 given that the conditional prior for θ_2 given θ_1 is available from subjective information sources.

These contexts resemble in several aspects our enlarged model after we pretend that the latent structure $\mathbf{z}$ is a parameter of the model. As we see in the next section, we consider separately the case where $h(\mathbf{z};\theta)$ actually depends on θ from the case where $h(\mathbf{z};\theta)$ does not really depend on θ. Section 2.2.1 is devoted to the illustration of our approach. In Section 2.2.2, we illustrate the method via several examples. In particular we derive the reference prior for the Warner model (Warner, 1965) for dealing with sensible questions, the general scale mixture model, and the probit model, for which no previous formal objective priors have been proposed. We also discuss, for pedagogical reasons, an example where our method cannot be used, namely the Student t model with an unknown number of degrees of freedom. In Section 2.2.3, we discuss a possible extension to situations where nuisance parameters are present.

The reference priors which we obtain in this section need not be equal to the "orthodox" ones. This happens because we are actually considering a "different model." However, in terms of frequentist coverage, they always show a similar and reasonable behavior. In some specific examples, such as the Warner model, our prior seems more natural than the "orthodox" Jeffreys prior; see the next section for details. Also, in complicated models, where there is no closed form reference prior, this method, when applicable, provides easy-to-handle solutions.

2.2.1 The Method

First we briefly recall, in our notation, the approach proposed in Sun and Berger (1998, Sections 2.2 and 2.3). First consider case A: suppose one knows the marginal prior for $\mathbf{z}$, say $\pi^s(\mathbf{z})$ and wants to derive the reference conditional prior $\pi^{(R)}(\theta \mid \mathbf{z})$. Following the reference prior approach, one must find that conditional prior that asymptotically maximizes the expected value of the Kullback-Leibler divergence between the conditional posterior density of θ, given $\mathbf{z}$ and the data, say $\mathbf{X}_n$, and the conditional prior of θ, given $\mathbf{z}$.

From an asymptotic expansion of the distance and using a result in Ghosh and Mukerjee (1992), one can argue that the maximizing conditional prior is

$$\pi^{(R)}(\theta \mid \mathbf{z}) \propto \mid I_{\theta,\theta} \mid^{1/2}. \tag{2.2.2}$$

where $I_{\theta,\theta} = I_{\theta,\theta}(\theta,\mathbf{z})$ is the Fisher information for θ, given that $\mathbf{z}$ is held fix, and $\mid I_{\theta,\theta} \mid$ denote the determinant of $I_{\theta,\theta}$.

If the above conditional prior is proper, then (2.2.2) is the conditional reference prior. Otherwise one should consider normalization concerns, following the approach in Sun and Berger (1998) or Berger, Liseo, and Wolpert (1999). Choose a nested sequence $\lambda_1 \subset \lambda_2 \subset \ldots$ of compact subsets of the enlarged parameter space $\Theta \times \mathbf{Z}$ such that $\cup_i \lambda_i = \Theta \times \mathbf{Z}$. For each i, define $G_i(\theta : (\theta,\mathbf{z}) \in \lambda_i)$. Also, let

$$K_i(\mathbf{z}) = \int_{G_i} \mid I_{\theta,\theta}(\theta,\mathbf{z}) \mid^{1/2} d\theta.$$

Sun and Berger (1998) show that the conditional reference prior is given by

$$\pi^{(R)}(\theta \mid \mathbf{z}) \propto \lim_{i \to \infty} \frac{K_i(\mathbf{z}^*)}{K_i(\mathbf{z})} \mid I_{\theta,\theta}(\theta, \mathbf{z}) \mid^{1/2},$$

where $\mathbf{z}^*$ is any fixed value in the interior of $\mathbf{Z}$. It is also easy to see that, when $I_{\theta,\theta}(\theta, \mathbf{z})$ factorizes into the product of two functions, one depending on θ and the other depending on $\mathbf{z}$, and the sequence of compact sets λ_i are chosen to be the Cartesian product of intervals in θ and $\mathbf{z}$ dimensions, then

$$\pi^{(R)}(\theta \mid \mathbf{z}) \propto \mid I_{\theta,\theta}(\theta, \mathbf{z}) \mid^{1/2} .$$

In this procedure the subjective input is concentrated on the choice of the sequence of the λ_i's: however, in many practical examples, there will be a natural, objective way of choosing them.

In case B, one needs to derive the marginal prior of θ in the presence of "conditional" prior information about $(\mathbf{z} \mid \theta)$. Following Sun and Berger (1998), the marginal reference prior is then

$$\pi^{(R)}(\theta) \propto \eta(\theta),$$

where

$$\eta(\theta) = \exp\left(\frac{1}{2} \int p(\mathbf{z} \mid \theta) \log\left(\frac{\mid I \mid}{\mid I_{\mathbf{zz}} \mid}\right) d\mathbf{z}\right), \tag{2.2.3}$$

and $I_{\mathbf{zz}}$ represents the lower right corner block of the Fisher matrix. In the special case that the quantity $\mid I \mid / \mid I_{\mathbf{zz}} \mid$ does not depend on $\mathbf{z}$, then

$$\pi^{(R)}(\theta) \propto (\mid I \mid / \mid I_{\mathbf{zz}} \mid)^{1/2}. \tag{2.2.4}$$

Note that, in this case, since the distribution of $\mathbf{z}$ depends on θ, it will be necessary to include this part of the model into the likelihood.

The transposition of the Sun and Berger approach into our framework would not be complete before considering the problem of the sample size. In fact, the completion formula (2.2.1) is almost always used at unit level: this implies that, for a sample of n i.i.d. observations, the vector length of latent variables $\mathbf{z}$ would depend on the sample size. Thus the conditional or the marginal reference prior for θ would depend on the sample size. However, in practical uses of the method, we will see that the dependence is rather weak. Alternatively one can argue that the Sun and Berger method is essentially asymptotic; therefore one can directly compute the conditional prior for θ by letting $n \to \infty$. We will discuss these alternative routes in the Examples section. Notice that all the examples related to case B which are discussed in this section share the common fact that the explicit introduction of the latent vector $\mathbf{z}$ into the likelihood makes the observed vector $\mathbf{y}$ conditionally independent of θ. However, more complex situations can be considered.

2.2.2 Examples

2.2.2.1 Sensible Answers

In order to protect confidentiality, questionnaires which contain sensitive questions may use some form of counfoundness: the following is a very simple example (Warner, 1965). Suppose we are interested in the fraction of regular drug users in a given population. In order to avoid non-responses, one can ask the person to be interviewed to toss a coin: if the coin gives H, he/she should answer the real question: "Do you make regular use of drugs?"; if the coin gives T, then he/she should answer to a more innocuous question (e.g., "Are you born in the first semester of the year?"). The interviewer, of course, does not know which question has been answered. The statistical model is then easily set up. Let Y_j, $j = 1, \ldots, n$ be the binary answer of the jth unit. Let θ be the true fraction of drug users in the population and let p be the fraction in the population which would answer YES to the innocuous question; p is assumed to be known. Then,

$$\Pr(Y_i = y; \theta, p) = \left(\frac{\theta + p}{2}\right)^y \left(1 - \frac{\theta + p}{2}\right)^{1-y}, \qquad y = 0, 1; \quad i = 1, \ldots, n.$$

Straightforward calculations lead to the Fisher information quantity for θ which is

$$I(\theta) = ((\theta + p)(2 - \theta - p))^{-1}.$$

The Jeffreys prior for the model is then proportional to

$$\pi^{(J)}(\theta) \propto ((\theta + p)(2 - \theta - p))^{-1/2}. \tag{2.2.5}$$

This prior clearly depends on p and it is symmetric only when $p = 0.5$. Note that p is supposed to be a known quantity and it represents the probability of answering YES to a completely instrumental question, which is not the object of our interest. That the prior distribution of θ would depend on p is obviously disturbing.

Consider now our alternative approach and, for each observation Y_i, introduce a Bernoulli latent variable Z_i representing the result of the coin toss for individual i, $i = 1, \ldots, n$. Also, treat the vector $\mathbf{z} = (Z_1, \ldots, Z_n)$ as an unknown parameter. Then, for each observation,

$$\Pr(Y_i = y \mid Z_i, \theta) = \left[\theta^y (1 - \theta)^{1-y}\right]^{Z_i} \left[p^y (1 - p)^{1-y}\right]^{1-Z_i}. \tag{2.2.6}$$

This is an example where we know the marginal distribution of the latent variables $\mathbf{z}$ and we look for the conditional reference prior for $\theta \mid \mathbf{z}$. From Section 2.2.1 we know that this is proportional to the square root of $I_{\theta\theta}$, the diagonal element of the Fisher matrix corresponding to θ. Now we show that $I_{\theta\theta} = n_z (\theta(1 - \theta))^{-1}$, where n_z is the number of Z_i equals to 1. With the augmented model (2.2.6), the log-likelihood function is

$$\ell(\theta, \mathbf{z}) \propto \sum_i y_i z_i \log\theta + \sum_i z_i \log(1-\theta) - \sum_i y_i z_i \log(1-\theta);$$

then

$$-\frac{\partial \ell(\theta, \mathbf{z})}{\partial \theta^2} = \frac{\sum_i y_i z_i}{\theta^2} + \frac{\sum_i (1-y_i) z_i}{(1-\theta)^2}.$$

Since $\mathbf{E}(Y_i \mid Z_i = z) = \theta^z p^{1-z}$, then

$$\mathbf{E}\left(\sum_i Y_i \mid \mathbf{z}\right) = \sum_i z_i \theta^{z_i} p^{1-z_i} = n_z \theta.$$

Finally,

$$I_{\theta,\theta} = \frac{n_z}{\theta} + \frac{n_z(1-\theta)}{(1-\theta)^2} = n_z\left(\frac{1}{\theta} + \frac{1}{1-\theta}\right).$$

This means that the conditional reference prior for θ given $\mathbf{z}$ is

$$\pi^{(R)}(\theta \mid \mathbf{z}) \propto \frac{1}{\sqrt{\theta(1-\theta)}} \tag{2.2.7}$$

Since the marginal distribution of the vector $\mathbf{z}$ does not depend on θ, (2.2.7) is also the marginal reference prior for the actual parameter of interest. The prior (2.2.7) is then the usual Jeffreys or reference prior for the binomial model and, contrary to (2.2.5), it does not depend on p. This is a simple example where the new approach does not provide the same result as the orthodox Jeffreys prior. We have compared priors (2.2.7) and (2.2.5) in terms of frequentist coverage of one tail credible intervals. In Tables 2.1 and 2.2, for three different values of n, we report the frequentist coverage at level 0.95 for the two priors, for different values of the parameters θ and p. Simulations are based on 10000 samples for each combination of true parameter values and sample sizes. The performances are definitely comparable.

TABLE 2.1. Frequentist coverage at level 0.95 of $\pi^{(R)}$ for different values of θ and p.

	$n=10$			$n=30$			$n=100$		
	0.10	0.50	0.90	0.10	0.50	0.90	0.10	0.50	0.90
$p=0.1$	1	1	0.97	1	0.97	0.98	0.94	0.95	0.97
$p=.25$	1	1	0.94	1	0.92	0.98	0.97	0.95	0.98
$p=5$	1	1	0.94	1	0.95	0.96	0.98	0.95	0.97
$p=0.75$	1	1	0.96	1	0.94	0.97	0.98	0.95	0.96
$p=0.9$	1	1	0.95	1	0.92	0.97	0.98	0.94	0.96

Beyond the simplicity of this example, we must notice that there exist several generalizations of this model for which the "orthodox" derivation of the reference prior can be quite involved (see, for example, van den Hout and van der Heijden, 2002; or Bockenholt, Barlas, and der Heijden, 2009). Our method, in a sense, deconvolves the problem and provides a reasonable "objective" prior.

TABLE 2.2. Frequentist coverage at level 0.95 of π^J for different values of θ and p.

	$n=10$			$n=30$			$n=100$		
	0.10	0.50	0.90	0.10	0.50	0.90	0.10	0.50	0.90
$p=0.1$	1	1	0.97	1	0.93	0.95	0.98	0.95	0.95
$p=.25$	1	1	1.94	1	0.97	0.92	0.99	0.95	0.95
$p=5$	1	1	0.94	1	0.95	0.92	1	0.95	0.95
$p=0.75$	1	1	0.96	1	0.94	0.94	1	0.95	0.94
$p=0.9$	1	1	0.95	1	0.96	0.93	1	0.94	0.96

2.2.2.2 Scale Mixtures

Suppose X is a random variable with density $f_X(x;\theta)$ which depends on some unknown parameter $\theta \in \mathbb{R}^d$. Let $I_{\theta\theta}$ be the corresponding Fisher matrix for the model. Let $Y = X/Z$ be a scale mixture arising from the above model; here X and Z are independent and Z is a positive random variable with some given density $g(z)$, which does not depend on θ. If we consider Z as an unknown parameter, the density of Y can be obviously written as

$$f_Y(y;\theta) = z f_X(zy;\theta). \tag{2.2.8}$$

The most popular example of this set up is the Student t distribution with (known) ν degrees of freedom; it is a scale mixture of Gaussian distributions and, in this case $Z^2 \sim \text{Gamma}(\nu/2, \nu/2)$. Another example is the scalar skew t distribution which can be seen as a scale mixture of skew normal random variables. The following result allows to directly derive the reference prior for scale mixture models.

Theorem 2.1. *Suppose $Y_1, \ldots, Y_n$ is a random sample from the density (2.2.8), and write each Y_j as X_j/Z_j with $\mathbf{Z} = (Z_1, \ldots, Z_n)$ unknown parameters. Then the upper left corner of the Fisher matrix related to θ does not depend on $\mathbf{Z}$, and the conditional reference prior for θ in the model (2.2.8) is the same as the reference prior for θ in the model $f_X(x;\theta)$.*

Proof. Without loss of generality, assume that $d = 1$, so θ is a scalar parameter. The log-likelihood function for $(\theta, \mathbf{Z})$ is

$$\ell(\theta, \mathbf{Z}) = \sum_{j=1}^{n} \left[\log Z_j + \log f_X(z_j y_j;\theta)\right].$$

Straightforward algebra shows that

$$-\frac{\partial^2}{\partial\theta^2}\ell(\theta,\mathbf{Z}) = \sum_{j=1}^{n} \frac{1}{f_X(z_j y_j;\theta)} \left\{ \frac{[(\partial/\partial\theta) f_X(z_j y_j;\theta)]^2}{f_X(z_j y_j;\theta)} + \frac{\partial^2}{\partial\theta^2} f_X(z_j y_j;\theta) \right\}.$$

Then

$$I_{\theta\theta} = \sum_{j=1}^{n} \int_{\mathbb{R}} \left\{ \frac{[(\partial/\partial\theta) f_X(z_j y_j;\theta)]^2}{f_X(z_j y_j;\theta)} + \frac{\partial^2}{\partial\theta^2} f_X(z_j y_j;\theta) \right\} Z_j dy_j =$$
$$= \sum_{j=1}^{n} \int_{\mathbb{R}} \left\{ \frac{[(\partial/\partial\theta) f_X(w_j;\theta)]^2}{f_X(w_j;\theta)} + \frac{\partial^2}{\partial\theta^2} f_X(w_j;\theta) \right\} dw_j = n I_{\theta\theta}.$$

The obvious consequence of this theorem is that the reference prior for the location and scale parameters of a Student t distribution with a known number of degrees of freedom is the same as the reference prior for the parameters of a Gaussian model. This results was already known. However it is not known what is the exact reference prior for the parameters of a skew t distribution (Azzalini and Capitanio, 2003). The above result suggests that this reference prior can be the same as the one derived in Liseo and Loperfido (2006) for the scalar skew normal model (see also Bayes and Branco, 2007, for a useful approximation).

2.2.2.3 The Degrees of Freedom of a Student t Density

Here we consider an example of type B, where the distribution of the vector $\mathbf{z}$ actually depends on θ and the conditional distribution of $\bar{Y}$ given $\mathbf{z}$ does not.

Suppose $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \text{St}(0,1,\nu)$, that is for the sake of simplicity we assume the location and scale parameters of the t distribution are known and the object of interest is the value of the degrees of freedom. It is well known that, in this case, Y_i can be represented as $Y_i = X_i/\sqrt{Z_i}$ where X_i is independent of Z_i, $X_i \sim \text{N}(0,1)$ and $Z_i \sim \text{Ga}(\nu/2, \nu/2)$. Using this representation, one can say that, for $i = 1, \ldots, n$,

$$Y_i \mid Z_i \sim N(0, 1/Z_i).$$

The augmented log-likelihood for the model is then

$$\ell(\mathbf{z},\nu) = c + \left(\frac{\nu-1}{2}\right) \sum_i \log Z_i - \frac{1}{2} \sum_i Z_i Y_i^2 + \frac{n\nu}{2} \log \frac{\nu}{2} - n \log \Gamma\left(\frac{\nu}{2}\right) - \frac{\nu}{2} \sum_i Z_i,$$

where c is constant with respect to $\mathbf{z}$ and ν. Now we consider the Fisher information matrix with respect to the enlarged model, where the expected values will be taken with respect the joint distribution of $(\bar{Y}, \mathbf{Z})$. Standard calculations show that

$$I_{\nu,\nu} = \mathbf{E}\left(-\frac{\partial^2}{\partial\nu^2} \ell(\mathbf{z},\nu) \right) = \frac{n}{4} \left(\psi'\left(\frac{\nu}{2}\right) - \frac{2}{\nu} \right),$$
$$I_{z_i,\nu} = \mathbf{E}\left(-\frac{\partial^2}{\partial z_i \partial\nu} \ell(\mathbf{z},\nu) \right) = -\frac{1}{\nu-2},$$
$$I_{z_i,z_i} = \mathbf{E}\left(-\frac{\partial^2}{\partial z_i^2} \ell(\mathbf{z},\nu) \right) = \frac{\nu^2(\nu-1)}{2(\nu-2)(\nu-4)},$$
$$I_{z_i,z_j} = \mathbf{E}\left(-\frac{\partial^2}{\partial z_i \partial z_j} \ell(\mathbf{z},\nu) \right) = 0,$$

where $\psi'(\cdot)$ is the Trigamma function, defined as the second derivative of the logarithm of the Eulero Gamma function.

Following the approach in Section 2.2.1, the marginal reference prior for ν is then given by

$$\pi^{(R)}(\nu) \propto (\mid I \mid / \mid I_{\mathbf{z},\mathbf{z}} \mid)^{1/2},$$

where $I_{\mathbf{z},\mathbf{z}}$ is the lower right corner of the matrix I, relative to the vector parameter $\mathbf{z}$. The ratio of the two determinants can be evaluated as a special case of the following theorem, which is stated without proof.

Theorem 2.2. *Let $(a_1, a_2, \ldots, a_n)$, $(b_1, b_2, \ldots, b_n)$ and c be all real numbers not equal to zero. Suppose D is a symmetric $(n+1) \times (n+1)$ matrix with the following structure*

$$\begin{aligned} d_{11} &= c;\ d_{1j} = d_{j1} = a_j,\ j = 2, \ldots, n+1, \\ d_{jj} &= b_j,\ j = 2, \ldots, n+1, \\ d_{ij} &= 0, \textit{ otherwise}. \end{aligned}$$

Then

$$\mid D \mid = c \prod_{i=1}^{n} b_i - \sum_{j=1}^{n} \left(a_j^2 \prod_{k \neq j} b_k \right). \tag{2.2.9}$$

When $a_j = a$ and $b_j = b$ for all j, (2.2.9) simplifies to

$$\mid D \mid = b^{n-1}(bc - na^2).$$

Then the ratio is equal to

$$\mid I \mid / \mid I_{\mathbf{z},\mathbf{z}} \mid = n \left[\psi' \left(\frac{\nu}{2} \right) - \frac{1}{\nu} \right] - \frac{2n(\nu - 4)}{\nu^2(\nu - 1)(\nu - 2)}. \tag{2.2.10}$$

The reference prior $\pi^R(\nu)$ is then defined as the square root of the above quantity. Now we study the asymptotic behavior of $\pi^R(\nu)$. To this end, we recall one property of the Trigamma function: for $a \to \infty$,

$$\psi'(a) = a^{-1} + (2a^2)^{-1} + (6a^3)^{-1} + o(a^{-4}).$$

Then, for large values of ν, (2.2.10) can be approximated by

$$\mid I \mid / \mid I_{\mathbf{z},\mathbf{z}} \mid \approx n \left(\frac{\nu + 2}{\nu^2} - \frac{2(\nu - 4)}{\nu^2(\nu - 2)(\nu - 1)} \right) = O(\nu^{-2}),$$

and the reference prior obtained in this way is improper. This result differs from the one obtained, in the orthodox way, by Fonseca, Ferreira, and Migon (2008). Since it is known that the likelihood function arising from an i.i.d. sample of t observation approaches a constant as $\nu \to \infty$, our prior cannot be used. Although this is

a negative result, we include it in this section as an example of a conjecture: our approximate reference prior tends to be more diffuse than the orthodox ones. As a consequence, when dealing with improper priors, one should always check whether the use of the prior may lead to improper posterior distributions, as in this example.

2.2.2.4 Probit Model

Consider the standard probit model, where n independent Bernoulli r.v.'s $(Y_1, \ldots, Y_n)$ are observed and, for $i = 1, \ldots, n$,

$$Y_i = \begin{cases} 1 & \Phi(\mathbf{x}_i'\beta), \\ 0 & 1 - \Phi(\mathbf{x}_i'\beta), \end{cases}$$

where $\mathbf{x}_i$ and β are p-dimensional vectors of covariates and parameters.

Textbook Bayesian analysis (Rossi, Allenby, and McCulloch, 2005) of this model is performed via the introduction of a latent variable Z_i for each observation, in such a way that

$$Z_i = \mathbf{x}_i'\beta + \varepsilon_i, \qquad \varepsilon_i \sim N(0,1),$$

and

$$Y_i = \begin{cases} 1 & Z_i > 0, \\ 0 & Z_i \leq 0. \end{cases}$$

The augmented likelihood function $f(\mathbf{y}, \mathbf{z}; \beta)$ then factorizes into

$$f(\mathbf{y}, \mathbf{z} \mid \beta) = f(\mathbf{y} \mid \mathbf{z}) f(\mathbf{z} \mid \beta) \tag{2.2.11}$$

with

$$f(\mathbf{z} \mid \beta) = \prod_{i=1}^{n} \varphi(z_i; \mathbf{x}_i'\beta, 1),$$

where $\varphi(x; b, c)$ denotes the density of a normal random variable with mean b and variance c, evaluated at x. The relation between the y_i's and the z_i's is then deterministic:

$$f(\mathbf{y} \mid \mathbf{z}, \beta) = \prod_{i=1}^{n} \left[I(z_i > 0) I(y_i = 1) + I(z_i \leq 0) I(y_i = 0) \right].$$

From equation (2.2.11), this model belongs to case B and the reference prior for β is given by (2.2.4). Simple calculations show that the extended Fisher matrix for $(\beta, \mathbf{z})$ is given by

$$I(\beta, \mathbf{z}) = \begin{pmatrix} \mathbf{X}'\mathbf{X} & 2\mathbf{X}' \\ 2\mathbf{X} & \mathbf{I}_n \end{pmatrix},$$

where $\mathbf{X}$ is the design matrix and $\mathbf{I}_n$ is the n-dimensional diagonal matrix. Since the above matrix does not depend on β it is clear that the reference prior for β is then

the flat prior

$$\pi^{(R)}(\beta) \propto 1.$$

This prior has been used for many years, although it has not been derived as "the" objective prior for the probit model.

2.2.3 The Case with Nuisance Parameters

In practice, one often use models with many nuisance parameters and our method should be generalized to these situations. Suppose that θ can be split into a vector ψ of parameters of interest and a vector λ, consisting of nuisance parameters. In this case, starting from the general expression of the model

$$p(y,z;\psi,\lambda) = g(y;z,\psi,\lambda)h(z;\psi,\lambda),$$

different special situations can arise. In particular, we will discuss the following two cases:

(C) g depends on ψ and h depends on λ, that is

$$p(y,z;\psi,\lambda) = g(y;z,\psi)h(z;\lambda); \tag{2.2.12}$$

(D) g depends on λ and h depends on ψ, that is

$$p(y,z;\psi,\lambda) = g(y;z,\lambda)h(z;\psi).$$

Consider first case C. Following in spirit the reference prior algorithm, one should first derive the conditional prior for $\lambda \mid \psi$ and then the marginal prior for ψ. In this case one starts taking ψ as known. From (2.2.12) it is clear that, conditionally on ψ, we are back to case B of Section 2.2.1. Then it is possible to derive the approximate reference prior for λ (notice that this a marginal prior with respect to $\mathbf{z}$ and a conditional prior with respect to ψ!), say $\pi^{(R)}(\lambda \mid \psi)$. Now, one could ideally integrate out λ and obtain

$$\pi(\mathbf{z};\psi) = \int_{\lambda} h(\mathbf{z};\lambda)\pi^{(R)}(\lambda \mid \psi)d\lambda$$

and use again the approach for case B in 2.2.1. However the above integral rarely has a closed and simple form: then it might be preferable to consider $\mathbf{z}$ as a vector of observations and derive the marginal prior for ψ as in the usual reference prior algorithm. In this case one could start from the model $p(\mathbf{y},\mathbf{z};\lambda,\psi)$ and from the conditional reference prior $\pi^{(R)}(\lambda \mid \psi)$, and directly apply the approach described in Sun and Berger (1998).

Consider now case D: here the augmented model can be written as $p(\mathbf{y},\mathbf{z};\psi,\lambda) = g(\mathbf{y};\mathbf{z},\lambda)h(\mathbf{z};\psi)$. Taking ψ as fixed, one falls again in case A of Section 2.2.1; it is

then easy to derive the conditional (on ψ and $\mathbf{z}$) reference prior for λ as

$$\pi^{(r)}(\lambda \mid \psi, \mathbf{z}) \propto \sqrt{I_{\lambda,\lambda}(\lambda, \psi)}.$$

In this case the use of the latent variables makes the observations $\mathbf{y}$ independent of ψ (conditionally on $\mathbf{z}$). Then the marginal reference prior for ψ can be derived using $h(\mathbf{z}; \psi)$ as the likelihood and the conditional reference prior for λ as our partial prior information.

Negative binomial regression example. The negative binomial regression model is very popular when overdispersion is suspected in regression analysis for count data. It is often expressed in hierarchical terms as follows: let $(Y_1, \ldots, Y_n)$ be independent Poisson r.v. with

$$\mathbf{E}(Y_i) = Z_i \exp\left\{\mathbf{x}_i'\beta\right\}, \quad i = 1, \ldots n.$$

For simplicity set $\mu_i = \exp\{\mathbf{x}_i'\beta\}$. The overdispersion is introduced by assuming that

$$Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} \text{Gamma}(\delta, \delta).$$

This way, while the unconditional mean of Y_i is still μ_i, the unconditional variance is

$$\text{Var}(Y_i) = \mu_i \left(1 + \mu_i \delta^{-1}\right).$$

The latent variable Z_i will be then included in the augmented log-likelihood which is then, up to an additive constant, equal to

$$\ell(\mu, \delta) = \sum_{i=1}^{n} \left[-z_i(\mu_i + \delta) + y_i \log(z_i \mu_i) + \delta \log \delta - \log \Gamma(\delta) + (\delta - 1) \log z_i\right].$$

Simple calculations show that the Fisher matrix for (δ, μ) is of the form

$$I(\delta, \mu) = \begin{pmatrix} c & a & a & \ldots & a \\ a & b_1 & 0 & \ldots & 0 \\ a & 0 & b_2 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a & 0 & 0 & \ldots & b_n \end{pmatrix}$$

with

$$b_i = \mu_i \frac{\delta}{\delta - 1} + \frac{\delta^2}{\delta - 2}, \qquad i = 1, \ldots n,$$

and

$$c = n(\psi'(\delta) - \delta^{-1}); \qquad a = -(\delta - 1)^{-1}.$$

Using Theorem 2.2, one can see that the conditional (on μ) reference prior for δ is

$$\pi^{(R)}(\delta \mid \mu) \propto \left\{ n\psi'(\delta) - \frac{n}{\delta} - \frac{\delta - 2}{\delta(\delta - 1)} \sum_{i=1}^{n} \{\mu_i(\delta - 2) + \delta(\delta - 1)\}^{-1} \right\}^{1/2}.$$

Under the assumption of no covariates ($\mu_i = \mu = e^{\beta}$) the above expression simplifies to

$$\pi^{(R)}(\delta \mid \mu) \propto \left\{ \psi'(\delta) - \frac{1}{\delta} - \frac{\delta - 2}{\delta(\delta - 1)} \{\mu(\delta - 2) + \delta(\delta - 1)\}^{-1} \right\}^{1/2}.$$

Now we derive the marginal prior for μ (or β) under the simplifying assumption of no covariates. To this end it is easier to work with the negative binomial model with parameters (δ, μ) and directly apply the Sun and Berger approach, which is based on (2.2.3) and (2.2.4). The log-likelihood is then

$$\ell(\delta, \mu) = \text{const.} + \sum_{i=1}^{n} \left[y_i \log \frac{\mu}{\mu + \delta} + \delta \log \frac{\delta}{\mu + \delta} + \log \Gamma(y_i + \delta) - \log \Gamma(\delta) \right].$$

Standard calculations show that the Fisher matrix in (δ, μ) is diagonal and

$$\frac{|I|}{I_{\delta,\delta}} = I_{\mu,\mu} = n \frac{\delta}{\mu(\mu + \delta)},$$

which depends on δ. Then the marginal prior for μ is

$$\pi^{(R)}(\mu) = \exp\left(\frac{1}{2} \int \pi(\delta \mid \mu) \log \left\{ n \frac{\delta}{\mu(\mu + \delta)} \right\} d\delta \right), \tag{2.2.13}$$

which must be evaluated numerically for each value of μ. Notice that the use of this approach avoids the explicit evaluation of the expected value of $\psi'(Y + \delta)$ which would be necessary under the regular reference prior algorithm for this model.

2.2.4 Conclusions

We have discussed a simple method to derive (approximate) reference priors when the statistical model is expressed in terms of a latent structure. In many situations the method produces the "orthodox" reference priors. In some cases, however, the result is different and our conjecture is that the "true" reference prior is never more vague than the one we obtain.

We start from the basic simple idea that the latent variables are turned into unknown parameters and the Fisher information matrix has then the dimension of $(\theta, \mathbf{z})$. Consequently, the expected values of the likelihood function should be calculated with respected to the conditional distribution of Y given $(\theta, \mathbf{z})$. In the Student t example with an unknown number of degrees of freedom, we calculated the Fisher matrix by taking expectation also with respect to $\mathbf{Z}$. We did that just for computational reasons, since taking expectation only with respect to $\bar{Y}$ would produce messy results.

Our future work will be to establish, in terms of some probability metrics, a distance between the proposed prior and the true reference prior. For example it will be interesting to explore whether the priors we obtained are still matching priors.

2.3 Reference Priors for Empirical Likelihoods

Bertrand Clarke and Ao Yuan

Since Owen (1988, 1990, 1991), empirical likelihood (EL) techniques have gained popularity largely because they incorporate information for parameter estimation into a non-parametric context by constrained optimization. Despite extensive use in the Frequentist context, EL has only recently come into use in Bayesian analysis. Lazar (2003) observed that the properties of EL are in many respects similar to those of parametric likelihoods and proposed ways they could be used in Bayesian inference. She presented several simulations under different conditions to show the effect of prior selection in ELs was much the same as in independence likelihoods. Further similarities between ELs and parametric likelihoods have been delineated in Yuan, Zheng, and Xu (2009). The implication of this is that reference priors for ELs may behave similarly to the way reference priors in independence likelihoods do. Specifically, they may give slightly narrower credibility sets than the normal priors with large variances as studied in Lazar (2003). In econometrics, Moon and Schorfheide (2004) used ELs for a Bayesian analysis by choosing priors that put most of their mass on parameter values for which the moment constraint was approximately satisfied. Recently, Grendar and Judge (2009) established that ELs can be regarded as a posterior mode in an asymptotic sense.

Here, our main contribution is the identification of reference priors for empirical likelihood. This is important because, in principle, once a model and prior have been chosen, the posterior is determined and Bayesian analysis can proceed computationally. Indeed, automating prior selection — regardless of whether the resulting priors are used to form credibility sets — can help with posterior exploration. We recall that reference priors are merely one class of objective priors, see Liu and Ghosh (2009) for a recent survey.

In the next section we briefly review the formulation of empirical likelihoods. Then, in Section 2.3.2, we review the concept of reference priors for IID likelihoods, set up the corresponding optimization problem for the EL, and quote a result that will help us solve it. In Sections 2.3.3–2.3.5, we state our main results giving asymptotic expansions for three distances between priors and posteriors obtained from ELs and identify the reference priors they generate. In Section 2.3.6, we discuss the implications of our work.

2.3.1 Empirical Likelihood

The basic formulation of EL as given by Owen (1988) can be expressed as in Qin and Lawless (1994) and stated as follows. Let $X^n = (X_1, ..., X_n)$ be IID d-dimensional random vectors with unknown distribution function F and suppose the q-dimensional parameter $\theta = (\theta_1, \ldots, \theta_q)'$ is a functional value of F, i.e., there is a function T so that $T(F) = \theta$. Write $x^n = (x_1, \ldots, x_n)$ to denote outcomes of X^n and x to denote outcomes of an individual random variable X. Suppose that additional information linking θ and F is available from a set of functions $g(x, \theta) = (g_1(x, \theta), ..., g_r(x, \theta))'$ where $r \geq q$ and $E[g(X, \theta)] = 0$. The expectation is taken in the distribution $F = F_T$ taken to be true and it is assumed that the true θ, θ_T satisfies $T(F_T) = \theta_T$. Indeed, here we assume that θ is identifiable with respect to g, i.e., $\theta^* \neq \theta$ implies $E[g(X, \theta^*)] \neq 0$.

An expression for the EL can be given as follows. Let F be a distribution function varying over a class and consider the likelihood

$$L(F) = \prod_{i=1}^{n} F(\{x_i\}). \tag{2.3.1}$$

Now, write $w_i = F(\{x_i\})$ and $w = (w_1, ..., w_n)$. The EL is subject to the auxilliary information constraints from g and achieves

$$\max_{w} \prod_{i=1}^{n} w_i \quad \text{subject to} \quad \sum_{i=1}^{n} w_i = 1 \quad \text{and} \quad \sum_{i=1}^{n} w_i g(x_i, \theta) = 0.$$

Let $t = (t_1, ..., t_r)'$ be the Lagrange multipliers corresponding to the constriant with $g(x, \theta)$, then one can derive

$$w_i = \frac{1}{n} \frac{1}{1 + t' g(x_i, \theta)}$$

and that $t = t_n(x_1, ..., x_n, \theta)$ is determined by

$$\sum_{i=1}^{n} \frac{g(x_i, \theta)}{1 + t' g(x_i, \theta)} = 0. \tag{2.3.2}$$

Note that $t = 0$ satisfies (2.3.2), but then $w_i = 1/n$ and the side information from g does not enter the EL. So, we henceforth require $t \neq 0$ to avoid trivality. Thus, the empirical likelihood assumes the form

$$p_\theta^n = p(x^n | \theta) = \prod_{i=1}^{n} \frac{1}{n} \frac{1}{1 + t' g(x_i, \theta)} = \prod_{i=1}^{n} w_i. \tag{2.3.3}$$

Note that even though the data is IID F, (2.3.3) does not in general factor into a product of terms each depending on only one of the x_is. Thus, ELs are not in general independence likelihoods (unlike (2.3.1)) even though they may be regarded

as identical. They are a generalization of IID to permit the dependence structure induced by the constraint. Indeed, the data enter the constraint symmetrically so we expect that they will remain symmetric in the empirical likelihood itself. Because of this dependence, it is difficult to assign priors to ELs and so Bayesian analysis has been limited. Our main contribution is the extension of objective Bayes methods to the EL context by deriving reference priors for them.

2.3.2 Reference Priors

In a pair of seminal papers, Shannon (1948a, 1948b) gave an outline of the general theory of communication. One of the basic ideas was to reinterpret the conditional density given a parameter, or likelihood, as an "information theoretic channel." The idea is that θ is a message drawn from a source distribution of messages, say Π with density π, and the sender wants to send a randomly chosen message θ to a collection of receivers. The receivers, however, do not receive θ exactly. Each receiver for $i = 1, \ldots, n$ receives a noisy version of θ, say $X_i = x_i$, from which they want to decode θ. The relationship between the θ sent and the x received is given by $p(x|\theta)$; the difference between a channel and likelihood is that the channel is a conditional density that will be used repeatedly (with both arguments redrawn), whereas a likelihood is a function of θ for fixed x^n. Now, assume each of the n receivers receives an x_i independently of the rest, but they pool their x_is to decode θ. If this process occurs many times, Shannon showed the rate of information transmission is

$$I(\Theta; X^n) = \int \int \pi(\theta) p(x^n|\theta) \log \frac{p(x^n|\theta)}{m(x^n)} \mu(dx^n) \mu(d\theta), \tag{2.3.4}$$

(in nats per symbol sent) where μ generically denotes a dominating measure for its argument. The quantity in (2.3.4) is the (Shannon) mutual information. The natural question is how large it can be. This is answered by maximizing over W to find the maximal rate, the capacity of the channel $p(\cdot|\cdot)$. The result is

$$\Pi_{cap}(\cdot) = \arg\max_{\Pi} I(\Theta; X^n),$$

the capacity achieving source distribution. Asymptotically in n, Ibragimov and Hasminsky (1973) showed Π_{cap} was Jeffreys prior for regular finite-dimensional parametric families.

Bernardo (1979) wrote

$$I(\Theta, X^n) = E_m D(\pi(\cdot) \| \pi(\cdot|X^n)),$$

where, for densities p and q with respect to a common dominating measure, the relative entropy is

$$D(p||q) = \int p(x) \log \frac{p(x)}{q(x)} \mu(dx).$$

That is, the capacity achieving source distribution is the prior that makes the asymptotic distance between a prior and its corresponding posterior as far apart as possible, on average, in relative entropy. Bernardo (1979) also called Π_{cap} a reference prior on the grounds that it could be used as a prior, or more typically, used as a way to assess the amount of information in a subjective prior.

We comment that reference prior results are asymptotic in n and we assume this without further comment apart from noting that reference priors obtained for fixed n are usually discrete, see Berger, Bernardo, and Mendoza (1989). Even so, Zhang (1994) provided a convergence result ensuring the discrete priors converge to Jeffreys prior for many regular parametric families.

Berger and Bernardo (1989) examined a conditional form of the Shannon mutual information to identify

$$\arg\max_{\Pi} I(\Theta; X^n|\Psi) = \arg\max_{\Pi} \int \pi(\theta|\psi)p(x^n|\theta,\psi)\log\frac{p(x^n|\theta,\psi)}{\int p(x^n|\theta,\psi)\pi(\theta|\psi)\mu(d\theta)} \mu(dx^n)\mu(d\theta)\pi(\psi)\mu(d\psi),$$

where $(\Theta,\Psi) = (\Theta_1,\ldots,\Theta_q,\Psi_1,\ldots,\Psi_\ell)$ and Ψ is a nuisance parameter. A proof for regular finite-dimensional families can be found in Ghosh and Mukerjee (1992). Further treatment of the multiparameter case can be found in Berger and Bernardo (1992a, 1992b, 1992c). Sun and Berger (1998) examined conditional mutual information further, and Clarke and Yuan (2004) gave a complete treatment.

Of recent interest is the work done by Ghosh, Mergel, and Liu (2009) and Liu and Ghosh (2009) to obtain reference priors under alternative measures of distance. They establish that Jeffreys prior is the reference prior for almost all members of the power divergence family. The exception is the Chi-square distance for which the prior turns out to be proportional to the fourth root of the determinant of the Fisher information.

To be precise about the quantities we examine for the EL, write

$$m_n(x^n) = m(x^n) = \int p(x^n|\theta)\pi(\theta)d\theta,$$

where $p(x^n|\theta)$ is an in (2.3.3), giving posterior

$$\pi(\theta|x^n) = \frac{\pi(\theta)p(x^n|\theta)}{m(x^n)}.$$

Then, the relative entropy between $\pi(\theta|x^n)$ and $\pi(\theta)$ is

$$D(\pi(\cdot|x^n)||\pi(\cdot)) = \int \pi(\theta|x^n)\log\frac{\pi(\theta|x^n)}{\pi(\theta)}d\theta;$$

the Hellinger distance between $\pi(\theta|x^n)$ and $\pi(\theta)$ is

$$H(\pi(\cdot|x^n),\pi(\cdot)) = \int (\sqrt{\pi(\theta)} - \sqrt{\pi(\theta|x^n)})^2 d\theta;$$

and the Chi-square distance between $\pi(\theta|x^n)$ and $\pi(\theta)$ is

$$\chi^2(\pi(\cdot|x^n), \pi(\cdot)) = \int \frac{(\pi(\theta|x^n) - \pi(\theta))^2}{\pi(\theta)} d\theta.$$

Here, we examine the expectation of the above three quantities, namely

$$E_{m_n} D(\pi(\cdot|x^n)||\pi(\cdot)), \quad E_{m_n} H(\pi(\cdot|x^n), \pi(\cdot)), \quad \text{and} \quad E_{m_n} \chi^2(\pi(\cdot|x^n), \pi(\cdot)).$$

These three distance measures have interpretations that may make them more or less useful in a given setting. The relative entropy occurs in probabilistic coding theory and usually represents an amount of information (in nats). The Chi-square distance is familiar from goodness-of-fit testing (see Clarke and Sun, 1997; Hervé, 2007). The Hellinger distance originates from geometry in which the square root converts a great circle on the unit sphere to a line segment in a plane. It can be verified that as distances, $\chi^2(p,q) \geq D(p||q) \geq H(p,q)$ for any two densities p, q for which they are defined.

Observe that $m(x^n)$ is the Bayes action for estimating P_θ under relative entropy i.e.,

$$m(x^n) = \arg\min_Q \int w(\theta) D(P_\theta^n||Q) \mu(d\theta).$$

and the chain rule for relative entropy gives

$$D(P_\theta^n||M_n) = \sum_{k=1}^{n} E_m D(P_\theta||M_k(\cdot|X^{k-1})). \tag{2.3.5}$$

However, under Hellinger distance the Bayes action for estimating of P_θ is

$$m_H(x^n) = \left\{ \int w(\theta) p(x^n|\theta)^{1/2} \mu(d\theta) \right\}^2 = \arg\min_Q \int w(\theta) H(P_\theta^n||Q) \mu(d\theta),$$

and under Chi-square distance the Bayes action for estimating P_θ is

$$m_{\chi^2}(x^n) = \int w(\theta) p(x^n|\theta)^2 \mu(d\theta) = \arg\min_Q \int w(\theta) \chi^2(P_\theta^n||Q) \mu(d\theta).$$

It is seen that neither is a probability (unless additional constraints are imposed) and neither satisfies an additive risk condition like (2.3.5). This means that the reference priors under Hellinger or Chi-square distance are not for the Bayes action and so need not be least favorable. However, they do provide priors maximally changed on average by the data.

Note that, to date, almost all reference prior work has been in the regular parametric family context. However, there are cases, such as EL, in which we do not have a well-defined IID parametric likelihood. Indeed, as can be seen from (2.3.3), the EL is stationary but not independent. However, the stationarity is close enough to independence that MLEs are consistent and Laws of Large Numbers and Central Limit Theorems hold.

To see this more formally, define the following notation. Let θ be the "true" parameter value for the observed data and assume θ is in an open set whose closure is compact. Write $l_i(\theta) = \log w_i(\theta)$ with first derivative denoted $l_i^{(1)}(\theta) = \partial l_i(\theta)/\partial\theta$ and second derivative denoted $l_i^{(2)}(\theta) = \partial l_i(\theta)/[\partial\theta\partial\theta']$. Next, consider the following regularity conditions.

- R1. The constraint function has bounded moments, i.e., $E||g(X,\theta)||^\alpha < \infty$ for some $\alpha > 2$.
- R2. The outer product matrix $\Omega = E[g(X,\theta)g'(X,\theta)]$ is positive definite.
- R3. The Jacobian matrix $D = E[\partial g(X,\theta)/\partial\theta]$ is of rank r.
- R4. The norms $||g(x,\theta)||$ and $||g'(x,\theta)g(x,\theta)||$ are bounded by an integrable function function $G(x)$, in each neighborhood of θ.
- R5. The prior $\pi(\cdot)$ is continuous and the matrix $\Lambda(\theta) = D'(\theta)\Omega^{-1}(\theta)D(\theta)$ is invertible.
- R6. The prior $\pi(\cdot)$ and the $l_i^{(2)}(\cdot)$ for $i = 1,\ldots,n$ are bounded.

Let $\hat{\theta}_n = \arg\sup_\theta \log p(x^n|\theta)$ denote the maximum empirical likelihood estimate of θ. Then, we have the following asymptotic results which parallel the corresponding results for regular IID likelihoods.

Theorem 2.3. *Assume R1–R4. Then $\hat{\theta}$ is consistent and asymptotically normal with asymptotic variance matrix $\Lambda^{-1}(\theta)$. That is,*

$$\hat{\theta}_n \to \theta \;\; a.s. \;\; and \;\; \sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{D} N(0, \Lambda^{-1}).$$

Proof. See Yuan, Zheng, and Xu (2009).

2.3.3 Relative Entropy Reference Priors

Equipped with the EL setting of Section 2.3.1 and the reference prior formulation of Section 2.3.2, we can now state the first of our main results.

Theorem 2.4. *Assume R1–R6. Then*

$$E_{m_n} D(\pi(\cdot|x^n)||\pi(\cdot)) = \frac{q}{2}\log\frac{n}{2\pi e} - \int \pi(\theta)\log\frac{\pi(\theta)}{|\Lambda^{-1}(\theta)|^{1/2}}d\theta + o(1).$$

So, the reference prior for the EL under relative entropy is $\pi^*_{KL}(\theta) \propto |\Lambda^{-1}(\theta)|^{1/2}$.

We comment that the proof of the asymptotic expression is a sequence of asymptotically valid approximations whose convergence identifies the leading terms. The highest-order term depending on the prior is optimized in the usual way to give the reference prior. The same comment applies to Theorems 2.5 and 2.6 below.

Proof. Recall $l_i(\theta) = \log w_i(\theta)$ and $l_i^{(2)}(\theta) = \partial^2 l_i(\theta)/(\partial\theta\partial\theta')$ and consider the limit of the mean of the $l_i^{(2)}$s. As in the proofs of Theorem 1 and 2 in Yuan, Zheng, and Xu (2009), we have

$$\begin{aligned} w_i(\theta) &= \frac{1}{n}\frac{1}{1+t'g(x_i,\theta)} = \frac{1}{n}\Big\{1 - t'g(x_i,\theta) + g'(x_i,\theta)g(x_i,\theta)O(n^{-1}(\log\log n))\Big\} \\ &= \frac{1}{n}\Big\{1 - B_n'g(x_i,\theta) + ||g(x_i,\theta)||o(n^{-(1-1/\alpha)}(\log\log(n)) \\ &\quad + g'(x_i,\theta)g(x_i,\theta)O(n^{-1}(\log\log n))\Big\}, \end{aligned} \tag{2.3.6}$$

where $B_n = \left\{\frac{1}{n}\sum_{i=1}^n g(x_i,\theta)g'(x_i,\theta)\right\}^{-1}\frac{1}{n}\sum_{i=1}^n g(x_i,\theta)$.

By using the Taylor expansion $\log(1+x) \approx x$ on (2.3.6) we get

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n l_i(\theta) = &-\log n - B_n'\frac{1}{n}\sum_{i=1}^n g(x_i,\theta) + \frac{1}{n}\sum_{i=1}^n ||g(x_i,\theta)||o(n^{-(1-1/\alpha)}(\log\log(n)) \\ &\frac{1}{n}\sum_{i=1}^n g'(x_i,\theta)g(x_i,\theta)O(n^{-1}(\log\log n)) + O(n^{-1}\log\log n). \end{aligned} \tag{2.3.7}$$

Taking second derivatives in (2.3.7) by using the product rule gives

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} l_i^{(2)}(\theta) \\
&= -\Big[2\Big\{\frac{\partial^2}{\partial\theta\partial\theta'}\frac{1}{n}\sum_{i=1}^{n} g'(x_i,\theta)\Big\}\Big\{\frac{1}{n}\sum_{i=1}^{n} g(x_i,\theta)g'(x_i,\theta)\Big\}^{-1}\frac{1}{n}\sum_{i=1}^{n} g(x_i,\theta) \\
&\quad +2\Big\{\frac{\partial}{\partial\theta}\frac{1}{n}\sum_{i=1}^{n} g'(x_i,\theta)\Big\}\Big\{\frac{\partial}{\partial\theta'}\Big(\frac{1}{n}\sum_{i=1}^{n} g(x_i,\theta)g'(x_i,\theta)\Big)^{-1}\Big\}\frac{1}{n}\sum_{i=1}^{n} g(x_i,\theta) \\
&\quad +\frac{1}{n}\sum_{i=1}^{n} g'(x_i,\theta)\Big\{\frac{\partial^2}{\partial\theta\partial\theta'}\Big(\frac{1}{n}\sum_{i=1}^{n} g(x_i,\theta)g'(x_i,\theta)\Big)^{-1}\Big\}\frac{1}{n}\sum_{i=1}^{n} g(x_i,\theta) \\
&\quad +\Big\{\frac{\partial}{\partial\theta}\frac{1}{n}\sum_{i=1}^{n} g'(x_i,\theta)\Big\}\Big\{\frac{1}{n}\sum_{i=1}^{n} g(x_i,\theta)g'(x_i,\theta)\Big\}^{-1}\Big\{\frac{\partial}{\partial\theta'}\frac{1}{n}\sum_{i=1}^{n} g(x_i,\theta)\Big\} \\
&\quad +\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta\partial\theta'}||g(x_i,\theta)||o(n^{-(1-1/\alpha)}(\log\log(n)) \\
&\quad +\Big\{\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta\partial\theta'}g'(x_i,\theta)g(x_i,\theta)+1\Big\}O(n^{-1}(\log\log n))\Big]. \qquad (2.3.8)
\end{aligned}
$$

By the strong law of large numbers, $\frac{1}{n}\sum_{i=1}^{n} g(x_i,\theta) \to Eg(X,\theta) = 0$ a.s., thus for any $\theta_n \to \theta$ (P or a.s.), only the fourth term on the right of (2.3.8) above is asymptotically non-zero. This gives

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} l_i^{(2)}(\theta_n) \\
&\to -\lim_n\Big\{\frac{\partial}{\partial\theta}\frac{1}{n}\sum_{i=1}^{n} g'(x_i,\theta)\Big\}\Big\{\frac{1}{n}\sum_{i=1}^{n} g(x_i,\theta)g'(x_i,\theta)\Big\}^{-1}\Big\{\frac{\partial}{\partial\theta'}\frac{1}{n}\sum_{i=1}^{n} g(x_i,\theta)\Big\} \\
&= -D'(\theta)\Omega^{-1}(\theta)D(\theta) = -\Lambda(\theta), \quad (P \text{ or } a.s.). \qquad (2.3.9)
\end{aligned}
$$

We use (2.3.9) in the following Laplace expansion argument.

By a second order Taylor expansion in Lagrange form, we have

$$
p(x^n|\theta) = \exp\Big[\sum_{i=1}^{n} l_i(\hat\theta_n) + \frac{1}{2}n(\hat\theta_n-\theta)'\Big\{\frac{1}{n}\sum_{i=1}^{n} l_i^{(2)}(\theta_n)\Big\}(\hat\theta_n-\theta)\Big], \quad (2.3.10)
$$

where θ_n is betwen $\hat\theta_n$ and θ. Similarly,

$$
m(x^n) = \int \pi(\theta)\exp\Big[\sum_{i=1}^{n} l_i(\hat\theta_n) + \frac{1}{2}n(\hat\theta_n-\theta)'\Big\{\frac{1}{n}\sum_{i=1}^{n} l_i^{(2)}(\theta_n)\Big\}(\hat\theta_n-\theta)\Big]d\theta. \qquad (2.3.11)
$$

So,

$$
\log\frac{p(x^n|\theta)}{m(x^n)} = \log\frac{\exp\Big[\frac{1}{2}n(\hat\theta_n-\theta)'\big\{\frac{1}{n}\sum_{i=1}^{n} l_i^{(2)}(\theta_n)\big\}(\hat\theta_n-\theta)\Big]}{\int\pi(\alpha)\exp\Big[\frac{1}{2}n(\hat\theta_n-\alpha)'\big\{\frac{1}{n}\sum_{i=1}^{n} l_i^{(2)}(\theta_n)\big\}(\hat\theta_n-\alpha)\Big]d\alpha}.
$$

Let $\phi(\cdot|\hat{\theta}_n, \Lambda)$ be the q-dimensional normal density with mean $\hat{\theta}_n$ and covariance matrix Λ. Now, for any $\delta > 0$,

$$\begin{aligned}
&\int \pi(\alpha) \exp\Big[\frac{1}{2} n(\hat{\theta}_n - \alpha)' \Big\{\frac{1}{n}\sum_{i=1}^{n} l_i^{(2)}(\theta_n)\Big\}(\hat{\theta}_n - \alpha)\Big] d\alpha \\
&= \int_{||\alpha - \hat{\theta}_n|| \le \delta} \pi(\alpha) \exp\Big[\frac{1}{2} n(\hat{\theta}_n - \alpha)' \Big\{\frac{1}{n}\sum_{i=1}^{n} l_i^{(2)}(\theta_n)\Big\}(\hat{\theta}_n - \alpha)\Big] d\alpha \\
&\quad + \int_{||\alpha - \hat{\theta}_n|| > \delta} \pi(\alpha) \exp\Big[\frac{1}{2} n(\hat{\theta}_n - \alpha)' \Big\{\frac{1}{n}\sum_{i=1}^{n} l_i^{(2)}(\theta_n)\Big\}(\hat{\theta}_n - \alpha)\Big] d\alpha. \qquad (2.3.12)
\end{aligned}$$

Write $\Lambda_n(\theta_n)^{-1} = -\frac{1}{n}\sum_{i=1}^{n} l_i^{(2)}(\theta_n)$, with θ_n and $\theta_{1,n}$ in the ball $B(\hat{\theta}_n, \delta) = \{\alpha : ||\alpha - \hat{\theta}_n|| \le \delta\}$. Then, the first term on the right in (2.3.12) is

$$\begin{aligned}
&\pi(\theta_{1,n}) \int_{||\alpha - \hat{\theta}_n|| \le \delta} \exp\Big\{-\frac{1}{2} n(\hat{\theta}_n - \alpha)' \Lambda_n^{-1}(\theta_n)(\hat{\theta}_n - \alpha)\Big\} d\alpha \\
&= \pi(\theta_{1,n})(2\pi)^{q/2} n^{-q/2} |\Lambda_n(\theta_n)|^{-1/2} \int_{||\alpha|| \le \delta\sqrt{n}} \phi(\alpha|0, I_q) d\alpha \\
&\sim \pi(\theta)(2\pi)^{q/2} n^{-q/2} |\Lambda^{-1}(\theta)|^{1/2}, \qquad (2.3.13)
\end{aligned}$$

since $\delta > 0$ is arbitrary, by R6. To deal with the second term in (2.3.12), note that it admits the bound

$$0 \le e^{(-n/4)\delta^2 ||\Lambda_n||} \int_{||\alpha'|| > \delta\sqrt{n}} \pi(\alpha') \exp\Big[\frac{1}{4} n\alpha' \Big\{\frac{1}{n}\sum_{i=1}^{n} l_i^{(2)}(\theta_n)\Big\}\alpha'\Big] d\alpha' \qquad (2.3.14)$$

with probability going to one, where $\alpha' = \hat{\theta}_n - \alpha$, since Λ_n converges to Λ. Doing the normal integral in (2.3.14) gives a factor of order $\mathcal{O}(n^{-q/2})$ which is multiplied by a factor of order $\exp[(-n\delta^2)||\Lambda_n(\theta_n)||]$. The result is of lower order than the first term in (2.3.12).

By Theorem 2.3, observe that $Y_n = n(\hat{\theta}_n - \theta)'\Lambda(\theta)(\hat{\theta}_n - \theta) \xrightarrow{D} \chi_q^2$ under $p(x^n|\theta)$. Since $E\chi_q^2 = q$ for $\varepsilon > 0$, we can find $M > 0$ such that $|E[\chi_q^2 I(\chi_q^2 \le M)] - q| < \varepsilon$. Weak convergence gives $E[Y_n I(Y_n \le M)] \to E[\chi_q^2 I(\chi_q^2 \le M)]$. Provided $n(\hat{\theta} - \theta)$ is uniformly integrable in P_θ, uniformly for θ, we have $E(Y_n) \to E(\chi_q^2) = q$ as $\varepsilon \to 0$.

Using (2.3.13) and (2.3.14) in (2.3.12) and the result from Theorem 2.3, we have

$$\begin{aligned}
&E_{m_n} D(\pi(\cdot|X^n)||\pi(\cdot)) = \int\int \pi(\theta) p(x^n|\theta) \log \frac{p(x^n|\theta)}{m(x^n)} d\mu(x^n) d\theta \\
&\sim -\int \pi(\theta) E_{p(x^n|\theta)}\Big\{\frac{1}{2} n(\hat{\theta}_n - \theta)'\Lambda(\theta)(\hat{\theta}_n - \theta)\Big\} d\theta \\
&\quad + \frac{q}{2}\log\frac{n}{2\pi} - \int \pi(\theta)\log\pi(\theta) d\theta - \frac{1}{2}\int \pi(\theta)\log|\Lambda^{-1}(\theta)| d\theta \\
&\sim -\frac{q}{2} + \frac{q}{2}\log\frac{n}{2\pi} - \int \pi(\theta)\log\pi(\theta) d\theta - \frac{1}{2}\int \pi(\theta)\log|\Lambda^{-1}(\theta)| d\theta.
\end{aligned}$$

2.3.4 Hellinger Reference Prior

Next, we state and prove the analogous result for the Hellinger distance.

Theorem 2.5. *Assume R1–R6. Then*

$$\begin{aligned} &E_{m_n} H(\pi(\cdot|x^n), \pi(\cdot)) \\ &= (2\pi/n)^{q/4} E\big[\exp\big(\frac{1}{4}\chi_q^2\big)\big] \int |\Lambda^{-1}(\theta)|^{1/4} \pi^{3/2}(\theta) d\theta + o(\frac{1}{n^{q/4}}). \end{aligned}$$

So, the reference prior for the EL under the Hellinger metric is $\pi_H^*(\theta) \propto |\Lambda(\theta)|^{1/2}$.

Note that the reference prior under the Hellinger distance is the inverse of the reference prior under the relative entropy.

Proof. It is easy to see that

$$E_{m_n} H(\pi(\cdot|x^n), \pi(\cdot)) = 2\Big\{1 - \int\int \pi(\theta)\sqrt{m(x^n)p(x^n|\theta)} d\mu(x^n) d\theta\Big\}. \quad (2.3.15)$$

Recalling (2.3.10) and (2.3.11) from the proof of Theorem 2.4, we set up a slight extension of Laplace's method by expanding the prior to second order. Let $\pi^{(1)}(\alpha) = \partial\pi(\alpha)/\partial\alpha$ and $\pi^{(2)}(\alpha) = \partial^2\pi(\alpha)/[\partial\alpha\partial\alpha']$ and recall that $\int \alpha' A\alpha \phi(\alpha|0, I_q) d\alpha = tr(A)$. Then, taking convergences in $p(x^n|\theta)$, we have

$$\begin{aligned} &\int \pi(\alpha) \exp\Big[\frac{1}{2} n(\hat\theta_n - \alpha)' \Big\{\frac{1}{n}\sum_{i=1}^n l_i^{(2)}(\theta_n)\Big\}(\hat\theta_n - \alpha)\Big] d\alpha \\ =&(2\pi/n)^{q/2} |\Lambda^{-1}(\theta_n)|^{1/2} \\ &\times \int \Big\{\pi(\hat\theta_n) + (\alpha - \hat\theta_n)'\pi^{(1)}(\hat\theta_n) + \frac{1}{2}(\alpha - \hat\theta_n)'\pi^{(2)}(\theta_{2,n})(\alpha - \hat\theta_n)\Big\} \\ &\times \phi(\alpha|\hat\theta_n, \Lambda^{-1}(\theta_n)/n) d\alpha \\ =&(2\pi/n)^{q/2} |\Lambda^{-1}(\theta_n)|^{1/2} \int \Big\{\pi(\hat\theta_n) + \frac{1}{n}\alpha' \Lambda^{-1/2'}(\theta_n)\pi^{(2)}(\theta_{2,n})\Lambda^{-1/2}(\theta_n)\alpha\Big\} \\ &\times \phi(\alpha|0, I_q) d\alpha \\ \sim&(2\pi/n)^{q/2} |\Lambda^{-1}(\theta_n)|^{1/2} \pi(\hat\theta_n)\Big[1 + \frac{1}{2n} tr\big\{\Lambda^{-1/2'}(\theta)\frac{\pi^{(2)}(\theta)}{\pi(\theta)}\Lambda^{-1/2}(\theta)\big\}\Big] \\ \sim&(2\pi/n)^{q/2} |\Lambda^{-1}(\theta)|^{1/2} \pi(\hat\theta_n)\Big[1 + \frac{1}{2n} tr\big\{\Lambda^{-1/2'}(\theta)\frac{\pi^{(2)}(\theta)}{\pi(\theta)}\Lambda^{-1/2}(\theta)\big\}\Big] \\ \sim&(2\pi/n)^{q/2} |\Lambda^{-1}(\theta)|^{1/2} \pi(\theta). \end{aligned} \quad (2.3.16)$$

The key term in (2.3.15) is

$$\int\int \pi(\theta)\sqrt{m(x^n)p(x^n|\theta)} d\mu(x^n) d\theta = \int\int \sqrt{\frac{m(x^n)}{p(x^n|\theta)}} \pi(\theta) p(x^n|\theta) d\mu(x^n) d\theta.$$

So, using the square root of the ratio of (2.3.10) to (2.3.11), and (2.3.16) we have that

$$\int\int \pi(\theta)\sqrt{m(x^n)p(x^n|\theta)}d\mu(x^n)d\theta \sim (2\pi/n)^{q/4}\int\Big[\int |\Lambda^{-1}(\theta)|^{1/4}\pi^{3/2}(\theta) \times \exp\big\{\frac{1}{4}n(\hat{\theta}_n-\theta)'\Lambda(\theta)(\hat{\theta}_n-\theta)\big\}p(x^n|\theta)d\mu(x^n)\Big]d\theta.$$

By Theorem 2.3, $\hat{\theta}_n \to \theta$ a.s. and $n(\hat{\theta}_n-\theta)'\Lambda(\theta_n)(\hat{\theta}_n-\theta) \xrightarrow{D} \chi^2_q$ under $p(x^n|\theta)$. So, if the convergence of the exponent to χ^2_q is uniform over θ, we get

$$\int\int \pi(\theta)\sqrt{m(x^n)p(x^n|\theta)}d\mu(x^n)d\theta \sim (2\pi/n)^{q/4}E\Big[\exp\big(\frac{1}{4}\chi^2_q\big)\Big]\int |\Lambda^{-1}(\theta)|^{1/4}\pi^{3/2}(\theta)d\theta,$$

as claimed and

$$\pi^*(\theta) = \arg\min_{\pi}\Big\{\int |\Lambda^{-1}(\theta)|^{1/4}\pi^{3/2}(\theta)d\theta \quad \text{subject to} \quad \int \pi(\theta)d\theta = 1\Big\}.$$

Using Lagrange multipliers and taking derivatives of $\int |\Lambda^{-1}(\theta)|^{1/4}\pi^{3/2}(\theta)d\theta - \lambda\int\pi(\theta)d\theta$ with respect to $\pi(\theta)$ for fixed θ, we get

$$\frac{3}{2}|\Lambda(\theta)|^{-1/4}\pi^{1/2}(\theta) - \lambda = 0, \quad \text{or} \quad \pi(\theta) \propto |\Lambda(\theta)|^{1/2}.$$

2.3.5 Chi-square Reference Prior

The following result under the Chi-square distance is analogous to Clarke and Sun (1997) and Ghosh, Mergel, and Liu (2009), however, the solution is hard to obtain explicitly.

Theorem 2.6. *Assume R1–R6. Then*

$$\begin{aligned} &E_{m_n}\chi^2(\pi(\cdot|x^n),\pi(\cdot)) \\ &= \Big(\frac{n}{2\pi}\Big)^{q/2}E\Big[\exp\big(-\frac{1}{2}\chi^2_q\big)\Big]\int |\Lambda(\theta)|^{1/2}d\theta - n^{(q-2)/2}2^{(q+2)/2}\pi^{q/2} \\ &\quad \times E\Big[\exp\big(-\frac{1}{2}\chi^2_q\big)\Big]\int |\Lambda(\theta)|^{1/2}tr\Big\{\Lambda^{-1/2'}(\theta)\frac{\pi^{(2)}(\theta)}{\pi(\theta)}\Lambda^{-1/2}(\theta)\Big\}d\theta \\ &\quad + o(n^{(q-2)/2}). \end{aligned}$$

So, the reference prior for the EL under the Chi-square distance is

$$\pi(\cdot) = \arg\min_{\pi(\cdot)} \int |\Lambda(\theta)|^{1/2} tr[\Lambda^{-1/2'}(\theta) \frac{\pi^{(2)}(\theta)}{\pi(\theta)} \Lambda^{-1/2}(\theta)] d\theta,$$

subject to $\int \pi(\theta) d\theta = 1$.

Proof. As in Theorem 2.5, let $\pi^{(1)}(\theta) = \partial \pi(\theta)/\partial\theta$ and $\pi^{(2)}(\theta) = \partial^2 \pi(\theta)/[\partial\theta\partial\theta']$ and recall that $\int \theta' A \theta \phi(\theta|0, I_q) d\theta = tr(A)$. Now, when $p(x^n|\theta)$ defines the mode of convergence, Taylor expanding gives

$$\begin{aligned}
&\int \pi(\alpha) \exp\Big[\frac{1}{2} n(\hat{\theta}_n - \alpha)' \big\{\frac{1}{n}\sum_{i=1}^{n} l_i^{(2)}(\theta_n)\big\}(\hat{\theta}_n - \alpha)\Big] d\alpha \\
&\sim (2\pi/n)^{q/2} |\Lambda^{-1}(\theta)|^{1/2} \\
&\quad \times \int \Big\{\pi(\hat{\theta}_n) + (\alpha - \hat{\theta}_n)' \pi^{(1)}(\hat{\theta}_n) + \frac{1}{2}(\alpha - \hat{\theta}_n)' \pi^{(2)}(\theta_{2,n})(\alpha - \hat{\theta}_n)\Big\} \\
&\quad \times \phi(\alpha|\hat{\theta}_n, \Lambda^{-1}(\theta)/n) d\alpha \\
&= (2\pi/n)^{q/2} |\Lambda^{-1}(\theta)|^{1/2} \\
&\quad \times \int \Big\{\pi(\hat{\theta}_n) + \frac{1}{\sqrt{n}} \alpha' \Lambda^{-1/2'}(\theta) \pi^{(1)}(\hat{\theta}_n) + \frac{1}{2n} \alpha' \Lambda^{-1/2'}(\theta) \pi^{(2)}(\theta_{2,n}) \Lambda^{-1/2}(\theta) \alpha\Big\} \\
&\quad \times \phi(\alpha|0, I_q) d\alpha \\
&\sim (2\pi/n)^{q/2} |\Lambda^{-1}(\theta)|^{1/2} \pi(\theta) \Big[1 + \frac{1}{2n} tr\big\{\Lambda^{-1/2'}(\theta) \frac{\pi^{(2)}(\theta)}{\pi(\theta)} \Lambda^{-1/2}(\theta)\big\}\Big].
\end{aligned} \tag{2.3.17}$$

Using the inverse of (2.3.17), we have that $E_{m_n} \chi^2(\pi(\cdot|x^n), \pi(\cdot))$ equals

$$\begin{aligned}
&\int\int \pi(\theta) \frac{p^2(x^n|\theta)}{m(x^n)} d\mu(x^n) d\theta - 1 \\
&= \int\int \pi(\theta) p(x^n|\theta) \frac{\exp\Big[\frac{1}{2} n(\tilde{\theta}_n - \theta)' \big\{\frac{1}{n}\sum_{i=1}^{n} l_i^{(2)}(\theta_n)\big\}(\tilde{\theta}_n - \theta)\Big]}{\int \pi(\alpha) \exp\Big[\frac{1}{2} n(\tilde{\theta}_n - \alpha)' \big\{\frac{1}{n}\sum_{i=1}^{n} l_i^{(2)}(\theta_n)\big\}(\tilde{\theta}_n - \alpha)\Big] d\alpha} \\
&\quad \times d\mu(x^n) d\theta - 1 \\
&\sim \big(\frac{n}{2\pi}\big)^{q/2} \int\int |\Lambda(\theta)|^{1/2} \Big[1 - \frac{1}{2n} tr\big\{\Lambda^{-1/2'}(\theta) \frac{\pi^{(2)}(\theta)}{\pi(\theta)} \Lambda^{-1/2}(\theta)\big\}\Big] \\
&\quad \times \exp\Big[\frac{1}{2} n(\tilde{\theta}_n - \theta)' \big\{\frac{1}{n}\sum_{i=1}^{n} l_i^{(2)}(\tilde{\theta}_n)\big\}(\tilde{\theta}_n - \theta)\Big] p(x^n|\theta) d\mu(x^n) d\theta - 1 \\
&\sim \big(\frac{n}{2\pi}\big)^{q/2} E\big[\exp\big(-\frac{1}{2}\chi_q^2\big)\big] \\
&\quad \times \int |\Lambda(\theta)|^{1/2} \Big[1 - \frac{1}{2n} tr\big\{\Lambda^{-1/2'}(\theta) \frac{\pi^{(2)}(\theta)}{\pi(\theta)} \Lambda^{-1/2}(\theta)\big\}\Big] d\theta.
\end{aligned}$$

2.3.6 Discussion

It is seen that the reference prior for ELs under Hellinger is based on the reciprocal of the reference prior under relative entropy and that these differ from the reference prior under χ^2 which is hard to obtain explicitly. This is somewhat different from the treatment given in Ghosh, Mergel, and Liu (2009) who obtained the Jeffreys prior for all members of the power divergence family except the Chi-square distance. Here, it is only in the relative entropy case that the reference prior is based on the transformation that makes an efficient CAN estimator converge to $N(0, I_q)$. Nevertheless, the role of Jeffreys prior is roughly analogous to $\Lambda^{-1}(\theta) = (D'(\theta)\Omega^{-1}(\theta)D(\theta))^{-1}$.

An examination of the proof of all three theorems reveals a common structure: Approximate the ratio $p(x^n|\theta)/m(x^n)$ by a Laplace's method argument, take a function of the density ratio, and examine its limiting expectation using standard results and assumptions. Consequently, we conjecture that our basic technique extends to any Csiszár f-divergence, see Csiszár (1967), defined as $D_f(p||q) = E_p f(p/q)$ for some convex f where p and q are densities. The power divergence family (whose members often play a role in goodness-of-fit testing) is contained in this class. Many reference priors with different interpretations can be generated this way. Their divergence from the least favorable priors outside the relative entropy case means reference priors reflect a notion of minimal information rather than any decision-theoretic concept.

Acknowledgments: Yuan's works is partly supported by the National Center for Research Resources by NIH grant 2G12RR003048.

Chapter 3
Bayesian Decision Based Estimation and Predictive Inference

Shrinkage estimation is a traditional research topic in Bayesian analysis. The three sections in this chapter convincingly argue that this venerable topic remains a current research frontier with many open problems. The chapter starts with a review of the current state of research and concludes with an insightful discussion of a very specific form of shrinkage estimation arising in recent work on inference in gene-environment interaction studies.

3.1 Bayesian Shrinkage Estimation

William E. Strawderman

This section gives a decision theoretic account of Bayesian shrinkage estimation largely concentrating on multivariate normal location models with respect to squared error loss. There is considerable emphasis on Stein-type shrinkage and minimaxity in three and higher dimensions. Specifically, consider the problem of estimating the mean, θ of a p-dimensional multivariate normal distribution with covariance matrix equal to a multiple of the identity,

$$X \sim N_p(\theta, \sigma^2 I), \tag{3.1.1}$$

and loss equal to the sum of squared errors loss

$$(\theta, d) = ||d - \theta||^2 = \Sigma(d_i - \theta_i)^2. \tag{3.1.2}$$

Much of the paper will deal with this most simple version of the problem.

Section 3.1.1 is devoted to giving intuition into why shrinkage seems desirable from both a frequentist and a Bayesian perspective. In particular, we show that a particular form of shrinkage estimator (the James-Stein estimator) (James and Stein, 1961) arises quite naturally from both frequentist and Bayesian perspectives and

M.-H. Chen et al. (eds.), *Frontiers of Statistical Decision Making and Bayesian Analysis*, DOI 10.1007/978-1-4419-6944-6_3,

does not rely on normality of the underlying population. Section 3.1.2 is devoted to the theory of minimax Bayesian shrinkage for the simplest model mentioned above and lays the groundwork for results in more complex models. We give conditions under which various forms of (empirical, generalized, hierarchical, and pseudo-) Bayes estimators improve on the "usual" estimator X, which already has a nice collection of associated optimality properties (such as minimaxity, best equivariance, and best unbiasedness). Section 3.1.3 studies some of the details for the more complex case of a general covariance matrix and general quadratic loss, and section 3.1.4 gives some concluding remarks. Much of the development follows Strawderman (2003).

3.1.1 Some Intuition into Shrinkage Estimation

Suppose we have a p-dimensional vector of data, $X = (X_1, \ldots, X_p)'$, such that $E[X_i] = \theta_i$, and $cov(X) = \sigma^2 I_p$, where σ^2 is known. Suppose also that X has a known density, $f(x)$, with respect to Lebesgue measure ($f(\cdot)$ is not necessarily assumed to be normal). Assume we want to estimate the mean vector, $\theta = (\theta_1, \ldots, \theta_p)'$ and that the loss is equal to the sum of squared errors loss ($L(\theta, d) = ||d - \theta||^2$).

The vector X is an obvious unbiased estimator of θ that we will occasionally refer to as the "usual" estimator. As a first frequentist pass at understanding why we might want to shrink X, consider the class of linear estimators of the form bX. Is there a best value for b? A simple calculation shows that the risk,

$$R(\theta, bX) = E[L(\theta, bX)] = E[||bX - \theta||^2] = pb^2\sigma^2 + (b-1)^2||\theta||^2,$$

and that the value which minimizes this expression is

$$b(||\theta||^2) = ||\theta||^2/(p\sigma^2 + ||\theta||^2) = (1 - p\sigma/(p\sigma^2 + ||\theta||^2).$$

Unfortunately this optimizing b depends on the unknown mean, so it does not seem particularly useful. However, upon realizing that $E[||X||^2] = p\sigma^2 + ||\theta||^2$, it appears that we may estimate the optimal b by the quantity $(1 - p\sigma^2/||X||^2)$, and hence the optimal "estimator," bX, by

$$\delta(X) = (1 - p\sigma^2/||X||^2)X. \tag{3.1.3}$$

If "p" is replaced by "$p-2$" in (3.1.3), the resulting estimator is known as the James-Stein estimator. One striking aspect of the above intuitive argument is that it does not depend on an underlying assumption of normality.

There is an equally general empirical Bayes argument (see Strawderman, 1992) that arrives at the same estimator. We assume the above model for X with density $f(x)$. Also assume that the prior distribution of θ is given by $\pi(\theta) = f^{*n}(\theta)$, the n fold convolution of $f(\cdot)$ with itself. Hence the prior distribution of θ can be repre-

sented as the distribution of a sum of n iid variables $u_i, i = 1, \ldots, n$, where each u is distributed as $f(u)$.

Also, the distribution of $u_0 = (X - \theta)$ has this same distribution and is independent of each of the other u's. Then the Bayes estimator can be represented as

$$\delta(X) = E[\theta|X] = E[\theta|X - \theta + \theta] = E\Big[\sum_{i=1}^{n} u_i \Big| \sum_{i=0}^{n} u_i\Big],$$

and hence by exchangeability,

$$\delta(X) = nE\Big[u_j \Big| \sum_{i=0}^{n} u_i\Big] = \frac{nE\Big[\sum\limits_{i=0}^{n} u_i \Big| \sum_{i=0}^{n} u_i\Big]}{n+1} = \frac{nE[X|X]}{n+1} = \frac{nX}{n+1},$$

or equivalently,

$$\delta(X) = E[\theta|X] = \Big(1 - \frac{1}{n+1}\Big)X.$$

To find an empirical Bayes version of this estimator, suppose that n is unknown. We may estimate it from the marginal distribution of X, which has the same distribution as $X - \theta + \theta = \sum_{i=0}^{n} u_i$. In particular, since $E[u_i] = 0$ and $\text{cov}(u_i) = \sigma^2 I$, $E[||u_i||^2] = p\sigma^2$ and $E_\theta[||X||^2] = E\Big[||\sum_{i=0}^{n} u_i||^2\Big] = \sum_{i=0}^{n} E\Big[\big\|u_i\big\|^2\Big] = (n+1)p\sigma^2$. Therefore, $n+1$ can be estimated by $||X||^2/p\sigma^2$. Substituting this estimator of $n+1$ in the expression for the Bayes estimator, we have an empirical Bayes estimator

$$\hat{\delta}(X) = (1 - p\sigma^2/||X||^2)X,$$

which is again (3.1.3), the James-Stein estimator save for the substitution of p for $p-2$.

Recall again, that in both of the above developments, the only assumptions were that $E_\theta(X) = \theta$, and Cov(X)=$\sigma^2 I$ (the assumption of a density, $f(\cdot)$, was just a notational convenience to give the prior density as a convolution).

It seems worth remarking also that the above development is closely related to a standard argument giving the James-Stein estimator as an Empirical Bayes estimator in the normal case, except that in the above development, the variance of the prior distribution is restricted to be an integer multiple of variance of the individual X's. In the case of a normal distribution, the prior distribution has covariance matrix $\tau^2 I$, the resulting Bayes estimator is $(1 - \sigma^2/(\sigma^2 + \tau^2))X$. The marginal distribution of X is p-variate normal with mean vector 0 and covariance matrix $(\sigma^2 + \tau^2)I$. Hence (based on the marginal distribution of X) the best unbiased estimator of $1/(\sigma^2 + \tau^2)$ is $(p-2)/||X||^2$, and an empirical Bayes estimator is given by $(1 - (p-2)\sigma^2/||X||^2)X$, which is the classical James-Stein estimator.

The Stein-type estimator thus appears intuitively, at least, to be a reasonable estimator in a general location problem, as an approximation either to the best linear estimator or a conjugate prior Bayes estimator. We'll show, in the next sec-

tion, that the James-Stein estimator of the form $(1 - a\sigma^2/||X||^2)X$, is minimax for $0 \le a \le 2(p-2)$, provided $p \ge 3$.

3.1.2 Some Theory for the Normal Case with Covariance $\sigma^2 I$

3.1.2.1 Minimaxity and Unbiased Estimators of Risk

In this section we indicate basic developments for the relatively simple case of a normal distribution with covariance matrix equal to $\sigma^2 I$. We will develop this further for more complex covariance structures in Section 4. We use the following notation,

$$\nabla' g = \sum \nabla_i g_i, \text{ where } \nabla_i = \frac{\partial}{\partial X_i}.$$

Consider model (3.1.1) with loss (3.1.2). Stein's Lemma (Lemma 3.2) is a basic tool in evaluating the risk function of a (nearly) general estimator when the loss function is quadratic. A simple one dimensional version of the lemma is the following whose proof is trivial by using integration by parts.

Lemma 3.1. *Let $X \sim N(\theta, \sigma^2)$ and let $h(X)$ be a continuously differentiable function such that $E[h'(X)]$ is finite. Then $E[(X-\theta)h(X)] = \sigma^2 E[h'(X)]$.*

The multivariate extension given next follows from a one variable at a time application of Lemma 3.1 if stringent conditions are placed on h$(\cdot)$, but it follows under the much weaker stated conditions using Stoke's theorem. Stein's (1981) development does not use Stokes theorem but is essentially equivalent.

Lemma 3.2. *Suppose X has distribution (3.1.1). If $h : R^p \to R^p$ is an almost differentiable function with $E_\theta[\nabla' h(X)] < \infty$, then*

$$E_\theta[(X-\theta)' h(X)] = \sigma^2 E_\theta[\nabla' h(X)].$$

An expression for the risk function follows easily from this result.

Theorem 3.1. *Suppose $X \sim N_p(\theta, \sigma^2 I)$, and consider the estimator $\delta(X) = X + \sigma^2 g(X)$ of θ under loss (3.1.2). Assume $g : R^p \to R^p$ is an almost differentiable function for which $E_\theta[\nabla' g(X)] < \infty$, then*

$$R(\theta, \delta) = p\sigma^2 + \sigma^4 E_\theta[||g(X)||^2 + 2\nabla' g(X)].$$

Proof. $R(\theta,\delta) = E_\theta[||X + \sigma^2 g(X) - \theta||^2] = E_\theta[||X-\theta||^2 + \sigma^4 ||g(X)||^2 + 2\sigma^2 (X - \theta)' g(X)]$, and so, by using Lemma 3.2 on the final term,

$$R(\theta, \delta) = p\sigma^2 + \sigma^4 E_\theta[||g(X)||^2 + 2\nabla' g(X)],$$

which completes the proof. □

The next result also follows immediately from the Theorem 3.1.

Corollary 3.1. *Under the conditions of Theorem 3.1,*

(a) $p\sigma^2 + \sigma^4(||g(X)||^2 + 2\nabla' g(X))$ *is the UMVUE of the risk function of* (X)*, and*
(b)A sufficient condition for $\delta(X)$*, to dominate* X *(and hence be minimax) is that* $||g(X)||^2 + 2\nabla' g(X) \leq 0$ *a.e., and that strict inequality hold on a set of positive measure.*

A very useful version of the result follows easily for estimators of Baranchik's (1964) form, which include the James-Stein class.

Corollary 3.2. *Let* $X \sim N_p(\theta, \sigma^2 I)(p \geq 3)$*, and* $\delta(X) = X + \sigma^2 g(X)$ *where* $g(X) = -\left(\frac{r(\|X\|^2)}{\|X\|^2}\right) X$*, for* $r(\cdot)$ *a non-decreasing function such that* $0 < r(\cdot) \leq 2(p-2)$*, and assume there is strict inequality on a set of positive measure. Then, for quadratic loss (3.1.2),* $\delta(X)$ *has smaller risk than* X *and is thus minimax.*

Proof. To prove this result, from Corollary 3.1, we need only show that $[||g(X)||^2 + 2\nabla' g(X)] \leq 0$. Notice that

$$||g(X)||^2 + 2\nabla' g(X) = \left\| -\left(\frac{r(\|X\|^2)}{\|X\|^2}\right) X \right\|^2 + 2\nabla' \left(\frac{-r(\|X\|^2) X}{\|X\|^2}\right).$$

Now

$$2\nabla' \left(\frac{r(\|X\|^2) X}{\|X\|^2}\right) = 2(p-2)\left(\frac{r(\|X\|^2)}{\|X\|^2}\right) + 4\left(\frac{r'(\|X\|^2)}{\|X\|^2}\right) \geq 2(p-2) E_\theta \left[\frac{r(\|X\|^2)}{\|X\|^2}\right]$$

since $r(\cdot)$ is a nondecreasing function. Further, since $0 < r(\cdot) \leq 2(p-2)$,

$$\left\| -\left(\frac{r(\|X\|^2)}{\|X\|^2}\right) X \right\|^2 = \frac{r^2(\|X\|^2)}{\|X\|^2} \leq 2(p-2)\left(\frac{r(\|X\|^2)}{\|X\|^2}\right).$$

Thus

$$||g(X)||^2 + 2\nabla' g(X) \leq 2(p-2) E_\theta \left(\frac{r(\|X\|^2)}{\|X\|^2}\right) - 2(p-2) E_\theta \left[\frac{r(\|X\|^2)}{\|X\|^2}\right] = 0,$$

and the result follows, since the inequality is strict on a set of positive measure. □

If we take the function $r(t) \equiv a$, then the estimator is a James-Stein type estimator and is seen to be minimax for $0 \leq a \leq 2(p-2)$. A closer look at the proof for this case also gives the risk (since $r'(t)$ is 0) as $p\sigma^2 + a(a - 2(p-2))\sigma^4 E_\theta[1/||X||^2]$, and hence $a = (p-2)$ is the uniformly best choice (since $a = (p-2)$ minimizes $a(a - 2(p-2))$). Further, a simple calculation, shows that the risk at $\theta = 0$, is equal to $2\sigma^2$ (compared to $p\sigma^2$ for the estimator X) regardless of the dimension, p, provided that $p \geq 3$. Hence a substantial savings in risk is possible in a neighborhood of 0.

3.1.2.2 Bayes Minimax Shrinkage Estimators

Under the model (3.1.1) and loss (3.1.2), a (generalized) Bayes estimator with respect to a prior distribution $\pi(\theta)$ on θ is given by the posterior expectation of $\theta, E[\theta|X]$, which has the following representation:

$$\delta(X) = E[\theta|X] = X + \sigma^2 E[(\theta - X)/\sigma^2|X] = X + \sigma^2 g(X),$$

where $g(X)$ is given by

$$g(X) = \frac{\int \{(\theta - X)/\sigma^2\} \exp\{-(x-\theta)^2/2\sigma^2\}\pi(\theta)d\theta}{\int \exp\{-(x-\theta)^2/2\sigma^2\}\pi(\theta)d\theta} = \nabla m(X)/m(X),$$

and $m(X) = \int \exp\{-(x-\theta)^2/2\sigma^2\}\pi(\theta)d\theta/(2\pi\sigma)^{p/2}$ is the marginal distribution of X.

We will variously refer to such estimators of the form

$$\delta(X) = X + \sigma^2 \nabla m(X)/m(X) \tag{3.1.4}$$

as (proper) Bayes, or generalized Bayes according as $m(X)$ arises from a proper or generalized prior $\pi(\theta)$. We will use the term Pseudo-Bayes (see, e.g., Bock, 1988) if the function $m(X)$ is (weakly) differentiable, but may not be the marginal corresponding to any (proper or generalized) prior distribution. For example, the James-Stein type estimators are Pseudo-Bayes in this sense, corresponding to $m(X) = \|X\|^{-2a}$, where $a = (p-2)/2$ corresponds to the James-Stein estimator itself.

The next result is a basic minimaxity result for such estimators.

Theorem 3.2. *Suppose $X \sim N_p(\theta, \sigma^2 I)$, and consider the estimator (3.1.4) of θ under loss (3.1.2). The risk function (provided $E[||g(X)||^2] < \infty$) is given by*

$$R(\theta, \delta) = p\sigma^2 + \sigma^4 E\left[\frac{2m(X)\nabla^2 m(X) - ||\nabla m(X)||^2}{(m(X))^2}\right] = p\sigma^2 + 4\sigma^4 E\left[\frac{\nabla^2 (m(X))^{1/2}}{(m(X))^{1/2}}\right].$$

The proof of the first expression is a straightforward application of Theorem 3.1 to the estimator of the form (3.1.4). The second expression follows from the first by direct calculation.

An immediate and useful corollary is the following.

Corollary 3.3. *An estimator of finite risk of the form (3.1.4) is minimax (and dominates X) provided either*

(a) $m(X)$ is superharmonic (i.e., $\nabla^2 m(X) \leq 0$) (and strictly superharmonic on a set of positive measure),

(b) $\mathrm{m(X)}^{1/2}$ is superharmonic (and strictly superharmonic on a set of positive measure), or

(c) If the estimator is generalized Bayes and $\pi(\theta)$ is superharmonic (and strictly superharmonic on a set of positive measure).

Further, (c) implies (a) implies (b).

Proof. The proof of (a) and (b), as well as the implication that (a) implies (b), is immediate from Theorem 3.2. Also the proof of (c), and that (c) implies (a), follows from the fact that the average of superharmonic functions is superharmonic ($m(X)$ is the average of $\pi(\theta)$ with respect to a $N_p(X, \sigma^2 I)$ distribution). □

It is interesting to note that estimators of the James-Stein type (corresponding to $m(X) = ||X||^{-2a}$) are shown to be minimax for $0 < a \leq (p-2)$ using part (a), and for $0 < a \leq 2(p-2)$ using part (b). This follows since, as can easily be checked, $m(X) = ||X||^{-2a}$ is superharmonic for $0 \leq a \leq (p-2)$. Hence part (b), although more complicated, is substantially more powerful than part (a).

Incidentally, the inclusion of $a = 0$ in the range of values for minimaxity is justified since $a = 0$ corresponds to the estimator X itself, which is minimax, but of course does not (strictly) dominate itself. Interestingly, the same holds true of the inclusion of $a = 2(p-2)$ for minimaxity, but not domination over X. This is so since X and the James-Stein estimator with $a = 2(p-2)$ have the same risk function (e.g., see the discussion after Corollary 3.2).

While Theorem 3.2 gives a quite general result for minimaxity of Bayes estimators, it is sometimes (often) more convenient to use Corollary 3.2 directly to prove minimaxity of (generalized, proper, or pseudo-) Bayes estimators. As an example, consider the following class of hierarchical generalized and proper Bayes minimax estimators.

Example 3.1. (Strawderman, 1971). Suppose the prior distribution has the two stage prior.

First stage: $\theta|\lambda \sim N_p(0, \{(1-\lambda)/\lambda\}\sigma^2 I)$;
Second stage: $\lambda \sim (1+b)\lambda^b$ for $0 < \lambda < 1$.

The proper (if $b > -1$) or generalized Bayes estimator is given by

$$\delta_\pi(X) = E(\theta|X) = E[E(\theta|X,\lambda)|X] = \{1 - E(\lambda|X)\}X.$$

The second equality follows since the first stage conditional distribution of θ given X and λ is $N_p((1-\lambda)X, (1-\lambda)\sigma^2 I)$ by a standard conjugate Bayes calculation. A direct way to calculate the Bayes estimator $\delta_\pi(X)$ is to use the expression (3.1.4).

Note that the distribution of X conditional on λ is equal to the distribution of $(X-\theta)+\theta$ conditional on λ or $N_p(0(\sigma^2/\lambda)I)$. Hence $m(x) \propto \int_0^1 \lambda^{\frac{p}{2}+b} \exp\{-\frac{\lambda||x||^2}{2\sigma^2}\}d\lambda$, and so

$$\delta_\pi(X) = X + \sigma^2 \frac{\nabla m(X)}{m(X)} = \left\{ 1 - \sigma^2 \frac{\int_0^1 \lambda^{\frac{p}{2}+b+1} \exp(-\frac{\lambda||x||^2}{2\sigma^2})d\lambda}{\sigma^2 \int_0^1 \lambda^{\frac{p}{2}+b} \exp(-\frac{\lambda||x||^2}{2\sigma^2})d\lambda} \right\} X. \quad (3.1.5)$$

The following result gives the conditions under which the above Bayes estimator is minimax.

Theorem 3.3. *(a) For $p \geq 3$ the estimator (3.1.5) is generalized Bayes and minimax provided $-\frac{(p+2)}{2} < b \leq \frac{(p-6)}{2}$. (b) For $p \geq 5$ the estimator (3.1.5) is proper Bayes and minimax provided $-1 < b \leq \frac{(p-6)}{2}$.*

Proof. The estimator (3.1.5) is of the Baranchik form as in Corollary 3.2 with

$$r(||X||^2) = (||X||^2/\sigma^2)\left\{ \frac{\int_0^1 \lambda^{\frac{p}{2}+b+1} \exp\left(-\frac{\lambda\|x\|^2}{2\sigma^2}\right) d\lambda}{\int_0^1 \lambda^{\frac{p}{2}+b} \exp\left(-\frac{\lambda\|x\|^2}{2\sigma^2}\right) d\lambda} \right\}.$$

Note that $p/2+b > -1$ is required for the integrals to exist, and that this is the first inequality in part (a), i.e., is a condition for the estimator to be generalized Bayes. Note too, that the first inequality in part (b) is the condition for the estimator to be proper Bayes.

Integration by parts in the numerator gives

$$\begin{aligned} r(||X||^2) &= p+2+2b - \frac{2\exp\left(-\frac{\|x\|^2}{2\sigma^2}\right)}{\int_0^1 \lambda^{\frac{p}{2}+b} \exp\left(-\frac{\lambda\|x\|^2}{2\sigma^2}\right) d\lambda} \\ &= p+2+2b - \frac{2}{\int_0^1 \lambda^{\frac{p}{2}+b} \exp\left\{\frac{(1-\lambda)\|x\|^2}{2\sigma^2}\right\} d\lambda}. \end{aligned}$$

It is apparent that $r(t) \leq (p+2+2b)$ and is non increasing. Hence the estimator (3.1.5) is minimax by Corollary 3.2 provided $0 < p+2+2b \leq 2(p-2)$, which is equivalent to the second inequality in both (a) and (b). Also, the inequality in (a) requires that $p \geq 3$ and in (b), that $p \geq 5$. This completes the proof. □

Strawderman (1972) showed that no proper Bayes minimax estimators exist for $p < 5$. The following result extends the above class of Bayes minimax estimators to include shrinkage functions that are not necessarily monotone. It is essentially the main result in Fourdrinier, Strawderman, and Wells (1998) (FSW).

Theorem 3.4. *Suppose θ has a prior distribution with the following hierarchical structure: $\theta|\lambda \sim N_p(0, \sigma^2(1-\lambda)\lambda^{-1}I)$ and, $\lambda \sim h(\lambda), 0 < \lambda < 1$. Then $\sqrt{m(X)}$ is superharmonic (and hence $\delta_\Pi(X)$ is minimax) provided $h(\lambda)$ satisfies $\frac{\lambda h'(\lambda)}{h(\lambda)} = \ell_1(\lambda) + \ell_2(\lambda)$, where $\ell_1(\lambda) \leq A$ is nonincreasing in λ, $0 \leq \ell_2(\lambda) \leq B$, $\frac{1}{2}A + B \leq \frac{(p-6)}{4}$, $\lim_{\lambda\to 0} \lambda^{p/2} h(\lambda) = 0$, and $\lim_{\lambda\to 1} h(\lambda) < \infty$. Further, the estimator is proper Bayes if $h()$ is integrable.*

An interesting and useful class of priors to which this result applies are certain scaled multivariate-t priors. See FSW for the proof and further examples.

George (1986a) studied *multiple shrinkage estimators*, for example, estimators for which there are several possible points $v_i, i = 1, 2, \ldots k$, towards which to shrink. The goal is to shrink "adaptively" toward one of the points so that the resulting procedure will be minimax. Again, using the fact that mixtures of superharmonic functions are superharmonic the following version of one of George's results follows immediately from Corollary 3.3.

Theorem 3.5. *Suppose $\pi_\alpha(\theta)$, $\alpha \in A$, is a collection of superharmonic (generalized) priors with corresponding marginals $m_\alpha(X)$. Let $\lambda(\alpha)$ be any finite mixing distribution on $\alpha \in A$. Then*

(a)$\pi(\theta) = \int \pi_\alpha(\theta)h(\alpha)d\alpha$ is superharmonic with corresponding superharmonic marginal $m(x) = \int m_\alpha(X)h(\alpha)d\alpha$.
(b)The resulting generalized (or pseudo-, if $m_\alpha(X)$ are given but don't correspond to a prior) Bayes estimator is minimax.

Example 3.2. Suppose $v_i, i = 1,2,\ldots k$ are vectors in R^p, and $m_i(X) = \frac{1}{\|X-v_i\|^{2b}}$ for $0 < b < (p-2)/2$. Let $m(X) = \frac{1}{k}\sum_{i=1}^{k} m_i(X)$. Then $m_i(X)$ and $m(X)$ are superharmonic and the resulting pseudo-Bayes estimator given by

$$\delta_m(X) = X - \frac{2b\sigma^2 \sum_{i=1}^k \{(X-v_i)/\|X-v_i\|^{2b+2}\}}{\sum_{i=1}^k (1/\|X-v_i\|^{2b})}$$

is minimax. This estimator "adaptively" shrinks X toward the "closest" v_i or alternatively, $\delta_m(X)$ is a weighted combination of James-Stein like estimators shrinking toward the v_i with greater weights $1/||X-v_i||^{2b}$ put on the v_i closest to X. While this estimator is pseudo-Bayes, a similar generalized Bayes is easily given whereby $\pi(\theta) = \sum_1^k (1/\|\theta - v_i\|^{2b})$ and $0 \le b \le (p-2)/2$.

It is interesting to note that the result as given is confined to priors or marginals which are superharmonic $(0 \le b \le (p-2)/2)$, and not "square-root" superharmonic $(0 \le b \le p-2)$. This is because it need not be that a mixture of square-root superharmonic functions is itself square root superharmonic. As FSW show, a superharmonic marginal or prior cannot be proper, and hence the above result doesn't apply to mixtures of proper priors.

3.1.3 Results for Known Σ and General Quadratic Loss

3.1.3.1 Results for the Diagonal Case

Much of this section is based on the discussion in Strawderman (2003). We begin with a discussion of the multivariate normal case where Σ is diagonal. Let

$$X \sim N_p(\theta, \Sigma), \tag{3.1.6}$$

where Σ is diagonal, $\Sigma = diag(\sigma_1^2, \ldots, \sigma_p^2)$ and loss equal to a weighted sum of squared errors loss

$$L(\theta, \delta) = (\delta - \theta)' D(\delta - \theta) = \Sigma(\delta_i - \theta_i)^2 d_i \tag{3.1.7}$$

The results of the previous section extend by the use of Stein's Lemma in a straightforward way to give the following basic theorem.

Theorem 3.6. *Let X have the distribution (3.1.6) and let the loss be given by (3.1.7).*

(a)If $\delta(X) = X + \Sigma g(X)$, where $g(X)$ is weakly differentiable and $E||g||^2 < \infty$, then the risk of δ is

$$R(\delta,\theta) = E_\theta[(\delta-\theta)'Q(\delta-\theta)] = tr(\Sigma D) + E_\theta\Big[\sum_{i=1}^{p}\sigma_i^4 d_i\Big\{g_i^2(X) + 2\frac{\partial g_i(X)}{\partial X_i}\Big\}\Big].$$

(b)If $\theta \sim \pi(\theta)$, then the Bayes estimator of θ is $\delta_\Pi(X) = X + \Sigma\frac{\nabla m(X)}{m(X)}$, where $m(X)$ is the marginal distribution of X.

(c)If $\theta \sim \pi(\theta)$, then the risk of a proper (generalized, pseudo-) Bayes estimator of the form $\delta_m(X) = X + \Sigma\frac{\nabla m(X)}{m(X)}$ is given by

$$\begin{aligned} R(\delta_m,\theta) &= tr(\Sigma D) + E_\theta\left[\frac{2m(X)\sum_{i=1}^{p}\sigma_i^4 d_i\frac{\partial m^2(X)}{\partial^2 X_i}}{m^2(X)} - \frac{\sum_{i=1}^{p}\sigma_i^4 d_i\Big(\frac{\partial m^2(X)}{\partial^2 X_i}\Big)^2}{m^2(X)}\right] \\ &= tr(\Sigma D) + 4E_\theta\Big[\frac{\sum_{i=1}^{p}\sigma_i^4 d_i\partial^2\sqrt{m(X)}/\partial^2 X_i}{\sqrt{m(X)}}\Big]. \end{aligned}$$

(d)If $\frac{\sum_{i=1}^{p}\sigma_i^4 d_i\partial^2\sqrt{m(X)}/\partial^2 X_i}{\sqrt{m(X)}}$ is non-positive, the proper (generalized, pseudo-) Bayes $\delta_m(X)$ is minimax.

The proof follows closely that of corresponding results in Section 3. The result is basically from Stein (1981).

A key observation that allows us to construct Bayes minimax procedures for this situation, based on the procedures for the case $\Sigma = D = I$, is the following.

Lemma 3.3. *Suppose $\eta(X)$ is such that $\nabla^2\eta(X) = \sum_{i=1}^{p}\partial^2\eta(X)/\partial^2 X_i^2 \le 0$ (i.e. $\eta(X)$ is superharmonic) then $\eta^*(X) = \eta(\Sigma^{-1}D^{-1/2}X)$ is such that*

$$\sum_{i=1}^{p}\sigma_i^4 d_i\partial^2\eta^*(X)/\partial^2 X_i \le 0.$$

The proof is a straightforward calculation. Details can be found in Strawderman (2003).

Note, also, that for any scalar, a, if $\eta(X)$ is superharmonic, then so is $\eta(aX)$. This all leads to the following main result.

Theorem 3.7. *Suppose X has distribution (3.1.6) and loss is given by (3.1.7).*

(a)Suppose $\sqrt{m(X)}$ is superharmonic ($m(X)$ is a proper, generalized, or pseudo-marginal for the case $\Sigma = D = I$). Then

$$\delta_m(X) = X + \Sigma\left(\frac{\nabla m(\Sigma^{-1}D^{-1/2}X)}{m(\Sigma^{-1}D^{-1/2}X)}\right)$$

is a minimax estimator.

(b)If $\sqrt{m(\|X\|^2)}$ *is spherically symmetric and superharmonic, then*

$$\delta_m(X) = X + \frac{2m'(X'\Sigma^{-1}D^{-1}\Sigma^{-1}X)D^{-1}\Sigma^{-1}X}{m(X'\Sigma^{-1}D^{-1}\Sigma^{-1}X)}$$

is minimax.

(c)Suppose $\pi(\theta)$ *has the hierarchical structure* $\theta|\lambda \sim N_p(0,A_\lambda)$ *for* $\lambda \sim h(\lambda),\ 0 < \lambda < 1$, *where* $A_\lambda = (c/\lambda)\Sigma D\Sigma - \Sigma$ *and* c *is such that* A_1 *is positive definite and* $h(\lambda)$ *satisfies the conditions Theorem 3.3. Then*

$$\delta_\pi(X) = X + \Sigma\frac{\nabla m(X)}{m(X)}$$

is minimax.

(d)Suppose $m_i(X)$, $i = 1,2,\ldots,k$, *are superharmonic, then the multiple shrinkage estimator*

$$\delta_m(X) = X + \Sigma\left\{\frac{\sum_{i=1}^k \nabla m_i(\Sigma^{-1}D^{-1/2}X)}{\sum_{i=1}^k m_i(\Sigma^{-1}D^{-1/2}X)}\right\}$$

is a minimax multiple shrinkage estimator.

Proof.

(a)This follows directly from Theorem 3.6 parts (c) and (d) and Lemma 3.3.

(b)This follows from part (a) and Theorem 3.6 part (b) with a straightforward calculation.

(c)First note that $\theta|\lambda \sim N_p(0,A_\lambda)$ and $X-\theta|\lambda \sim N_p(0,\Sigma)$. Thus, $X-\theta$ and θ are therefore conditionally independent given λ. Hence $X|\lambda \sim N_p(0,A_\lambda+\Sigma)$. It follows that

$$m(X) \propto \int_0^1 \lambda^{p/2}\exp\left\{-\frac{\lambda}{c}(X'\Sigma^{-1}D^{-1}\Sigma^{-1}X)\right\}h(\lambda)d\lambda$$

but $m(X) = \eta(X'\Sigma^{-1}D^{-1}\Sigma^{-1}X/c)$, where $\sqrt{\eta(X'X)}$ is superharmonic by Theorem 3.4. Hence by part (b) $\delta_\pi(X)$ is minimax (and proper or generalized Bayes depending on whether $h(\lambda)$ is integrable or not).

(d)Since superharmonicity of $\eta(X)$ implies that of $\sqrt{\eta(X)}$, part (d) follows from part (a) and superharmonicity of mixtures of superharmonic functions. □

Example 3.3. Pseudo-Bayes minimax estimators. When $\Sigma = D = \sigma^2 I$, we saw in Section 3.1.2 that by choosing $m(X) = \frac{1}{\|X\|^{2b}}$, the pseudo-Bayes estimator was the James-Stein estimator $\delta_m(X) = (1 - \frac{2b\sigma^2}{\|X\|^2})X$. It now follows from this and part (b) of Theorem 3.7 that $m(X'\Sigma^{-1}D^{-1}\Sigma^{-1}X) = (1/X'\Sigma^{-1}D^{-1}\Sigma^{-1}X)^b$ has associated

with it the pseudo-Bayes estimator $\delta_m(X) = (1 - \frac{2bD^{-1}\Sigma^{-1}}{X'\Sigma^{-1}D^{-1}\Sigma^{-1}X}X$. This estimator is minimax for $0 < b \leq 2(p-2)$.

Example 3.4. Hierarchical proper Bayes minimax estimator. As suggested by Berger (1976) suppose the prior distribution has the hierarchical structure $\theta|\lambda \sim N_p(0, A_\lambda)$, where $A_\lambda = c\Sigma D\Sigma - \Sigma$, $c > 1/\min(\sigma_i^2 d_i)$ and $h(\lambda) = (1+b)\lambda^b$, $0 < \lambda < 1$ for $-1 < b \leq \frac{(p-6)}{2}$. The resulting proper Bayes estimator will be minimax for $p \geq 5$ by Theorem 3.7 part (c) and Example 3.1. For $p \geq 3$ the estimator $\delta_\pi(X)$ given in part (c) of Theorem 3.7 is a generalized Bayes minimax provided $-\frac{(p+2)}{2} < b \leq \frac{(p-6)}{2}$.

It can be shown to be admissible if the lower bound is replaced by -2, by the results of Brown (1971) (See also Berger and Strawderman (1996) and Kubokawa and Strawderman (2007)).

Example 3.5. Multiple shrinkage minimax estimator. It follows from Example 3.3 and Theorem 3.7 that $m(X) = \sum_{i=1}^k \left\{\frac{1}{(X-v_i)'\Sigma^{-1}D^{-1}\Sigma^{-1}(X-v_i)}\right\}^b$ satisfies the conditions of Theorem 3.7 (d) for $0 < b \leq (p-2)/2$ and hence that

$$\delta_m(X) = X - \frac{2b\sum_{i=1}^k \left\{[D^{-1}\Sigma^{-1}(X-v_i)]/[(X-v_i)'\Sigma^{-1}D^{-1}\Sigma^{-1}(X-v_i)]^{b+1}\right\}}{\sum_{i=1}^k \left\{1/[(X-v_i)'\Sigma^{-1}D^{-1}\Sigma^{-1}(X-v_i)]^b\right\}}$$

is a minimax multiple shrinkage (pseudo-Bayes) estimator.

If, as in Example 3.2, we used the generalized prior

$$\pi(\theta) = \sum_{i=1}^k \left\{\frac{1}{(\theta-v_i)'\Sigma^{-1}D^{-1}\Sigma^{-1}(\theta-v_i)}\right\}^b,$$

the resulting generalized Bayes (as opposed to pseudo-Bayes) estimators would be minimax for $0 < b \leq (p-2)/2$.

3.1.3.2 General Σ and General Quadratic Loss

Here we generalize the above results to the case of

$$X \sim N_p(\theta, \Sigma), \tag{3.1.8}$$

where Σ is a general positive definite covariance matrix, and

$$L(\theta, \delta) = (\delta - \theta)'Q(\delta - \theta), \tag{3.1.9}$$

where Q is a general positive definite matrix. We will see that this case can be reduced to the canonical form $\Sigma = I$ and $Q = diag(d_1, d_2 \ldots . d_p) = D$. We continue to follow the development in Strawderman (2003).

The following well known fact will be used repeatedly to obtain the desired generalization.

Lemma 3.4. *For any pair of positive definite matrices, Σ and Q, there exits a nonsingular matrix A such that $A\Sigma A' = I$ and $(A')^{-1}QA^{-1} = D$ where D is diagonal.*

Using this fact we can now present the canonical form of the estimation problem.

Theorem 3.8. *Let $X \sim N_p(\theta,\Sigma)$ and suppose that the loss is $L_1(\delta,\theta) = (\delta - \theta)'Q(\delta-\theta)$. Let A and D be as in Lemma 3.4, and let $Y = AX \sim N_p(v,I)$, where $v = A\theta$, and $L_2(\delta,v) = (\delta - v)'D(\delta - v)$.*

(a) If $\delta_1(X)$ is an estimator with risk function $R_1(\delta_1,\theta) = E_\theta L_1(\delta_1(X),\theta)$, then the estimator $\delta_2(Y) = A\delta_1(A^{-1}Y)$ has risk function $R_2(\delta_2,v) = R_1(\delta_1,\theta) = E_\theta L_2(\delta_2(Y),v)$.

(b) $\delta_1(X)$ is proper or generalized Bayes with respect to the proper prior distribution $\pi_1(\theta)$ (or pseudo-Bayes with respect to the pseudo-marginal $m_1(X)$) under loss L_1 if and only if $\delta_2(Y) = A\delta_1(A^{-1}Y)$ is proper or generalized Bayes with respect to $\pi_2(v) = \pi_1(A^{-1}v)$ (or pseudo-Bayes with respect to the pseudo-marginal $m_2(Y) = m_1(A^{-1}Y)$).

(c) $\delta_1(X)$ is admissible (or minimax or dominates $\delta_1^*(X)$) under L_1 if and only if $\delta_2(Y) = A\delta_1(A^{-1}Y)$ is admissible (or minimax or dominates $\delta_2^*(Y) = A\delta_1^*(A^{-1}Y)$ under L_2).

Proof.

(a) The risk function

$$\begin{aligned} R_2(\delta_2,v) &= E_\theta L_2[\delta_2(Y),v] = E_\theta[(\delta_2(Y)-v)'D(\delta_2(Y)-v)] \\ &= E_\theta[\{A\delta_1(A^{-1}(AX)) - A\theta\}'D\{A\delta_1(A^{-1}(AX)) - A\theta\}] \\ &= E_\theta[(\delta_1(X)-\theta)'A'DA(\delta_1(X)-\theta)] \\ &= E_\theta[(\delta_1(X)-\theta)'Q(\delta_1(X)-\theta)] = R_1(\delta_1,\theta). \end{aligned}$$

(b) The Bayes estimator for any quadratic loss is the posterior mean. Hence, since $\theta \sim \pi_1(\theta)$ and $v = A\theta \sim \pi_2(v) = \pi_1(A^{-1}v)$ (ignoring constants) then

$$\delta_2(Y) = E[v|Y] = E[A\theta|Y] = E[A\theta|AX] = A\,E[\theta|X] = A\,\delta_1(X) = A\delta_1(A^{-1}Y).$$

(c) This follows directly from part (a). □

Note: If $\Sigma^{1/2}$ is the positive definite square root of Σ and $A = P\Sigma^{-1/2}$ where P is orthogonal and diagonalizes $\Sigma^{1/2}Q\Sigma^{1/2}$, then this A and $D = P\Sigma^{1/2}Q\Sigma^{1/2}P'$ satisfy the requirements of the theorem.

Example 3.6. Proceeding as we did in Example 3.3 and applying Theorem 3.8,

$$m(X'\Sigma^{-1}Q^{-1}\Sigma^{-1}X) = (X'\Sigma^{-1}Q^{-1}\Sigma^{-1}X)^{-b}$$

has associated with it, for $0 < b \le 2(p-2)$, the pseudo-Bayes minimax James-Stein estimators is

$$\delta_m(X) = \left\{1 - \frac{2bQ^{-1}\Sigma^{-1}}{X'\Sigma^{-1}Q^{-1}\Sigma^{-1}X}\right\}X.$$

Generalizations of Example 3.4 to hierarchical Bayes minimax estimators and Example 3.5 to multiple shrinkage estimators are straightforward. We omit the details.

3.1.4 Conclusion and Extensions

I have attempted to present a development of much of what is known regarding Bayes minimax estimation of the mean vector of a multivariate normal mean vector under quadratic loss when the covariance is known. The presentation is designed to highlight the essential fact that the basic case is that of the identity covariance matrix and diagonal loss, and that general results flow with relative ease from that case. The approach has been more decision theoretic than purely Bayesian and puts substantial emphasis on minimaxity and Stein estimation. It has followed fairly closely the development in Strawderman (2003) (see also Brandwein and Strawderman, 2005).

Below are some of the areas we have not covered together with a few references that contain additional related references so that the interested reader may learn more.

The unknown covariance case. Strawderman (1973) found proper Bayes minimax estimators for the multivariate normal case when $\Sigma = \sigma^2 I$ (σ^2 unknown) and the loss is scaled squared error loss. Berger et al. (1977), Gleser (1979), and Lin and Tsai (1973) found minimax Stein-like estimators for the case of completely unknown Σ and general quadratic loss. See also Maruyama and Strawderman (2005).

The non-normal location vector case. Strawderman (1974) developed Baranchik-type and also generalized Bayes minimax estimators of the mean vector of a scale mixture of multivariate normal distributions under squared error loss. Brandwein and Strawderman (1978) and Brandwein (1979) gave Baranchik-type minimax estimators for spherically symmetric unimodal and spherically symmetric distributions respectively under squared error loss. Berger (1975) also gave similar results but with different methods of proof in the spherically symmetric case. See also Shinozaki (1984), Brandwein and Strawderman (1991), Cellier and Fourdrinier (1995), Fourdrinier and Strawderman (1996), Maruyama (2003a and 2003b), Fourdrinier, Kortbi, and Strawderman (2008), and Fourdrinier and Strawderman (2008).

Non-quadratic loss. Brandwein and Strawderman (1978) gave minimax Baranchik-type estimators for spherically symmetric distributions when the loss function is a concave function of quadratic loss. Hwang and Casella (1982) and Casella and Hwang (1983) gave shrinkage based confidence sets.

Admissibility. Brown (1966) establishes admissibility (for $p < 3$) and inadmissibility conditions (for $p \geq 3$), for the best equivariant estimator of a location parameter

under a very general loss function. Brown (1971) gave conditions on the prior distribution under which a Bayes estimator is admissible or inadmissible for the known covariance, quadratic loss case. See also Berger and Strawderman (1996), Berger, Strawderman, and Tang (2005), and Kubokawa and Strawderman (2007).

Non-location problems. See Clevensen and Zidek (1975) for Bayes minimax shrinkage estimators for several Poisson parameters, Berger (1980) for estimation of multiple Gamma scale parameters, and Tsui (1979) and Hwang (1982) for discrete exponential families.

3.2 Bayesian Predictive Density Estimation

Edward I. George and Xinyi Xu

Predictive analysis, which extracts information from historical and current data to predict future trends and behavior patterns, is one of the most fundamental and important areas in statistics. Of the many possible forms a prediction can take, the richest is a predictive density, a probability distribution over all possible outcomes. Such a comprehensive description of future uncertainty opens the door to sharper risk assessment and better decision making. The statistical challenge of course is how to estimate an unknown predictive density from historical or current data. For this purpose, the Bayesian approach of introducing a prior on the unknowns provides a natural and immediate answer. For example, suppose we observe data $X \sim p(x|\theta)$ with unknown parameter θ and wish to predict $Y \sim p(y|\theta)$. Given a prior π on θ, it follows from purely probabilistic considerations that a natural estimate of $p(y|\theta)$ is the predictive density

$$\hat{p}_\pi(y|x) = \int p(y|\theta)\pi(\theta|x)d\theta, \tag{3.2.1}$$

where $\pi(\theta|x)$ is the posterior distribution of θ. The sheer generality of this formulation provides a systematic approach to estimating $p(y|\theta)$ in a wide variety of setups. For instance, in subsequent sections we will illustrate how such predictive density estimates can borrow strength by combining information across dimensions in a multivariate setting and how they can adapt under model uncertainty in a regression setup. Furthermore, modern developments in numerical and simulation methods, such as Markov Chain Monte Carlo, and the rapid growth in computing power have unleashed the potential of these Bayesian predictive methods even in rather complicated settings.

Although a subjective Bayesian would find the predictive formulation above to be compelling, a skeptical frequentist might wonder how one should go about selecting a "good" prior or, for that matter, why should one even restrict attention to a Bayesian predictive density in the first place. At it turns out, these questions can be answered within a statistical decision theory framework, at least for certain for-

mulations. In such a framework, the performance potential of a density estimator $\hat{p}(y|x)$ of $p(y|\theta)$ is evaluated by a loss $L(p,\hat{p})$ which is typically averaged over x or θ or both (Berger, 1985). An appealing loss function here is the Kullback-Leibler (KL) or entropy loss,

$$L(p,\hat{p}) = \int p(y|\theta)\log\frac{p(y|\theta)}{\hat{p}(y|x)}dy, \tag{3.2.2}$$

which when averaged with respect to $p(x|\theta)$ leads to a measure of average long run performance, the KL risk criterion

$$R_{KL}(p,\hat{p}) = \int p(x|\theta)L(p,\hat{p})dx. \tag{3.2.3}$$

Aitchison (1990) noted that the KL loss is coherent here in the sense that for a given $\pi(\theta)$, the Bayes rule under $R_{KL}(p,\hat{p})$ is $\hat{p}_\pi(y|x)$, a property not shared for example by the symmetrized KL loss. For further discussion of the many attractive properties of KL loss, including considerations of information theory, proper local scoring and invariance, see Bernardo and Smith (1994) and the references therein. A more general class of loss functions, the divergence losses, have been considered for prediction in Ghosh, Mergel, and Datta (2008).

A traditional approach to predictive density estimation has been to substitute an estimator $\hat{\theta}$ for θ and then use $\hat{p}(y|x) = p(y|\hat{\theta})$. Although appealing in its simplicity, this commonly used "plug-in" approach has been shown by many to often lead to inferior predictive density estimators (Aitchison, 1975; Levy and Perng, 1986; Geisser, 1993; Komaki, 1996; Barberis, 2000; Tanaka and Komaki, 2005; Tanaka, 2006). In particular, Aitchison (1975) showed that maximum likelihood plug-in density estimators for Gamma models and for normal models are uniformly dominated under $R_{KL}(p,\hat{p})$ by Bayesian predictive estimators based on flat priors ($\pi(\theta) \equiv 1$). Intuitively, the problem with plug-in estimators is that they ignore the uncertainty about θ by treating it as if were known and equal to $\hat{\theta}$. In contrast, the Bayesian approach directly addresses this parameter uncertainty by margining out θ with respect to a prior distribution, thereby incorporating it into the density estimator.

We note in passing that for plug-in estimators, KL predictive risk is closely related to squared error estimation risk since by a Taylor expansion

$$R_{KL}(p(y|\theta), p(y|\hat{\theta})) \approx \frac{I(\theta)}{2}E(\theta-\hat{\theta})^2, \tag{3.2.4}$$

where $I(\theta)$ is the Fisher information. However, for Bayesian predictive estimators, this simple relationship does not hold. In fact, a Bayes rule does not necessarily belong to the class $\{p(y|\theta) : \theta \in R^p\}$, i.e., $\hat{p}_\pi(y|x)$ does not correspond to a "plug-in" estimator for θ, although under suitable conditions on π, $\hat{p}_\pi(y|x) \to p(y|\theta)$ as the sample size $n \to \infty$. Interestingly, as will be described in the next section, for Bayesian predictive densities under the multivariate normal model, there is a direct relationship between the KL predictive risk and the squared error estimation risk, a

connection that was established using Stein's unbiased estimate of risk in George, Liang, and Xu (2006).

The main challenge for the implementation of the Bayesian predictive approach is the choice of an appropriate prior π. Ideally, such a choice would be guided by meaningful subjective information. However, such information is often not available, especially in complicated problems with many unknown parameters. As noted by Liang et al. (2008), "Subjective elicitation of priors for model-specific coefficients is often precluded, particularly in high-dimensional model spaces, such as in nonparametric regression using spline and wavelet bases. Thus, it is often necessary to resort to specification of priors using some formal method (Kass and Wasserman, 1995; Berger and Pericchi, 2001)." Perhaps the simplest such "objective" approach is to attempt to reduce prior influence by using a diffuse prior such as a flat prior. Although such priors may yield reasonable procedures in low dimensional settings, such priors can also lead to inadequate predictive estimators, especially in high dimensional settings (see, e.g., Jeffreys, 1961; Berger and Bernardo, 1989).

Ultimately, a criterion such as the KL risk function described above provides a statistical decision theory framework in which the performance properties of Bayesian predictive densities can be compared and evaluated. Recent work using this approach has been fruitful for a number of high dimensional problems. In particular, work by Komaki (2001), Liang and Barron (2004), George, Liang, and Xu (2006), and Brown, George, and Xu (2008) has established conditions for minimaxity and admissibility as well as complete class results for Bayesian predictive density estimators in the fundamental multivariate normal setup. For distributions beyond the normal, new KL risk results for Bayesian predictive densities have been developed by Aslan (2006), Hartigan (1998), Komaki (1996, 2001, 2004), and Sweeting, Datta, and Ghosh (2006). In the following sections, we begin by describing the multivariate normal results in more detail, showing how they lead to uniformly improved Bayesian predictive density estimators over those based on uniform priors. We then proceed to describe how these results can be extended to the linear regression setting. After a simulated illustration of the potential of some of these Bayesian predictive estimators, we conclude with a discussion of directions for future research in this area.

3.2.1 Prediction for the Multivariate Normal Distribution

We now focus exclusively on predictive density estimation for the multivariate normal distribution, the centerpiece of parametric models. For this setup, we observe $X|\mu \sim N_p(\mu, v_x I)$ and wish to predict $Y|\mu \sim N_p(\mu, v_y I)$, two independent p-dimensional multivariate normal vectors with common unknown mean μ. Here $v_x > 0$ and $v_y > 0$ are assumed to be known. By a sufficiency and transformation reduction, this problem is equivalent to estimating the predictive density of X_{n+1} based on observing $X_1, \cdots, X_n$ where $X_1, \cdots, X_n|\theta$ i.i.d. $\sim N_p(\theta, \Sigma)$ with unknown θ and known Σ.

The Bayesian predictive density $\hat{p}_U$ under the uniform prior $\pi_U(\theta) \equiv 1$, namely

$$\hat{p}_U(y|x) = \frac{1}{\{2\pi(v_x + v_y)\}^{\frac{p}{2}}} \exp\left\{-\frac{\|y-x\|^2}{2(v_x+v_y)}\right\}, \tag{3.2.5}$$

dominates the plug-in rule $p(y|\hat{\theta}_{MLE})$, which substitutes the maximum likelihood estimate $\hat{\theta}_{MLE} = x$ for θ (Aitchison, 1975). Moreover, it is best invariant and minimax with constant risk (Murray, 1977; Ng, 1980; Liang and Barron, 2004), and is admissible when the model dimension $p = 1$ or 2 (Liang and Barron, 2004; Brown, George, and Xu, 2008). However, when $p \geq 3$, it turns out that $\hat{p}_U(y|x)$ can be further dominated by other predictive estimators. Indeed, Komaki (2001) showed that $\hat{p}_H$, the Bayesian predictive density under the Harmonic prior $\pi_H(\beta) \propto \|\beta\|^{-(p-2)}$ dominates $\hat{p}_U$ when the number of potential predictors $p \geq 3$. Similarly, Liang and Barron (2004) showed that proper Bayes rules $\hat{p}_a$ under Strawderman priors $\pi_a(\beta)$, which are defined hierarchically as $\beta|s \sim N_p(0, sv_0 I), s \sim (1+s)^{a-2}$, also dominate $\hat{p}_U$ when $p \geq 5$.

It is interesting to note that these results closely parallel some key developments concerning minimax estimation of a multivariate normal mean under quadratic loss. Based on observing $X|\theta \sim N_p(\theta, I)$, that problem is to estimate θ under

$$R_Q(\theta, \hat{\theta}) = E\|\hat{\theta} - \theta\|^2. \tag{3.2.6}$$

The maximum likelihood estimator $\hat{\theta}_{MLE}$, which is best invariant, minimax and admissible when $p = 1$ or 2, is dominated by the Bayes rules $\hat{\theta}_\pi = \int \theta\, \pi(\theta|x) d\theta$ under the Harmonic prior (Stein, 1974) and under the Strawderman prior (Strawderman, 1971) in high dimensions. Note that in the predictive density estimation problem, $\hat{p}_U$ plays the same "straw man" role as $\hat{\theta}_{MLE}$ in the point estimation problem. A further connection between $\hat{\theta}_{MLE}$ and $\hat{p}_U$ is revealed by the fact that $\hat{\theta}_{MLE}$ can also be motivated as the Bayes rule under the uniform prior $\pi_U(\theta) \equiv 1$.

George, Liang, and Xu (2006) drew out these parallels by establishing a unifying theory that not only subsumes the specialized results of Komaki (2001) and Liang and Barron (2004), but can also be used to construct large new classes of improved minimax Bayesian predictive densities. Their developments began by showing that any Bayes predictive density $\hat{p}_\pi$ can be represented in terms of the uniform prior estimator $\hat{p}_U$ and the corresponding marginal m_π, namely

$$\hat{p}_\pi(y|x) = \frac{m_\pi(w; v_w)}{m_\pi(x; v_x)} \hat{p}_U(y|x), \tag{3.2.7}$$

where $W = \frac{v_y X + v_x Y}{v_x + v_y}$ is a weighted average of X and Y. The principal benefit of the representation (3.2.7) is that it reduces the KL risk difference between $\hat{p}_\pi$ and $\hat{p}_U$ to a simple functional of the marginal $m_\pi(z; v)$

$$R_{KL}(\theta,\hat{p}_U)-R_{KL}(\theta,\hat{p}_\pi) = E_{\theta,v_w}\log m_\pi(W;v_w)-E_{\theta,v_x}\log m_\pi(X;v_x)$$
$$= \int_{v_w}^{v_x}\frac{\partial}{\partial v}E_{\theta,v}\log m_\pi(Z;v)dv. \quad (3.2.8)$$

Using the heat equation, Brown's representation (Brown, 1971) and Stein's identity (Stein, 1981), this risk difference can be represented by

$$R_{KL}(\theta,\hat{p}_U)-R_{KL}(\theta,\hat{p}_\pi) = \int_{v_w}^{v_x}E_{\theta,v}\Big(\frac{\nabla^2 m_\pi(Z;v)}{m_\pi(Z;v)}-\frac{1}{2}\|\nabla\log m_\pi(Z;v)\|^2\Big)dv$$
$$= \int_{v_w}^{v_x}E_{\theta,v}\left[2\nabla^2\sqrt{m_\pi(Z;v)}/\sqrt{m_\pi(Z;v)}\right]dv. \quad (3.2.9)$$

It is easy to see from (3.2.9) that a sufficient condition for a Bayes predictive density $\hat{p}_\pi$ to be minimax is that $m_\pi(z;v)$ or $\sqrt{m_\pi(z;v)}$ is superharmonic, or as a direct collary, that the prior π is superharmonic. These conditions are essentially the same as the minimax condition for the quadratic risk estimation problem. In both problems, that the Bayes rules under the harmonic prior and the Strawderman prior are minimax in high dimensions now follows easily from the fact that their corresponding marginals or square rooted marginals are superharmonic.

Comparing (3.2.9) with Stein's unbiased estimate of risk (Stein, 1974, 1981), George, Liang, and Xu (2006) reveals a fascinating identity that provides a connection between KL risk reduction to quadratic risk reduction

$$R_{KL}(\theta,\hat{p}_U)-R_{KL}(\theta,\hat{p}_\pi)=\frac{1}{2}\int_{v_w}^{v_x}\frac{1}{v^2}\left[R_Q^v(\theta,\hat{\theta}_U^v)-R_Q^v(\theta,\hat{\theta}_\pi^v)\right]dv. \quad (3.2.10)$$

Ultimately, it is this connection identity that yields similar sufficient conditions for minimaxity and domination in these two problems.

Brown, George, and Xu (2008) used the connection identity (3.2.10) to investigate the admissibility of Bayesian predictive density estimators. As proper Bayes rules are easily shown to be admissible in the KL risk setting, see Berger (1985), the focus was on formal Bayes rules. They showed that under essentially the same tail conditions for π as in Brown and Hwang (1982), there exists a sequence of densities $\{\pi_n\}$ such that $\int_{\|\theta\|\le 1}\pi_n(\theta)d\theta = \int_{\|\theta\|\le 1}\pi(\theta)d\theta > 0$ and that $B_Q(\pi_n,\hat{\theta})-B_Q(\pi_n,\hat{\theta}_{\pi_n})\to 0$, which using (3.2.10) leads to

$$B_{KL}(\pi_n,\hat{p}_\pi)-B_{KL}(\pi_n,\hat{p}_{\pi_n})=\frac{1}{2}\int_{v_w}^{v_x}\frac{1}{v^2}\left[B_Q^v(\pi_n,\hat{\theta}_\pi)-B_Q^v(\pi_n,\hat{\theta}_{\pi_n})\right]dv\to 0.$$

Then by a variant of Blyth's method, the corresponding Bayes predictive estimator $\hat{p}_\pi$ is admissible. The admissibility of $\hat{p}_U$ when $p=1$ or 2, and the admissibility of the Bayes rule under the harmonic prior when $p\ge 3$ follow directly from these tail conditions.

Going beyond obtaining prior tail conditions for admissibility, Brown, George, and Xu (2008) established a compelling justification for restricting attention to Bayesian predictive density estimators for the multivariate normal setup. They

showed that for this setup, the class of all generalized Bayes rules forms a complete class under the KL risk criterion. Thus, any predictive estimator, including any plug-in estimator, can at least be matched if not dominated in risk, by some Bayesian predictive density estimator.

These recent results for the multivariate normal model have laid the foundations for the development of new predictive methods for more complicated settings. In particular, the connection identity (3.2.10) provides a bridge between the predictive density estimation problem and the classic point estimation problem, providing a tool to borrow strength from some important, beautiful, and fundamental results in the latter area.

3.2.2 Predictive Density Estimation for Linear Regression

Linear regression models are the mainstay of statistical modeling, in many scenarios at least providing useful approximations to the relationship between explanatory variables and the future outcome of interest (Gelman et al., 2003). George and Xu (2008) and Kobayashi and Komaki (2008) both independently studied the problem of predictive density estimation under KL loss in a linear regression setting where they successfully extended a variety of the results discussed in the previous section.

The predictive density estimation problem in this context begins with the canonical normal linear model

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \varepsilon_{n\times 1}, \tag{3.2.11}$$

where $\varepsilon \sim N_n(0, \sigma^2 I)$ and X is a full rank, fixed $n \times p$ matrix of p potential predictors where $n \geq p$. Based on observing $X = x$, the goal is to estimate the density of a future vector $\tilde{Y}$ where

$$\tilde{Y}_{m\times 1} = \tilde{X}_{m\times p}\beta_{p\times 1} + \tau_{m\times 1}.$$

Here $\tau \sim N_m(0, \sigma^2 I)$ is independent of ε and $\tilde{X}$ is a fixed $m \times p$ matrix of the same p potential predictors in X with possibly different values. Assume that σ^2 is known, and without loss of generality set $\sigma^2 = 1$ throughout.

Letting $\hat{\beta}_y$ be the traditional maximum likelihood estimate of β based on the observed data, it is tempting to consider the plug-in predictive estimate $\hat{p}_{plug-in}(\tilde{y}|\hat{\beta}_y)$, which simply substitutes $\hat{\beta}_y$ for β in $p(\tilde{y}|\beta)$. However, as shown by George and Xu (2008), it can be dominated by the Bayesian predictive density $\hat{p}_U(\tilde{y}|y)$ under the uniform prior $\pi(\beta) \equiv 1$, namely,

$$\hat{p}_U^L(\tilde{y}|y) = \frac{1}{(2\pi\sigma^2)^{\frac{p}{2}}|\Psi|}\exp\left\{\frac{(\tilde{y}-\tilde{X}\hat{\beta}_y)'\Psi^{-1}(\tilde{y}-\tilde{X}\hat{\beta}_y)}{2\sigma^2}\right\}, \tag{3.2.12}$$

where $\Psi = I + \tilde{X}(X'X)^{-1}\tilde{X}'$. Moreover, $\hat{p}_U^L$ has constant risk and is minimax under the KL loss (Liang and Barron, 2004). Thus, like $\hat{p}_U$ in (3.2.5), it plays the role of straw man in this linear regression setup and is a good default predictive estimator. But not surprisingly, it can be improved upon by other Bayesian predictive densities when $p \geq 3$.

Analogous to the development in the multivariate normal case, the key marginal representation for Bayesian predictive estimator $\hat{p}_\pi^L$ in linear regression can be expressed as

$$\hat{p}_\pi^L(\tilde{y}|y) = \frac{m_\pi(\hat{\beta}_{y,\tilde{y}}, (W'W)^{-1})}{m_\pi(\hat{\beta}_y, (X'X)^{-1})} \hat{p}_U^L(\tilde{y}|y), \tag{3.2.13}$$

where $W = (X', \tilde{X}')'$, $\hat{\beta}_y = (X'X)^{-1}X'y \sim N_p(\beta, (X'X)^{-1})$, and

$$\hat{\beta}_{y,\tilde{y}} = (W'W)^{-1}W'(x', y')' \sim N_p(\beta, (W'W)^{-1}).$$

The representation (3.2.13) facilitates the the KL risk comparison of $\hat{p}_U^L$ and $\hat{p}_\pi^L$, where the difference takes the form

$$\begin{aligned} &R_{KL}(\beta, \hat{p}_U^L) - R_{KL}(\beta, \hat{p}_\pi^L) \\ &= E_{\beta,(W'W)^{-1}} \log m_\pi(\hat{\beta}_{y,\tilde{y}}; (W'W)^{-1}) - E_{\beta,(X'X)^{-1}} \log m_\pi(\hat{\beta}_y; (X'X)^{-1}). \end{aligned}$$

Since $(W'W)^{-1}$ and $(X'X)^{-1}$ are both symmetric and positive definite, there exists an invertible $p \times p$ matrix P such that

$$(X'X)^{-1} = PP' \qquad \text{and} \qquad (W'W)^{-1} = P\Sigma_D P', \tag{3.2.14}$$

where $\Sigma_D = \text{diag}(d_1, \ldots, d_p)$. Moreover, $d_i \in (0, 1]$ for all $1 \leq i \leq p$ with at least one $d_i < 1$, because $(W'W)^{-1} = (X'X + \tilde{X}'\tilde{X})^{-1}$ and $\tilde{X}'\tilde{X}$ is nonnegative definite. Therefore, the KL risk difference between $\hat{p}_U$ and $\hat{p}_\pi$ can then be represented by

$$R_{KL}(\beta, \hat{p}_U^L) - R_{KL}(\beta, \hat{p}_\pi^L) = \sum_{i=1}^{p} (1 - d_i) \int_{d_i}^{1} \frac{\partial}{\partial v_i} E_{\beta,V} \log m_{\pi_P}(Z, V) dv_i, \tag{3.2.15}$$

where $\pi_P(\beta) = \pi(P\beta)$ and $V = \text{diag}(v_1, \cdots, v_p)$. Paralleling the development of (3.2.9), unbiased estimates of the components in (3.2.15) can be obtained. By combining the above results, George and Xu (2008) established that a sufficient condition for $\hat{p}_\pi^L$ to be minimax is trace $\{H(m_\pi(z; PV_wP'))[(X'X)^{-1} - (W'W)^{-1}]\} \leq 0$ or trace $\left\{H(\sqrt{m_\pi(z; PV_wP')})[(X'X)^{-1} - (W'W)^{-1}]\right\} \leq 0$ for all $0 \leq w \leq 1$, where $H(f(z_1, \cdots, z_p))$ is the Hessian matrix of a function $f(z_1, \cdots, z_p)$. These results provide substantial generalizations of those in George, Liang, and Xu (2006), and can be used to construct improved predictive predictive estimators for linear regression models using scaled harmonic priors, shifted inverted gamma priors, and generalized t-priors, following the development in Fourdrinier, Strawderman, and Wells (1998).

3.2.3 Multiple Shrinkage Predictive Density Estimation

As will be illustrated in the simulation examples of the next section, Bayesian predictive density estimators can achieve dramatic risk reduction, but only in relatively small neighborhoods of prior modes. Thus, a desirable prior will not only satisfy the minimax and domination conditions above, but will also concentrate prior probability in a neighborhood of β. Now although β will almost always be unknown, there will sometimes be good reason to believe that β may be close to a particular subspace. For example, in large regression problems, it will often be suspected that at least some subset of the predictors is irrelevant in the sense that their coefficients, the corresponding components of β, are very small or zero. In this case, this suspicion would translate into the belief that β might be close to a subspace of β values for which a subset of components is identically zero. To exploit this possibility, George and Xu (2008) proposed the following minimax multiple shrinkage predictive estimators that adaptively shrink β towards the subspace most favored by the data.

First consider the construction of a predictive density estimator that shrinks a particular subset of the β components towards 0. Let S be the subset of $\{1,\ldots,p\}$ corresponding to the indices of the irrelevant predictors, and let β_S be the subvector of β corresponding to the columns of X indexed by S. If the components of β_S were in fact small or zero, it would be have been effective to have used a prior, such as the harmonic prior, that was centered around 0 on β_S and was uniform on $\beta_{\bar{S}}$, where $\bar{S}$ denotes the complement of S. Denoting such a prior by π_S and letting π_S^* be the restriction of π_S to β_S, i.e., $\pi_S^*(\beta_S) = \pi_S(\beta)$ is a function of β_S only, the Bayesian predictive density $\hat{p}^L_{\pi_S}(y|x)$ can be expressed as

$$\hat{p}_{\pi_S^*}(\tilde{y}|y) = \frac{m_{\pi_S^*}(\hat{\beta}_{S,y,\tilde{y}}, (W_S'W_S)^{-1})}{m_{\pi_S^*}(\hat{\beta}_{S,y}, (X_S'X_S)^{-1})}\, \hat{p}_U(\tilde{y}|y).$$

This shrinkage predictive density estimator offers substantial risk reduction when the components of β_S are all very small or zero by shrinking the posterior on the corresponding coefficients of β towards 0.

As was mentioned above, there will typically be uncertainty about which subset of the p predictors in X should be included in the model. Rather than arbitrarily selecting S, an attractive alternative is to use a multiple shrinkage predictive estimator which uses the data to emulate the most effective $\hat{p}_{\pi_S}$. Let Ω be the set of all potentially irrelevant subsets S, possibly even the set of all possible subsets. For each $S \in \Omega$, let π_S be a shrinkage prior constructed as above, and assign it probability $w_S \in [0,1]$ such that $\sum_{S\in\Omega} w_S = 1$. Then the mixture prior

$$\pi^*(\beta) = \sum_{S\in\Omega} w_S\, \pi_S(\beta)$$

will yield a multiple shrinkage predictive estimator

$$\hat{p}^*(\tilde{y}|y) = \sum_{S \in \Omega} \hat{p}(S|y)\hat{p}_{\pi_S}(\tilde{y}|y), \tag{3.2.16}$$

where $\hat{p}(S|y)$ is the model posterior probability of the form

$$\hat{p}(S|y) = \frac{w_S m_{\pi_S^*}(\hat{\beta}_{S,y}, (X_S'X_S)^{-1})}{\sum_{S \in \Omega} w_S m_{\pi_S^*}(\hat{\beta}_{S,y}, (X_S'X_S)^{-1})}.$$

The expression (3.2.16) shows that $\hat{p}^*(\mathbf{y}|y)$ is an adaptive convex combination of the individual shrinkage predictive estimates $\hat{p}_{\pi_S}$. Note that through $\hat{p}(S|y)$, $\hat{p}^*$ doubly shrinks $\hat{p}_U(\tilde{y}|y)$ by putting more weight on the $\hat{p}_{\pi_S}$ for which $m_{\pi_S^*}$ is largest and $\hat{p}_{\pi_S}$ shrinks most. Thus $\hat{p}^*$ is adaptive in the sense that it automatically adjusts to the subset index S for which β_S corresponds exactly to the zero or very small components of β. We expect $\hat{p}^*$ to offer meaningful risk reduction whenever any β_S is small for $S \in \Omega$, and so the potential for risk reduction using $\hat{p}^*$ is far greater than the risk reduction obtained by using an arbitrarily chosen $\hat{p}_{\pi_S}$.

It should be pointed out that the allocation of risk reduction by $\hat{p}^*$ is in part determined by the w_S weights in $\hat{p}(S|x)$. Because each $\hat{p}(S|y)$ is so sensitive, through $m_{\pi_S^*}$, to the value of $\hat{\beta}_{S,y}$, choosing the weights to be uniform should be adequate. However, one may also want to consider some of the more refined suggestions in George (1986b) for choosing such weights.

3.2.4 Simulation Studies

In this section, we demonstrate the shrinkage properties of some Bayesian predictive densities and their risk improvements over the default procedure under the uniform prior. To make the illustration simple and easy to understand, we use the multivariate normal setup from Section 3.2.1 for our simulations. Similar results can be obtained for linear regression models through direct extensions.

Figure 3.1 illustrates the shrinkage property of the Bayesian predictive density $\hat{p}_H(y|x)$ under the harmonic prior when $v_x = 1, v_y = 0.2$ and $p = 5$. Analogous to Bayes estimators $E_\pi(\theta|x)$ of θ that "shrink" $\hat{\theta}_{MLE} = x$, the marginal representation (3.2.7) reveals that Bayes predictive densities $\hat{p}_\pi(y|x)$ "shrink" $\hat{p}_U(y|x)$ by a multiplicative factor of the form $m_\pi(w; v_w)/m_\pi(x; v_x)$. However, the nature of the shrinkage by $\hat{p}_\pi(y|x)$ is different than that by $E_\pi(\theta|x)$. To insure that $\hat{p}_\pi(y|x)$ remains a proper probability distribution, the factor cannot be strictly less than 1. In contrast to simply shifting $\hat{\theta}_{MLE} = x$ towards the mean of π, $\hat{p}_\pi(y|x)$ adjusts $\hat{p}_U(y|x)$ to concentrate more on the higher probability regions of π.

To study the potential risk improvements provided by Bayesian predictive densities, we illustrate the risk differences of $\hat{p}_U(y|x)$ with the Bayes rules under the harmonic prior π_H or the Strawderman prior π_a with $a = 0.5$. Because $\hat{p}_H$ and $\hat{p}_a$ are unimodal at 0, it intuitively seems that the risk functions $R_{KL}(\theta, \hat{p}_H)$ and $R_{KL}(\theta, \hat{p}_a)$ should take on their minima at $\theta = 0$, and then asymptote up to $R_{KL}(\theta, \hat{p}_U)$ as

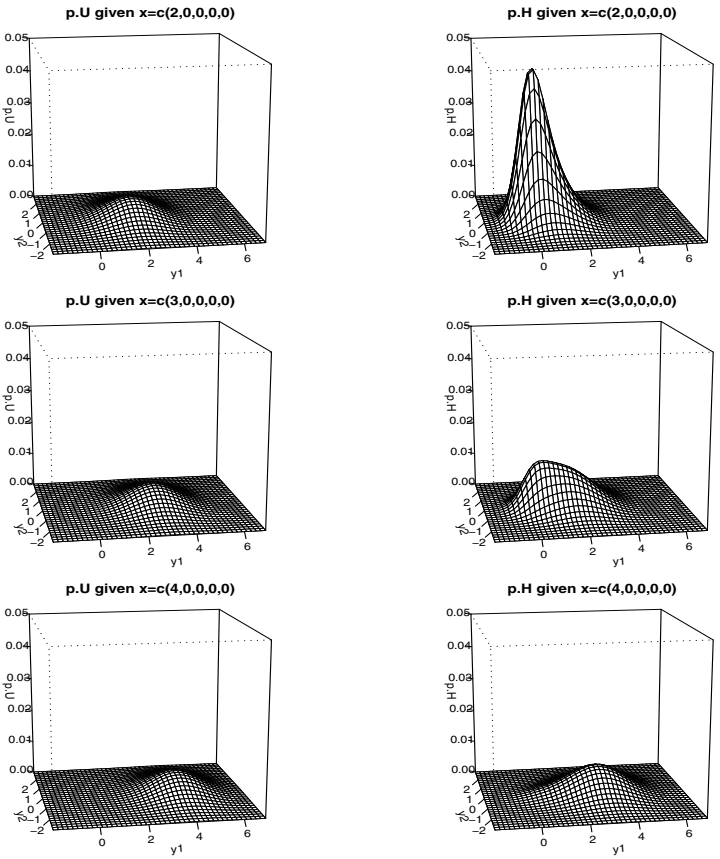


FIGURE 3.1. Shrinkage of $\hat{p}_U(y|x)$ to obtain $\hat{p}_H(y|x)$ when $v_x = 1, v_y = 0.2$ and $p = 5$. Here $y = (y_1, y_2, 0, 0, 0)$.

$\|\theta\| \to \infty$. That this is exactly what happens for these priors is illustrated in Figures 3.2 and 3.3, which display the difference at $\theta = (c, \ldots, c)'$, $0 \le c \le 4$ when $v_x = 1$ and $v_y = 0.2$ for dimensions $p = 3, 5, 7, 9$. The largest risk reduction in all cases occurs close to $\theta = 0$ and decreases rapidly to 0 as $\|\theta\|$ increases. (Recall that $R_{KL}(\theta, \hat{p}_U)$ is constant as a function of θ). At the same time, risk reduction by $\hat{p}_H$ and $\hat{p}_a$ is larger for larger p at each fixed $\|\theta\|$. Note that $\hat{p}_a$ offers more risk reduction than $\hat{p}_H$, apparently because it more sharply "shrinks $\hat{p}_U(y|x)$ towards 0." Note also that when $p = 3$, $[R_{KL}(\theta, \hat{p}_U) - R_{KL}(\theta, \hat{p}_a)]$ is negative for large θ, a manifestation of the non minimaxity of p_a when $a = 0.5$ and $p = 3$.

As we have seen in Section 3.2.3, the underlying priors and marginals of the Bayesian predictive densities can be readily modified to obtain minimax shrinkage towards subspaces, and linear combinations of superharmonic priors and marginals can be constructed to obtain minimax multiple shrinkage predictive densities $\hat{p}^*$ as

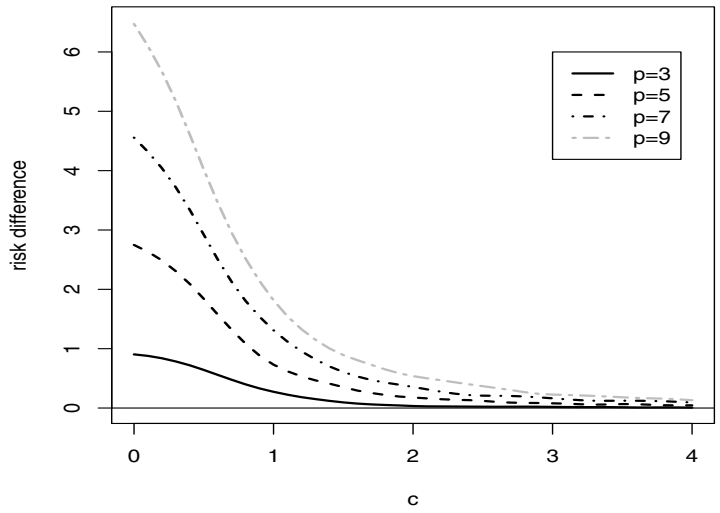


FIGURE 3.2. The risk difference between $\hat{p}_U$ and $\hat{p}_H$ when $\theta = (c, \cdots, c), v_x = 1, v_y = 0.2$.

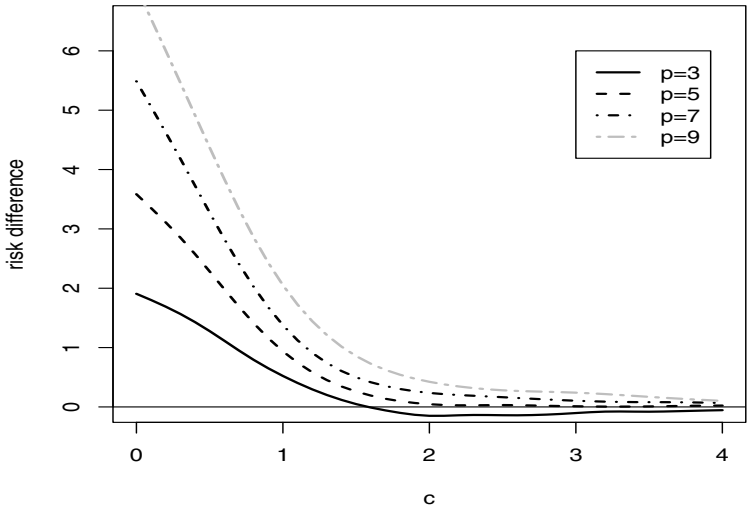


FIGURE 3.3. The risk difference between $\hat{p}_U$ and $\hat{p}_a$ with $a = 0.5$, $v_x = 1, v_y = 0.2$, and $\theta = (c, \cdots, c)$.

in (3.2.16), which are analogues of the minimax multiple shrinkage estimators of George (1986abc). As a result of the shrinkage behavior of $\hat{p}^*$, we would expect the risk reduction of $R_{KL}(\theta, \hat{p}^*)$ over $R_{KL}(\theta, \hat{p}_U)$ to be greatest wherever any β_S is small for $S \in \Omega$.

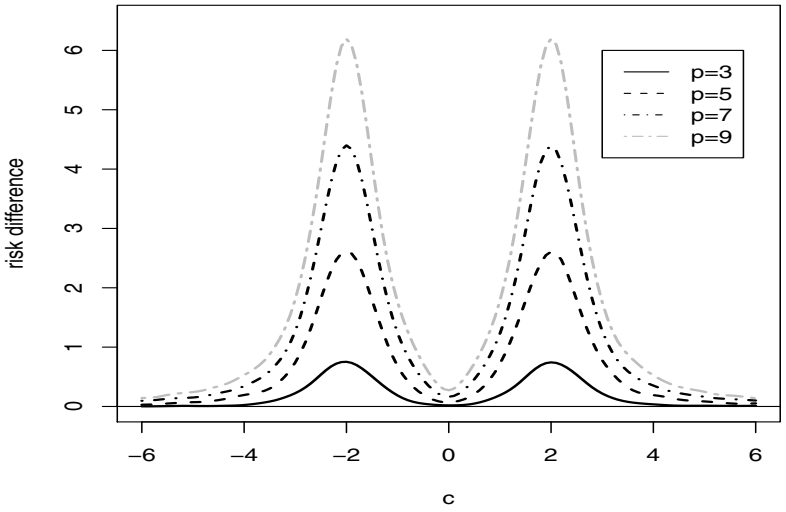


FIGURE 3.4. The risk difference between p_U and multiple shrinkage p_{H^*}, with $\theta = (c, \cdots, c), v_x = 1, v_y = 0.2, a_1 = 2, a_2 = -2$, and $w_1 = w_2 = 0.5$.

To see that this is precisely what would happen with $\hat{p}_{H^*}$, a multiple shrinkage version of $\hat{p}_H$ in the multivariate normal setting of Section 3.2.1, we consider $\hat{p}_{H^*}$ obtained analogously to (3.2.16) but using harmonic priors recentered at $s_1, s_2 \in R^p$, namely $\pi_{H_1}(\beta) \propto \|\beta - s_1\|^{-(p-2)}$ and $\pi_{H_2}(\beta) \propto \|\beta - s_2\|^{-(p-2)}$. Figure 3.4 illustrates the risk reduction $[R_{KL}(\theta, \hat{p}_U) - R_{KL}(\theta, \hat{p}_{H^*})]$ at various $\theta = (c, \ldots, c)'$ obtained by $\hat{p}_{H^*}$, which adaptively shrinks $\hat{p}_U(y|x)$ towards the closer of the two points $s_1 = (2, \ldots, 2)'$ and $s_2 = (-2, \ldots, -2)'$ using equal weights $w_1 = w_2 = 0.5$. As in Figures 3.2 and 3.3, we considered the case $v_x = 1, v_y = 0.2$ for $p = 3, 5, 7, 9$. As the plot shows, maximum risk reduction occurs when θ is close to either s_1 or s_2, and goes to 0 when θ moves away from these points. At the same time, for each fixed $\|\theta\|$, risk reduction by $\hat{p}_{H^*}$ is larger for larger p. It is impressive that the size of the risk improvement offered by $\hat{p}_{H^*}$ is nearly the same as each of its single target counterparts. The cost of multiple shrinkage enhancement seems negligible, especially compared to the benefits.

3.2.5 Concluding Remarks

Bayesian predictive densities have been widely used in many research areas. Besides predicting future trends and behavior patterns (Taylor and Buizza, 2004; Lewis and Whiteman, 2006; Weinberg, Brown, and Stroud, 2007), they have also been used in model checking and model diagnostics (Pardoe, 2001; Gelman et al., 2003; Sinharay, Johnson, and Stern, 2006), missing data analysis (Rubin, 1996; Gelman, King, Liu, 1998; Schafer, 1999; Gelman and Raghunathan, 2001; Little and Rubin, 2002), and data compression and information theory (Barron, Rissanen, and Yu, 1998; Clarke and Yuan, 1999; Liang and Barron, 2004).

Recent developments in Bayesian predictive density estimation for high-dimensional models provide valuable guidance for the construction of predictive estimators for particular setups. However, there are many open directions with much more to be done, especially for more general model setups. In this vein, Kato (2009) considered the predictive density estimation problem for a multivariate normal distribution where both the means and the variances are unknown. The Bayesian predictive estimator under an improper shrinkage prior was shown to dominate the default one under the right invariant prior when $p \geq 3$ and therefore be minimax. In another new direction, Xu and Liang (2010) explored the problem of estimating the predictive density of future observations from a nonparametric regression model. To evaluate the exact asymptotics of the minimax risk, they derived the convergence rate and constant for minimax risk among Bayesian predictive densities under Gaussian priors, and then showed that this minimax risk is asymptotically equivalent to that among all the density estimators. Such results provide not only powerful theoretical tools, but also easily implementable prior selection strategies for predictive analysis.

Acknowledgments: This work was supported by NSF grants DMS-0605102 and DMS-0907070.

3.3 Automated Bias-variance Trade-off: Intuitive Inadmissibility or Inadmissible Intuition?

Xiao-Li Meng

Prologue

Seeking an appropriate bias-variance trade-off is a common challenge for any sensible statistician, especially those at the forefront of statistical applications. Recently, addressing a class of bias-variance trade-off problems for studying gene-environment interactions, Mukherjee and Chatterjee (2008) adopted an approximate empirical partially Bayes approach to derive an estimator that amounts to using the following weighted estimator as a compromise:

$$\hat{\beta}_c = \frac{(\hat{\beta}-\hat{\beta}_0)^2}{\widehat{\mathrm{Var}}(\hat{\beta})+(\hat{\beta}-\hat{\beta}_0)^2}\hat{\beta} + \frac{\widehat{\mathrm{Var}}(\hat{\beta})}{\widehat{\mathrm{Var}}(\hat{\beta})+(\hat{\beta}-\hat{\beta}_0)^2}\hat{\beta}_0.$$

Here $\hat{\beta}$ and $\widehat{\mathrm{Var}}(\hat{\beta})$ are respectively our point estimator and its variance estimate of a parameter β under a model, and $\hat{\beta}_0$ is a more efficient estimator of β under a sub-model via fixing a nuisance parameter. The intuition here appears to be that since $\hat{B} = \hat{\beta}_0 - \hat{\beta}$ is an estimate of the bias in $\hat{\beta}_0$ when the sub-model fails, $\hat{\beta}_c$ should automatically give more weight to the robust $\hat{\beta}$ or the efficient $\hat{\beta}_0$ depending whether or not $\hat{B}^2$ is larger than $\widehat{\mathrm{Var}}(\hat{\beta})$. The implication here seems to be that the original $\hat{\beta}$ is inadmissible in terms of Mean-Squared Error (MSE) because it is dominated by $\hat{\beta}_c$, which appears to possess this magic *self-adjusting* mechanism for bias-variance trade-off without needing any assumption beyond those that guarantee the validity of the original $\hat{\beta}$. But is this intuition itself admissible? This question was posed as a Ph.D. qualifying exam problem at Harvard, in the context of a bivariate normal model. This section documents this examination, and concludes with a suggestion of revisiting the classic theory of admissibility, to which Professor Jim Berger has made fundamental contributions. The investigation also reveals a partial shrinkage phenomenon of the partially Bayes method, as well as a misguided insight in the literature of gene-environment interaction studies. Parts of this section adopt an interlacing style interweaving research investigations with pedagogical probes, honoring Berger's prolific contributions in both endeavors.

3.3.1 Always a Good Question ...

To many students, job candidates, and even some seasoned seminar speakers, a few faculty members are known to be "intimidating." We pose tough questions, demand intuitive explanations, challenge superficial answers, and we do so almost indiscriminately. A few students have expressed their surprise to me: "How could you guys be able to pick on almost any topic, and ask those penetrating questions even for things you apparently have never worked on?" I cannot speak for my fellow challengers, but the students are certainly correct that I have never worked on many of these topics, some of which I heard for the first time before I posed a question. If there is any secret—or bragging—here, it is the one that many senior statisticians work hard to pass on to our future generations. That is, there are only a very few fundamental principles in statistics, and the *bias-variance trade-off* is one of them. It is so deeply rooted in almost any statistical analysis, whether the investigator/speaker realizes it or not. Equipped with a few such powerful weapons, one can fire essentially in any situation, and almost surely not miss the target by too much.

The story I am about to tell is squarely a case of understanding bias-variance trade-off. In the context of a genetic study, a speaker mentioned a recent proposal by Mukherjee and Chatterjee (2008; hereafter M&C) for automatically achieving an appropriate bias-variance trade-off. The moment I saw the proposed formula, as given in the prologue, I knew silence would not be golden in this case. The method,

if it has the properties it was designed for, would have profound general implications given its simplicity and the practical demand for such "automated" methods. For the very same reason, however, it could do serious damage if it is applied indiscriminately but without its advertised properties in real terms.

As it happened, shortly after the seminar I needed to submit a problem for our Ph.D. qualifying examination. What could be more appropriate to test students' understanding of bias-variance trade-off, and at the same time their ability to carry out a rigorous investigation of a seemingly intuitive idea? The resulting qualifying exam problem is reproduced in Section 3.3.5 below, and the annotated solution is given in Section 3.3.6. Before presenting these materials *in verbatim*, which document an effort of integrating research investigation with pedagogical exploration, obviously the stage needs to be set. This is accomplished by Section 3.3.2, which discusses a gene-environment interaction study that motivates M&C; and by Section 3.3.3, which illustrates M&C's partially empirical Bayes approach via a bivariate normal example. Sections 3.3.2 and Section 3.3.3 also reveal, respectively, a misguided approximation in the literature of gene-environment interactions, and a partial shrinkage phenomenon of partially Bayes methods, and therefore they may be of independent interest. Indeed, Section 3.3.7 concludes with a suggestion of revisiting the classic theory of admissibility but with *partially Bayes risk*, which is also hoped to be a piece of admissible cake to the birthday-cake tasting (testing?) event for Jim Berger, an amazingly prolific scholar and Ph.D. adviser.

3.3.2 Gene-Environment Interaction and a Misguided Insight

3.3.2.1 Estimating Multiplicative Interaction Parameter

The motivation for M&C's proposal appears to be the need to address a bias-variance trade-off in studying gene-environment interactions. Following their setup, let E and G be respectively a binary environmental factor and a binary genetic factor, and D be the binary disease indicator; value "1" of any these binary variables indicates the presence (e.g., exposed, carrier, or with disease). One key interest here is to assess if there is a gene-environment (G-E) interaction in their impact on the odds of developing the disease. Let

$$O(G,E) = \frac{\Pr(D=1|G,E)}{\Pr(D=0|G,E)},$$

that is, the odds of disease in the sub-population defined by the pair $\{G,E\}$. Then the so-called *multiplicative interaction parameter* ψ is defined as

$$\psi = \frac{O(0,0)O(1,1)}{O(1,0)O(0,1)}, \tag{3.3.1}$$

which can be remembered as "odds ratio of odds," by analogy with the well-known ratio of cross-product expression of an odds ratio (OR), namely, the OR for a bivariate binary distribution $P(i,j) = \Pr(X = i, Y = j)$, expressed as

$$OR_{(X,Y)} = \frac{P(0,0)P(1,1)}{P(0,1)P(1,0)}.$$

This analogy also helps us to see why ψ is useful for assessing whether the factors G and E contribute to the odds of disease in a multiplicative fashion, that is, whether we can write $O(G,E) = g(G)e(E)$ for some functions g and e. This is because the mathematical reasoning behind the theorem "$OR_{(X,Y)} = 1$ if and only if $P(i,j)$ factors" is identical to that for "$\psi = 1$ if and only if $O(G,E)$ factors."

Consequently, by assessing whether $\beta = \log(\psi) = 0$, we can infer whether the effects of G and E are additive on the logit scale of the disease rate $\Pr(D = 1|G,E)$. In general, to estimate β directly (and therefore to assess it) would require a representative sample of $\{D,G,E\}$, as hinted by its expression in (3.3.1). Note however, by Bayes' Theorem,

$$O(G,E) = \frac{P(G,E|D=1)\Pr(D=1)}{P(G,E|D=0)\Pr(D=0)} \propto \frac{P(G,E|D=1)}{P(G,E|D=0)}.$$

It is then easy to verify that $\psi = OR_1/OR_0$, and hence

$$\beta = \log(\psi) = \log(OR_1) - \log(OR_0) \equiv \beta_0 - \theta, \tag{3.3.2}$$

where $OR_i = OR_{(G,E|D=i)}$ is the odds ratio for the conditional bivariate binary distribution $P(G,E|D=i), i = 0,1$. Consequently, if *G and E are conditionally independent given $D = 0$*, an assumption that will be labeled Assumption (0), then $\theta \equiv \log(OR_0) = 0$. This means that under Assumption (0), estimating β would be the same as estimating $\beta_0 \equiv \log(OR_1)$, the log odds ratio of *the diseased population* (a.k.a., the "cases"). This suggests the use of methods from retrospective sampling design, which typically is more effective, in terms of sampling cost and/or statistical efficiency, than prospective designs, especially when the disease prevalence is low; see Section 3.3.1 of M&C and the references therein.

3.3.2.2 A Potentially Misleading Insight

There is, of course, no free lunch. From a statistical inference perspective, the increased precision comes at the expense of possible serious bias when the assumption $\theta = 0$ fails. Incidently, in M&C, following an argument in Schmidt and Schaid (1999), this assumption is made as a consequence of another two assumptions: Assumption (1) G and E are independent in the general population (that is, not conditioning on the disease status) and Assumption (2) the disease is rare. Whereas these two assumptions do imply Assumption (0) hold *approximately* because when the diseased population is very small, the odds ratio between G and E for the disease-

free population can be approximated by that of the general population, these two assumptions were needed by Schmidt and Schaid (1999) apparently because they did not recognize that the second factor in their equation (1) is simply OR_0^{-1}, using our notation above. Consequently, instead of invoking the theoretically more insightful Assumption (0), they had to invoke the assumption that "*the disease risk is small at all levels of both study variables*" ("both study variables" here means the gene variable and environmental variable) in order to justify that the aforementioned second factor is (approximately) 1 (and hence $\theta \approx 0$). This unnecessary assumption apparently was inherited from Piegorsch, Weinberg, and Taylor (1994), who correctly pointed out the usefulness of the case-only studies.

This is a good demonstration of the value of precise theoretical derivation, because identity (3.3.2) shows clearly that $\beta = \beta_0$ if and only if $\theta = 0$, a condition that has little to do with the disease being rare. That is, it would be quite unfortunate if the quote above is interpreted as declaring that the so-called "case-only" approach for estimating ψ is useful only for rare diseases. Indeed, the only rationale for relying on Assumption (1) (and hence Assumption (2)) I can think of is if checking the independence of G and E in the general population is easier than that in the disease-free population. This could be a case when we do not trust the disease diagnosis, because the former does not require knowing each individual's disease status. But this advantage seems rather inconsequential in gene-environment interaction studies—if we do not have the disease status or do not trust them, then we have much more to worry about than assessing the independence between G and E.

Indeed, M&C's approach did not actually use Assumption (1) or Assumption (2). Instead, they directly use (3.3.2) by writing $\beta = \beta(\theta) = \beta_0 - \theta$ and then re-express (3.3.2) as

$$\beta(\theta) = \beta(0) - \theta. \tag{3.3.3}$$

This re-expression allows M&C to invoke a *partially Bayes* approach (Cox, 1975; McCullagh, 1990), which puts a prior on the nuisance parameter θ only. Since $\beta(0) = \log(OR_1)$ is a characteristic of the diseased population (i.e., $D = 1$), its inference does not involve $\theta = \log(OR_0)$, which is a characteristic of the disease-free population (i.e, $D = 0$). This separation allows M&C to first infer $\beta(0)$ via maximum likelihood estimation, and once $\beta(0)$ is replaced by its MLE, to infer $\beta = \beta(\theta)$ as a Bayesian inference problem of a *function* of the nuisance parameter θ. This is the essence of M&C's method, though their derivation contains a couple of theoretical complications that do not seem necessary (see Section 3.3.4).

Of course, for a pure Bayesian, such a hybrid "two-stage" method is neither necessary nor justifiable. However, as I argued in Meng (1994) in the context of a posterior predictive p-value (which is a posterior mean of a classic p-value as a function of a nuisance parameter under a prior on the nuisance parameter only, and hence a squarely partially Bayes entity), the value of such partially Bayesian methods should not be underestimated. Minimally, they allow some Bayesian perks to be enjoyed by those who do not wish to join the full B-club. For example, in the current setting, it allows the use of the prior knowledge/belief that the dependence between G and E is weak in the disease-free population. To see more clearly the

pros and cons of the partially Bayes framework, the next section will examine it in detail—and compare it with the fully Bayes approach—in the context of a normal regression model with one predictor.

3.3.3 Understanding Partially Bayes Methods

3.3.3.1 A Partially Bayes Approach for Bivariate Normal

M&C presented their general approach via a heuristic argument, which essentially amounts to assuming normality whenever needed, with variances treated as known. To avoid the distractions of the heuristics, which cannot be made precise in general because a Taylor expansion was invoked for approximating *a prior distribution*, let us assume directly that we have an i.i.d. sample $\{y_1, \ldots, y_n\}$ from the following bivariate normal model:

$$Y = \begin{pmatrix} X \\ Z \end{pmatrix} \sim N\left(\begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right), \tag{3.3.4}$$

where ρ is a known constant. Our interest here is to estimate β, with α being treated as a nuisance parameter.

As is well-known, without any prior knowledge, the MLE of β is $\hat{\beta}^{\text{MLE}} = \bar{Z}_n = \sum_i Z_i/n$, and the MLE for α is $\hat{\alpha}^{\text{MLE}} = \bar{X}_n = \sum_i X_i/n$. On the other hand, if we happen to know α, then the MLE of β is the regression estimator

$$\hat{\beta}(\alpha) = \bar{Z}_n + \rho(\alpha - \bar{X}_n). \tag{3.3.5}$$

Note that this definition of the $\hat{\beta}(\alpha)$ function allows us to reexpress (3.3.5) as

$$\hat{\beta}(\alpha) = \hat{\beta}(0) + \rho\alpha. \tag{3.3.6}$$

Clearly, given the data, the only unknown quantity in $\hat{\beta}(\alpha)$ is α (recall ρ is known here). Suppose we are willing to put down the prior $N(0, \tau^2)$ for α, where τ^2 represents our prior belief about how close α is to zero. Under this prior, the partially Bayes approach combines it with the (partial) likelihood from $\bar{X}_n|\alpha \sim N(\alpha, n^{-1})$ to arrive at the usual "shrinkage" posterior (e.g., Efron and Morris, 1973)

$$\alpha|\bar{X}_n \sim N\left(w_\tau \bar{X}_n,\ (n + \tau^{-2})^{-1}\right), \tag{3.3.7}$$

where $w_\tau = n/(n + \tau^{-2})$. Given this posterior of α, we can infer any of its functions, such as $\hat{\beta}(\alpha)$ of (3.3.5). In particular, M&C suggested to replace the α in (3.3.6) by the posterior mean in (3.3.7), which results in, after noting from (3.3.5) that $\hat{\beta}^{\text{MLE}} - \hat{\beta}(0) = \rho\bar{X}_n$, their estimator

$$\hat{\beta}^{\text{part}}_\tau \equiv \hat{\beta}(0) + w_\tau(\hat{\beta}^{\text{MLE}} - \hat{\beta}(0)) = w_\tau \hat{\beta}^{\text{MLE}} + (1 - w_\tau)\hat{\beta}(0). \tag{3.3.8}$$

Therefore, for a given hyperparameter τ^2, the *partially* Bayes estimator $\hat{\beta}_\tau^{\text{part}}$ for β is a compromise between the MLE under the restrictive model with $\alpha = 0$, $\hat{\beta}(0)$, and the MLE of β under the full model, $\hat{\beta}^{\text{MLE}} = \hat{\beta}(\hat{\alpha}^{\text{MLE}})$, as weighted by the usual shrinkage factor w_τ.

3.3.3.2 Comparing Full Bayes with Simultaneous Partially Bayes

Before we discuss the issue of choosing τ^2, it is informative to compare the above partially Bayes solution to a full Bayes one, which of course would require a joint prior for $\{\alpha, \beta\}$. To simplify the algebra, let us assume that *a priori* β and α are independent, and $\beta \sim N(0, \varsigma^2)$, with ς^2 given. Under this setup, the joint posterior of $\{\alpha, \beta\}$ obviously follows the usual regression calculation:

$$\begin{pmatrix} \alpha \\ \beta \end{pmatrix} \Bigg| \begin{pmatrix} \bar{X}_n \\ \bar{Z}_n \end{pmatrix} \sim N\left((\Omega^{-1} + \Sigma_n^{-1})^{-1} \Sigma_n^{-1} \begin{pmatrix} \bar{X}_n \\ \bar{Z}_n \end{pmatrix},\ (\Omega^{-1} + \Sigma_n^{-1})^{-1} \right), \tag{3.3.9}$$

where $\Sigma_n = \frac{1}{n} \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ and $\Omega = \begin{pmatrix} \tau^2 & 0 \\ 0 & \varsigma^2 \end{pmatrix}$.

To understand the difference between the full Bayes and the partially Bayes methods, however, it is more informative to invoke the following indirect derivation. Following the partially Bayes argument, we first treat α as a known constant. Then it is easy to see that the $\hat{\beta}(\alpha)$ of (3.3.5) is a sufficient statistic for β and

$$\hat{\beta}(\alpha) | \beta \sim N\left(\beta,\ n_\rho^{-1}\right),$$

where $n_\rho = n/(1-\rho^2)$ (larger than n due to gained information via regressing on α). Together with the prior $\beta \sim N(0, \varsigma^2)$, the identical calculation for (3.3.7) yields

$$\beta | \hat{\beta}(\alpha) \sim N\left(w_{\varsigma,\rho} \hat{\beta}(\alpha),\ \left(\varsigma^{-2} + n_\rho\right)^{-1} \right), \tag{3.3.10}$$

where $w_{\varsigma,\rho} = n_\rho/(n_\rho + \varsigma^{-2})$. The sufficiency of $\hat{\beta}(\alpha)$ for β for given α implies that $E[\beta|\bar{X}_n, \bar{Z}_n, \alpha] = E[\beta|\hat{\beta}(\alpha)]$, and hence, by iterated expectations and (3.3.6),

$$E[\beta|\bar{X}_n, \bar{Z}_n] = w_{\varsigma,\rho} E[\hat{\beta}(\alpha)|\bar{X}_n, \bar{Z}_n] = w_{\varsigma,\rho}(\hat{\beta}(0) + \rho E[\alpha|\bar{X}_n, \bar{Z}_n]). \tag{3.3.11}$$

At first glance, (3.3.11) achieves nothing because it simply transfers the calculation of $E[\beta|\bar{X}_n, \bar{Z}_n]$ to the equally difficult (or easy) problem of calculating $E[\alpha|\bar{X}_n, \bar{Z}_n]$. But this observation should also remind us that we can simply switch β with α (and accordingly ς with τ and $\bar{Z}_n$ with $\bar{X}_n$) to arrive at its dual identity

$$E[\alpha|\bar{X}_n, \bar{Z}_n] = w_{\tau,\rho}(\hat{\alpha}(0) + \rho E[\beta|\bar{X}_n, \bar{Z}_n]), \tag{3.3.12}$$

where $w_{\tau,\rho} = n_\rho/(n_\rho + \tau^{-2})$ and $\hat{\alpha}(\beta) = \bar{X}_n + \rho(\beta - \bar{Z}_n)$.

It is now a simple matter to solve (3.3.11)-(3.3.12) to arrive at

$$\hat{\beta}^{\text{full}}_{\tau,\varsigma} \equiv E[\beta|\bar{X}_n, \bar{Z}_n] = \frac{w_{\varsigma,\rho}[\hat{\beta}(0) + w_{\tau,\rho}\rho\hat{\alpha}(0)]}{1-\rho^2 w_{\varsigma,\rho}w_{\tau,\rho}}, \tag{3.3.13}$$

where the superscript "full" highlights the fact that it is identical to the fully Bayes answer from (3.3.9), as can be verified directly.

To express (3.3.13) in a more insightful way, we can use the fact that $\rho\hat{\alpha}(0) = \rho\bar{X}_n - \rho^2\bar{Z}_n = \hat{\beta}^{\text{MLE}} - \hat{\beta}(0) - \rho^2\hat{\beta}^{\text{MLE}}$ to arrive at

$$\hat{\beta}^{\text{full}}_{\tau,\varsigma} = w_{\varsigma\tau,\rho}\hat{\beta}^{\text{part}}_{\tau}, \tag{3.3.14}$$

where $\hat{\beta}^{\text{part}}_{\tau}$ is from (3.3.8),

$$w_{\varsigma\tau,\rho} = \frac{w_{\varsigma,\rho} - \rho^2 w_{\varsigma,\rho}w_{\tau,\rho}}{1-\rho^2 w_{\varsigma,\rho}w_{\tau,\rho}} = \frac{n_{\rho,\tau}}{n_{\rho,\tau}+\varsigma^{-2}},$$

and

$$n_{\rho,\tau} = \frac{n}{1-\rho^2(1-w_\tau)},$$

with $w_\tau = n/(n+\tau^{-2})$, as in (3.3.7). This means that, as far as point estimator goes, the full Bayes estimator $\hat{\beta}^{\text{full}}_{\varsigma,\tau}$ can be viewed as a further shrinkage of the partially Bayes estimator $\hat{\beta}^{\text{part}}_{\tau}$ towards zero. In particular, we notice that regardless of the value of ρ, $\lim_{\tau\to\infty} n_{\rho,\tau} = n$ and hence $\lim_{\tau\to\infty} w_{\varsigma\tau,\rho} = n/(n+\varsigma^{-2}) \equiv w_\varsigma$. This means that when $\tau = \infty$, the fully Bayes estimator for β would reduce to the usual shrinkage estimator $w_\varsigma\bar{Z}_n$ based on the $\bar{Z}_n$ margin alone. Intuitively, when $\tau = \infty$, there is no information to borrow from the prior knowledge of α for estimating β even if $\rho \neq 0$, and hence all the information is in the $\{Z, \beta\}$ margin.

3.3.3.3 Sequential Partially Bayes Methods and Partial Shrinkage

Intuitively, the fully Bayes method takes into account the prior information $\beta \sim N(0, \varsigma^2)$, which was not used by $\hat{\beta}^{\text{part}}_{\tau}$. The above derivation shows how one can achieve the full Bayes efficiency by performing two partially Bayes steps *simultaneously*, namely, by solving (3.3.11)-(3.3.12) as a pair, which is a special case of applying the "self-consistency" principle (Meng, Lee, and Li, 2009). In contrast, if we have carried out the partially Bayes method *sequentially*, that is, in two stages, then the full efficiency is not guaranteed even if priors for both β and α are used.

To see this more clearly, suppose we follow M&C's general argument and first treat the nuisance parameter α as known. Then conditioning on α, but taking into account the prior information on β via $N(0, \varsigma^2)$, our Bayes estimator for β is as given in (3.3.10),

$$\hat{\beta}_\varsigma(\alpha) \equiv E[\beta|\hat{\beta}(\alpha)] = w_{\varsigma,\rho}\hat{\beta}(\alpha) = w_{\varsigma,\rho}\left(\hat{\beta}(0) + \rho\alpha\right). \tag{3.3.15}$$

Now, unlike in the simultaneous method described above, if we follow the general argument as in M&C to treat $\hat{\beta}_\varsigma(\alpha)$ as the objective of our inference, we would replace α in the right most side of (3.3.15) by its (partial) posterior mean $E(\alpha|\bar{X}_n) = w_\tau \bar{X}_n$. This substitution then will lead to the sequential partially Bayes estimator

$$\hat{\beta}^{\text{seqe}}_{\varsigma\tau,\rho} = w_{\varsigma,\rho}\left(\hat{\beta}(0) + w_\tau(\rho\bar{X}_n)\right) = w_{\varsigma,\rho}\hat{\beta}^{\text{part}}_\tau. \tag{3.3.16}$$

Comparing (3.3.16) to (3.3.14), we see that although both of them are further shrinkages of the same $\hat{\beta}^{\text{part}}_\tau$ of (3.3.8) and both shrinkage factors depend on ς, $\hat{\beta}^{\text{seqe}}_{\varsigma\tau,\rho}$ shrinks less towards zero than the full Bayes estimator $\hat{\beta}^{\text{full}}_{\varsigma\tau,\rho}$. This is because

$$w_{\varsigma\tau,\rho} \equiv \frac{n_{\rho,\tau}}{n_{\rho,\tau} + \varsigma^{-2}} < \frac{n_\rho}{n_\rho + \varsigma^{-2}} \equiv w_{\varsigma,\rho}, \tag{3.3.17}$$

provided that

$$n_{\rho,\tau} \equiv \frac{n}{1-\rho^2(1-w_\tau)} < \frac{n}{1-\rho^2} \equiv n_\rho,$$

which is the case as long as $\rho \neq 0$ because $w_\tau = n/(n+\tau^{-2}) > 0$.

Intuitively, $\hat{\beta}^{\text{seqe}}_{\varsigma\tau,\rho}$ only achieves *partial shrinkage* compared to $\hat{\beta}^{\text{full}}_{\varsigma\tau,\rho}$ because it fails to take into account the prior information $\beta \sim N(0,\varsigma^2)$ *when estimating* α. Even when β and α are *a priori* independent, as long as X and Z are correlated conditional on the model parameter, X and Z are correlated with respect to the *predictive distribution*, that is, with the model parameter integrated out according to the prior. In our current setting, the correlation between (X,Z) with respect to their predictive distribution is $\rho_{\tau,\varsigma} = \rho/\sqrt{(1+\tau^2)(1+\varsigma^2)}$. As long as $\rho_{\tau,\varsigma} \neq 0$, the information on the marginal distribution of Z via α will have an impact on the marginal distribution of X (and vice versa). However, as $\varsigma \to \infty$, $\rho_{\tau,\varsigma} \to 0$ and hence this impact disappears as the prior information for β becomes diffuse. This can also be seen from (3.3.17), which becomes equality and hence $\hat{\beta}^{\text{seqe}}_{\varsigma\tau,\rho} = \hat{\beta}^{\text{fall}}_{\varsigma\tau,\rho}$ whenever $\varsigma = \infty$, regardless of the value of ρ or τ.

3.3.4 Completing M&C's Argument

For a given value of τ^2, M&C's general approach is essentially an approximate version of what is presented in Section 3.3.3.1, resulting in the same partially Bayes estimator general expression as in (3.3.8). I say essentially because M&C apparently introduced a technical complication that is not necessary. The derivation in Section 3.3.3.1 relies on treating $\hat{\beta}(\alpha)$ of (3.3.5) as our *estimand*. Note $\hat{\beta}(\alpha)$ actually depend on data, but from the Bayesian perspective, treating it as a known function of the unknown α only presents no conceptual or technical complication. However, M&C introduced $\beta(\theta)$ (using their generic notation θ, which is the same as α for the bivariate normal example), the limit of $\hat{\beta}(\theta)$, as the data-free estimand, and then

derive a partially Bayes estimator for $\beta(\theta)$ via the delta method $\beta(\theta) - \beta(0) \approx \beta'(0)\theta$ and the (partially Bayes) posterior on θ.

In the example of the gene-environment interaction, this definition of $\beta(\theta)$ worked well, because (3.3.3) holds for both the population version and sample version. However, for the bivariate normal example, although the sample version $\hat{\beta}(\alpha)$ of (3.3.5) is a linear function of α, the limit version, according to M&C's definition, would be a constant function because $\beta(\alpha) = \beta$ for all α and hence $\beta'(0) = 0$. Consequently, in general, the aforementioned delta method can be meaningless. Fortunately, this complication is really unnecessary, as we can work directly with $\hat{\beta}(\alpha)$ as the estimand for the partially Bayes method.

Another unnecessary complication is in M&C's treatment of estimating the prior variance as a hyperparameter. Given the prior $\theta \sim N(0, \tau^2)$, M&C first approximated the prior for $\phi \equiv \beta(\theta)$ by $N(\phi_0, \tau_\phi^2)$, where $\phi_0 = \beta(0)$ and $\tau_\phi^2 = [\beta'(0)]^2\tau^2$. (Perhaps this is where M&C felt the need to introduce the population version $\beta(\theta)$ because it might seem odd to put a prior on a data-dependent quantity $\hat{\beta}(\theta)$; but there is actually nothing incoherent in the partially Bayes framework for the latter operation.) To estimate τ_ϕ, M&C invoked an empirical Bayes argument, which estimates the hyperparameter τ^2 by $\max\{\hat{\theta}^2 - \hat{v}^2, 0\}$ when the approximation $\hat{v}^{-1}(\hat{\theta} - \theta)|\theta \sim N(0,1)$ holds for some statistic $\hat{v}$. A critical ingredient of M&C's proposal is to use $\hat{\theta}^2$ as a *conservative* estimate of τ^2, which then leads to a conservative estimator of the corresponding hyperparameter τ_ϕ^2 as $\hat{\tau}_\phi^2 = [\hat{\beta}'(0)]^2\hat{\theta}^2$. Substituting this estimator for the hyperparameter in a general version of (3.3.8) leads to M&C's general proposal. But $\hat{\beta}'(0)\hat{\theta}$ is nothing but the first-term Taylor expansion of $\hat{\beta}(\hat{\theta}) - \hat{\beta}(0) = \hat{\beta} - \hat{\beta}_0$ (though note the hidden assumption that $\hat{\beta}(\hat{\theta}) = \hat{\beta}$). This suggests that we can bypass the calculation of $\hat{\beta}'(0)$ and directly use $(\hat{\beta} - \hat{\beta}_0)^2$ as a conservative estimator of τ_ϕ^2. Indeed, with this modification, M&C's proposal has the simpler expression as given in the prologue.

What is necessary is that once the hyperparameter is estimated from the data, the operating characteristics of the resulting estimator must be evaluated specifically according to the estimation method used. That is, we can no longer rely on the established general properties of the (fully) Bayesian estimators to justify their corresponding empirical counterparts. We do tend to believe that such empirical estimators are reasonably accurate in a variety of situations in practice, as demonstrated in M&C via simulations. But the same belief sometimes can get us into deep trouble when we put too much faith on simulations, which are necessarily limited. Indeed, intuitively speaking, the idea that we can achieve a good universal compromise between $\hat{\beta}$ and $\hat{\beta}_0$ only using themselves plus an estimate of $\mathrm{Var}(\hat{\beta})$ (see the formula in prologue or (3.3.18) below) is just too good to be true. It is true that when $\hat{\beta}$ is an unbiased estimator of β, $\hat{B} \equiv \hat{\beta}_0 - \hat{\beta}$ provides an unbiased estimator of the bias in $\hat{\beta}_0$. But it would be illogical for us to worry about $\hat{\beta}$ having too large a variance—and hence the need to seek a reduction by bringing in a more efficient estimator $\hat{\beta}_0$—but not to worry about the large variability in $\hat{B}$, which depends on $\hat{\beta}$ critically. How can we be sure that the large error in the estimated weight $\hat{w}_{\tau_\phi} = w_{\hat{\tau}_\phi}$, which

in turn depends critically on $\hat{B}$, would not offset the gain in mean-squared error due to the (correct) weighting via w_{τ_ϕ}?

Indeed, we are not sure at all, as demonstrated in the following Ph.D. qualifying exam problem. (Again, both Section 3.3.5 and Section 3.3.6 are reproduced in verbatim from the actual exam, other than correcting a few typographical errors.)

3.3.5 *Learning through Exam: The Actual Qualifying Exam Problem*

During a recent departmental seminar, our speaker made an assertion along the following lines: "*I have two estimators, $\hat{\beta}$ and $\hat{\beta}_0$ for the same parameter β. The former is more robust because it is derived under a more general model, and the second is more efficient because it is obtained assuming a more restrictive model. The following is a compromise between the two:*

$$\hat{\beta}_c = \frac{(\hat{\beta}-\hat{\beta}_0)^2}{\widehat{\mathrm{Var}}(\hat{\beta})+(\hat{\beta}-\hat{\beta}_0)^2}\hat{\beta} + \frac{\widehat{\mathrm{Var}}(\hat{\beta})}{\widehat{\mathrm{Var}}(\hat{\beta})+(\hat{\beta}-\hat{\beta}_0)^2}\hat{\beta}_0, \tag{3.3.18}$$

where $\widehat{\mathrm{Var}}(\hat{\beta})$ is a consistent estimate of the variance of $\hat{\beta}$. This should work better because when the more restrictive model is true, $\hat{\beta}_c$ tends to give more weight to the more efficient $\hat{\beta}_0$, and at the same time, $\hat{\beta}_c$ remains consistent because asymptotically it is the same as $\hat{\beta}$."

As some of you might recall, I was both intrigued by and skeptical about this assertion. This problem asks you to help me to understand and investigate the speaker's assertion. To do so, let's first formalize the meaning of a general model and a more restrictive one.

Suppose we have i.i.d. data $\mathbf{Y} = \{y_1, \ldots, y_n\}$ from a model $f(y|\theta)$, where $\theta = \{\alpha, \beta\}$, both of which are scalar quantities, with β the parameter of interest, α the nuisance parameter, and the *meaning* of β does not depend on the value of α. Suppose the restrictive model takes the form $f_0(y|\beta) = f(y|\alpha = 0, \beta)$, i.e., under the restrictive model we know the true value of α is zero. Let $\hat{\theta} = \{\hat{\alpha}, \hat{\beta}\}$ be a consistent estimator of θ under the general model $f(y|\theta)$, and let $\hat{\beta}_0$ be a consistent estimator of β_0, which is guaranteed to be β only when the restrictive model $f_0(y|\beta)$ holds. We further assume all the necessary regularity conditions to guarantee their *joint* asymptotic normality, that is,

$$\sqrt{n}\left[\begin{pmatrix}\hat{\theta}\\ \hat{\beta}_0\end{pmatrix} - \begin{pmatrix}\theta\\ \beta_0\end{pmatrix}\right] \to N\left(\begin{pmatrix}\mathbf{0}\\ 0\end{pmatrix}, \Sigma = \begin{pmatrix}\Sigma_\theta & C^T\\ C & \sigma^2_{\hat{\beta}_0}\end{pmatrix}\right). \tag{3.3.19}$$

For simplicity of derivation, we will assume $\Sigma \geq 0$ (i.e., a semi-positive definite matrix) is *known*, and the convergence in (3.3.19) is in the L^2 sense (i.e., $X_n \to X$ means $\lim_{n\to\infty} E||X_n - X||^2 = 0$).

(A) The speaker clearly was considering a variance-bias trade-off, assuming that $\hat{\beta}_0$ is more efficient than $\hat{\beta}$ when the more restrictive model is true. Under the setup above, prove this is true asymptotically when $\hat{\theta}$ and $\hat{\beta}_0$ are maximum likelihood estimators (MLE, as in the superscript below) under the general model and restrictive model respectively and when we use the Mean-Squared Error (MSE) criterion (we can then assume Σ_θ and σ_β^2 are given by the inverse of the corresponding Fisher information). That is, prove that if the restrictive model holds, the (asymptotic) relative efficiency (RE) of $\hat{\beta}_0$ to that of $\hat{\beta}$ is no less than 1:

$$RE \equiv \lim_{n\to\infty} \frac{E[\hat{\beta}^{\text{MLE}} - \beta]^2}{E[\hat{\beta}_0^{\text{MLE}} - \beta]^2} \geq 1, \tag{3.3.20}$$

and give a necessary and sufficient condition for equality to hold. Provide an intuitive statistical explanation of this result, including the condition for equality to hold.

(B) Give a counterexample to show that (3.3.20) no longer holds if we drop the MLE requirement. What is the key implication of this result on the speaker's desire to improve $\hat{\beta}$ via $\hat{\beta}_0$?

(C) Since we assume Σ is known, we can replace $\widehat{\text{Var}}(\hat{\beta})$ in (3.3.18) by σ_β^2/n, where σ_β^2 is an appropriate entry of Σ_θ. We can therefore re-express (3.3.18) as

$$\hat{\beta}_c = (1 - W_n)\hat{\beta} + W_n\hat{\beta}_0, \quad \text{where} \quad W_n = \frac{\sigma_\beta^2}{\sigma_\beta^2 + n(\hat{\beta} - \hat{\beta}_0)^2}. \tag{3.3.21}$$

Prove that, under our basic setup (3.3.19), $\lim_{n\to\infty} E(W_n) = 0$ if and only if $\beta \neq \beta_0$.

(D) Using Part (C) to prove that whenever $\beta \neq \beta_0$,

$$\lim_{n\to\infty} \frac{E[\hat{\beta}_c - \beta]^2}{E[\hat{\beta} - \beta]^2} = 1. \tag{3.3.22}$$

Which aspect of the speaker's assertion this result helps to establish?

(E) To show that the condition $\beta \neq \beta_0$ cannot be dropped in Part (D), let us consider that our data $\{y_1, \ldots, y_n\}$ are i.i.d. samples from the following bivariate normal model:

$$Y = \begin{pmatrix} X \\ Z \end{pmatrix} \sim N\left(\begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right), \tag{3.3.23}$$

where ρ is *known*. Show that under this model, *when we use MLEs for* $\hat{\beta}$ and $\hat{\beta}_0$, $\sqrt{n}(\hat{\beta}_c - \beta)$ has exactly the same distribution as

$$\xi = Z_0 - \rho(X_0 + \sqrt{n}\alpha)\tilde{W}_n = (Z_0 - \rho X_0) + \rho[(1 - \tilde{W}_n)X_0 - \tilde{W}_n\sqrt{n}\alpha], \tag{3.3.24}$$

where $(X_0, Z_0)^\top$ has the same distribution as in (3.3.23) but with both α and β set to zero, and

$$\tilde{W}_n \equiv \tilde{W}_n(\rho,\alpha) = \frac{1}{1+\rho^2(X_0+\sqrt{n}\alpha)^2}.$$

Use the right-most expression in (3.3.24) to then show that

$$nE[\hat{\beta}_c - \beta]^2 = 1 - \rho^2 + \rho^2 G_n(\rho,\alpha), \tag{3.3.25}$$

where

$$G_n(\rho,\alpha) = E[(1-\tilde{W}_n(\rho,\alpha))X_0 - \tilde{W}_n(\rho,\alpha)\sqrt{n}\alpha]^2. \tag{3.3.26}$$

(F) Continuing the setting of Part (E), use (3.3.25) to prove that when $\alpha = 0$, for all n,

$$E[\hat{\beta}_0^{\text{MLE}} - \beta]^2 < E[\hat{\beta}_c - \beta]^2 < E[\hat{\beta}^{\text{MLE}} - \beta]^2,$$

as long as $\rho \neq 0$. Why does this result imply that $\beta \neq \beta_0$ cannot be dropped in Part (D)? What happens when $\rho = 0$?

(G) Still under the setting of Parts (E) and (F), verify that $G_n(0,\alpha) = n\alpha^2$, and then use this fact to prove that as long as $n\alpha^2 > 1$, there exists a $\rho^*_{n,\alpha} > 0$ such that for all $0 < |\rho| < \rho^*_{n,\alpha}$,

$$nE[\hat{\beta}_c - \beta]^2 > 1 = nE[\hat{\beta}^{\text{MLE}} - \beta]^2.$$

Does this contradict Part (D)? Why or why not?

(H) What do all the results above tell you about the speaker's proposed estimator $\hat{\beta}_c$? Does it have the desired property as the speaker hoped for? Would you or when would you recommend it? Give reasons for any conclusion you draw.

3.3.6 *Interweaving Research and Pedagogy: The Actual Annotated Solution*

(A) *This part tests a student's understanding of the most basic theory of likelihood inference, especially the calculation of Fisher information, and the fact that the MLE approach is efficient/coherent in the sense that when more assumptions are made its efficiency is guaranteed to be non-decreasing.*

The result (3.3.20) is easily established using the fact that if we write the expected Fisher information under the general model (with $n = 1$) as

$$I(\theta) = \begin{pmatrix} i_{\alpha\alpha} & i_{\alpha\beta} \\ i_{\alpha\beta} & i_{\beta\beta} \end{pmatrix}, \quad \text{and notationally} \quad I^{-1}(\theta) = \begin{pmatrix} i^{\alpha\alpha} & i^{\alpha\beta} \\ i^{\alpha\beta} & i^{\beta\beta} \end{pmatrix},$$

then $i^{\beta\beta} = [i_{\beta\beta} - i^2_{\alpha\beta} i^{-1}_{\alpha\alpha}]^{-1}$. The Fisher information under the restrictive model of course is given by $i_{\beta\beta}$ with $\alpha = 0$. Consequently, under our basic setup, when $\alpha = 0$,

$$RE = \frac{i^{\beta\beta}}{i_{\beta\beta}^{-1}} = \left[1 - \frac{i_{\alpha\beta}^2}{i_{\alpha\alpha} i_{\beta\beta}}\right]^{-1} \geq 1, \tag{3.3.27}$$

where equality holds if and only if $i_{\alpha\beta} = 0$ *when* $\alpha = 0$, that is, when β and α are *orthogonal* (asymptotically) under the restrictive model. Intuitively, the gain of efficiency of $\hat{\beta}_0^{\text{MLE}}$ over $\hat{\beta}^{\text{MLE}}$ is due to $\hat{\beta}^{\text{MLE}}$'s *covariance adjustment* via $\hat{\alpha}^{\text{MLE}} - \alpha$ when $\alpha = 0$. However, this adjustment can take place if and only if $\hat{\beta}^{\text{MLE}}$ is correlated with $\hat{\alpha}^{\text{MLE}}$ when $\alpha = 0$, which is the same as $i_{\alpha\beta} \neq 0$.

(B) *This part in a sense is completely trivial, but it carries an important message. That is, the common notation/intuition that "the more information (e.g., via model assumptions) or the more data, the more efficiency" can be true only when the procedure we use processes information/data in an efficient way (e.g., as with MLE).*

There are many trivial and "absurd" counterexamples. For example, in Part (A), if we use the same MLE under the general model, but only use 1/2 our samples when applying the MLE under the restrictive model, then the RE ratio in (3.3.27) obviously will be *deflated* by a factor 2, and hence it can easily be made to be less than one.

[A much less trivial or absurd example is when we want to estimate the correlation parameter ρ with bivariate normal data $\{(x_i, y_i), i = 1, \ldots, n\}$. Without making any restriction on other model parameters, we know the sample correlation is asymptotically efficient with asymptotic variance $(1-\rho^2)^2/n$ (see Ferguson, 1996, Chapter 8). Now suppose our restrictive model is that both X and Y have mean zero and variance 1. The Fisher information for this restrictive model is $(1+\rho^2)/(1-\rho^2)^2$, therefore $RE = 1 + \rho^2 \geq 1$, which confirms Part (A). However, since $E(XY) = \rho$ under the restrictive model, someone might be tempted to use the obvious moment estimator $\hat{r}_n = \sum_i x_i y_i / n$ for ρ. But one can easily calculate that the variance (and hence MSE) of $\hat{r}_n$ is $(1+\rho^2)/n$ for any n. Consequently, the RE of $\hat{r}_n$ compared to the sample correlation is (asymptotically) $(1-\rho^2)^2/(1+\rho^2)$, which is always less than one and actually approaches zero when ρ^2 approaches 1. So the additional assumption can hurt tremendously if one is not using an efficient estimator! (Students may recall that my qualifying exam problem from a previous year was about this problem.) Moments estimators are used frequently in practice because of their simplicity and robustness (to model assumptions), but this example shows that one must exercise great caution when using moment estimators, especially when making claims about their relative efficiency when adding assumptions or data.]

(C) *Intuitively this result is obvious, because when $\beta \neq \beta_0$, the denominator in W_n can be made arbitrarily large as n increases, and hence its expectation should go to zero. But this part tests a student's ability to make such "hand-waving" argument rigorous without invoking excessive technical details, which is an essential skill for theoretical research.*

Let $\Delta_n = \sqrt{n}(\hat{\beta} - \hat{\beta}_0 - \delta)$, where $\delta = \beta - \beta_0$. Then by (3.3.19), Δ_n converges in L^2 to $N(0, \tau^2)$, where $\tau^2 = a^\top \Sigma a$, with $a = (0, 1, -1)^\top$. Therefore, there exists a n_0 such that for all $n \geq n_0$, $\text{Var}(\Delta_n) \leq 2\tau^2$. Consequently, for any $\varepsilon > 0$, if we let

$M_\varepsilon = \sqrt{2\tau^2/\varepsilon}$, and $A_n = \{|\Delta_n| \geq M_\varepsilon\}$, then by Chebyshev's inequality, we have

$$\Pr(A_n) = \Pr(|\Delta_n| \geq M_\varepsilon) \leq \frac{\mathrm{Var}(\Delta_n)}{M_\varepsilon^2} \leq \varepsilon. \tag{3.3.28}$$

Now if $\delta \neq 0$, then as long as $n \geq \frac{M_\varepsilon^2}{\delta^2}$, we have, noting $0 < W_n = \frac{\sigma_\beta^2}{\sigma_\beta^2 + (\Delta_n + \sqrt{n}\delta)^2} \leq 1$,

$$0 \leq E(W_n) = E(W_n \mathbf{1}_{A_n}) + E(W_n \mathbf{1}_{A_n^c}) \leq \Pr(A_n) + \frac{\sigma_\beta^2}{\sigma_\beta^2 + (\sqrt{n}|\delta| - M_\varepsilon)^2}, \tag{3.3.29}$$

where in deriving the last inequality we have used the fact that $(u+v)^2 \geq (|u|-|v|)^2$. That $E(W_n) \to 0$ then follows from (3.3.28) and (3.3.29) by first letting $n \to \infty$ in (3.3.29), and then letting $\varepsilon \to 0$ in (3.3.28).

To prove the converse, we note that when $\delta = 0$, $W_n = \frac{\sigma_\beta^2}{\sigma_\beta^2 + \Delta_n^2}$. Therefore, by (Jensen's) inequality $E(X^{-1}) \geq [E(X)]^{-1}$, we have

$$E(W_n) \geq \frac{\sigma_\beta^2}{\sigma_\beta^2 + E(\Delta_n^2)} \to \frac{\sigma_\beta^2}{\sigma_\beta^2 + \tau^2} > 0.$$

(D) *This part is rather straightforward, as long as the student is familiar with the Cauchy-Schwarz inequality (which is a must!)*

From (3.3.21), we have $\sqrt{n}(\hat{\beta}_c - \beta) = \sqrt{n}(\hat{\beta} - \beta) - W_n D_n$, where $D_n = \sqrt{n}(\hat{\beta} - \hat{\beta}_0)$. It follows then

$$nE(\hat{\beta}_c - \beta)^2 = nE(\hat{\beta} - \beta)^2 + E(W_n^2 D_n^2) - 2E[\sqrt{n}(\hat{\beta} - \beta)(W_n D_n)]. \tag{3.3.30}$$

Under our assumptions, the first term on the right hand side of (3.3.30) converges to $\sigma_\beta^2 > 0$, so (3.3.22) follows if we can establish that the second term on the right hand side of (3.3.30) converges to zero. This is because, by the Cauchy-Schwarz inequality, the third term on the right hand side of (3.3.30) is bounded above in magnitude by $2\sqrt{nE(\hat{\beta} - \beta)^2 E(W_n^2 D_n^2)}$, and hence it must then converge to zero as well if the second term does so. But by the definition of W_n in (3.3.21),

$$E(W_n^2 D_n^2) = E\left[W_n \frac{\sigma_\beta^2 D_n^2}{\sigma_\beta^2 + D_n^2}\right] \leq \sigma_\beta^2 E(W_n),$$

which converges to zero by Part (C) when $\delta = \beta - \beta_0 \neq 0$. The implication of this result is that the speaker's assertion that $\hat{\beta}_c$ is asymptotically the same as $\hat{\beta}$ is correct, as long as $\beta \neq \beta_0$. [Note there is a subtle difference between $\beta = \beta_0$ and $\alpha = 0$. The latter implies the former, but the reverse may not be true because one can always choose $\hat{\beta}_0$ to be $\hat{\beta}$ even if the restrictive model is not true.]

(E) *This part tests a student's understanding of multi-variate normal models and the basic regression concepts, with which one can complete this part without any tedious algebra.*

The most important first step is to recoganize/realize that under the general model, $\hat{\beta}^{\text{MLE}}=\bar{Z}_n$, and under the restrictive model, $\hat{\beta}_0^{\text{MLE}}=\bar{Z}_n-\rho\bar{X}_n$, where $\bar{X}_n$ and $\bar{Z}_n$ are the sample averages; hence $D_n=\rho\sqrt{n}\bar{X}_n$. The first expression in (3.3.24) then follows from (3.3.21) when we re-write it as $\hat{\beta}_c=\bar{Z}_n-W_n(\rho\bar{X}_n)$ and let $X_0=\sqrt{n}(\bar{X}_n-\alpha)$ and $Z_0=\sqrt{n}(\bar{Z}_n-\beta)$, and the fact that (X_0,Z_0) has the same bivariate normal distribution as in (3.3.4) but with zero means. The second expression is there to hint at the independence of the two terms, because the first term $(Z_0-\rho X_0)$ is the residual after regressing out X_0, and the second term is a function of X_0 only. With this observation, (3.3.25) follows immediately because the residual variance is $1-\rho^2$.

(F) *Again, this part does not require any algebra if a student understands the most basic calculations with bivariate normal and regression.*

When $\alpha=0$, $\tilde{W}_n(\rho,0)=\frac{1}{1+\rho^2X_0^2}$, and

$$G_n(\rho,0)=E[X_0(1-\tilde{W}_n(\rho,0))]^2=E\left[X_0^2\left(\frac{\rho^2X_0^2}{1+\rho^2X_0^2}\right)^2\right]\equiv C_\rho,$$

where the constant $C_\rho>0$ is free of n and it is clearly less than $E(X_0^2)=1$. Therefore the identity (3.3.25) immediately leads to $nE[\hat{\beta}_c-\beta]^2=1-(1-C_\rho)\rho^2$, which is strictly larger than $nE[\hat{\beta}_0^{\text{MLE}}-\beta]^2=1-\rho^2$ and smaller than $nE[\hat{\beta}^{\text{MLE}}-\beta]^2=1$, as long as $\rho\neq0$. Clearly in this case (3.3.22) of Part (D) will not hold because the ratio there will be $1-(1-C_\rho)\rho^2<1$, hence the condition $\beta\neq\beta_0$ cannot be dropped in Part (D) – note when $\rho\neq0$, $\beta\neq\beta_0$ is equivalent to $\alpha\neq0$.

When $\rho=0$, $\hat{\beta}^{\text{MLE}}=\hat{\beta}_0^{\text{MLE}}$, and hence regardless of the value of α, Part (D) holds trivially even though the condition $\beta\neq\beta_0$ is violated. This also provides another (trivial) example that $\beta=\beta_0$ does not imply $\alpha=0$, as we discussed at the end of the solution to Part (D) above.

(G) *This part demonstrates the need of some basic mathematical skills in order to derive important statistical results (that cannot be just "hand-waved"!).*

When $\rho=0$, $\tilde{W}_n(0,\alpha)=1$, and hence $G_n(0,\alpha)=n\alpha^2$. From its expression (3.3.26), the (random) function under expectation is continuous in ρ and bounded above by $X_0^2+n\alpha^2$, which has the expectation $1+n\alpha^2$. Hence, by the Dominated Convergence Theorem, $G_n(\rho,\alpha)$ is a continuous function of ρ for any given α and n. Consequently, whenever $G_n(0,\alpha)=n\alpha^2>1$, there must exist a $\rho_{n,\alpha}^*>0$, such that for any $|\rho|\leq\rho_{n,\alpha}^*$, $G_n(\rho,\alpha)>1$ as well. It follows then, when $0<|\rho|\leq\rho_{n,\alpha}^*$, from (3.3.25),

$$nE[\hat{\beta}_c-\beta]^2=1-\rho^2+\rho^2G_n(\rho,\alpha)>1-\rho^2+\rho^2=1=nE[\hat{\beta}^{\text{MLE}}-\beta]^2.$$

This inequality, however, does not contradict Part (D) because the choice of $\rho_{n,\alpha}^*$ depends on n, so Part (D) implies that as n increases, $\rho_{n,\alpha}^*\to0$.

(H) Parts (A) and (B) demonstrate that in order for the proposed estimator (3.3.18) to achieve the desired compromise, a minimal requirement is that there should be some "efficiency" requirement on the estimation procedures, especially the one under the more restrictive model. Otherwise it would not be wise in general to bring in $\hat{\beta}_0$ to *contaminate* an already more efficient and more robust estimator $\hat{\beta}$.

Parts (C) and (D) proved that under quite mild conditions, the proposed $\hat{\beta}_c$ is equivalent asymptotically to the estimator under the general model, as long as the estimator under the more restrictive model is *inconsistent*, that is, as long as $\beta_0 \neq \beta$. So in that sense the speaker's proposal is not harmful but not helpful either asymptotically, and therefore any possible improvement must be a finite-sample one (which apparently is what the speaker intended and indeed the only possible way if one uses MLE to start with).

Parts (E)-(G) give an example to show that when the restrictive model is true, the speaker's proposal can achieve the desired compromise, that is, $\hat{\beta}_c$ beats $\hat{\beta}^{\text{MLE}}$ in terms of MSE for all n, but it is not as good as $\hat{\beta}_0^{\text{MLE}}$. The latter is not surprising at all because in this case $\hat{\beta}_0^{\text{MLE}}$ is the most efficient estimator (asymptotically, but also in finite sample given its asymptotic variance is also the exact variance). However, when the restrictive model is not true, then there is no longer any guarantee that $\hat{\beta}_c$ will dominate $\hat{\beta}$ (indeed this is not possible in general whenever $\hat{\beta}$ is admissible). The result in Part (G) also hinted that in order for $\hat{\beta}_c$ to beat $\hat{\beta}$, the "regression effect" of $\hat{\beta}$ on $\hat{\alpha}$ must be strong enough (e.g., expressed in this case via $|\rho| > \rho^*_{n,\alpha}$) in order to have enough borrowed efficiency from $\hat{\beta}_0$ to make it happen.

In summary, the speaker's proposal can provide the desired compromise when the restricted model is close to being true and the original two estimators are efficient in their own right, but it cannot achieve this unconditionally. In general, it is not clear at all as when one should use such a procedure, especially when the original two estimators are not efficient to start with.

3.3.7 A Piece of Inadmissible Cake?

M&C's $\hat{\beta}_c$ evidently was proposed as an improvement on the original $\hat{\beta}$, with MSE as the intended criterion. Adopting the classic framework of decision theory (Berger, 1985), the hope is that $\hat{\beta}_c$ is *R-better* than $\hat{\beta}$ in terms of the squared loss:

$$R(\hat{\beta};(\alpha,\beta)) = \int_y (\hat{\beta}(y)-\beta)^2 f(y|\alpha,\beta)\mu(dy),$$

where $f(y|\beta,\alpha)$ is the sampling density and μ is its corresponding baseline measure. But for $R(\hat{\beta}_c;(\alpha,\beta)) \leq R(\hat{\beta};(\alpha,\beta))$ to hold for all β and α (and with strict inequality for at least one (α,β)) means $\hat{\beta}$ is not admissible under the squared loss. The simple normal problem investigated in Section 3.3.5 and Section 3.3.6 demonstrates clearly that this would be wishful thinking in general. The question then is:

How do we quantify the apparently good properties of $\hat{\beta}_c$, as suggested by the empirical evidences in M&C?

If we have a joint prior on $\{\alpha, \beta\}$, of course we can compare the Bayesian risks of $\hat{\beta}_c$ and $\hat{\beta}$. But the partially Bayes approach precisely wants to avoid any prior specification about β. This leads to the notion of *partially Bayes risk*

$$r_\pi(\hat{\beta};\beta) = \int R(\hat{\beta};(\alpha,\beta))\pi(d\alpha).$$

If we adopt such a measure, then one fundamental question is: Under which prior $\pi(\alpha)$ the original $\hat{\beta}$ is dominated by $\hat{\beta}_c$, that is, $r_\pi(\hat{\beta}_c;\beta) \le r_\pi(\hat{\beta};\beta)$ for all β?

Intuitively, it is possible for $\hat{\beta}_c$ to dominate $\hat{\beta}$ in terms of r_π when π puts enough mass on or near $\alpha = 0$, as suggested by Part (F) of Section 3.3.5. The trouble is that in practice we will not know how close the restrictive model is to the truth *when we wish for an automated bias-variance trade-off*, because if we knew, then we surely should have included the information in our model to improve our estimator (e.g., via an informative prior), just as if we know $\alpha = 0$ for sure, then we should just use $\hat{\beta}_0$ (assuming it is an efficient estimator under the sub-model). We therefore seem to run into a circular situation. The information we need to evaluate $\hat{\beta}_c$ meaningfully makes $\hat{\beta}_c$ unnecessary, but without it, there does not seem to exist a meaningful way to establish the superiority of $\hat{\beta}_c$.

This was the main reason that I suspected that $\hat{\beta}_c$ was more a craving than a creation. I of course hope my suspicion is groundless and that M&C's proposal can lead to a real advancement at the frontier of methods for accomplishing appropriate bias-variance tradeoff. But this is a case where only hard theory, not simulations nor intuitions, can settle the matter. After all, the whole industry of shrinkage estimation came out of the counter-intuitive—at least initially—Stein's paradox established by rigorous theory (Stein, 1956; James and Stein, 1961; Efron and Morris, 1977). There might be an empirical partially Bayes theory in parallel to the elegant one established by Efron and Morris (1973) for shrinkage via empirical Bayes, but the key ingredient in M&C, that is, estimating the prior variance via the *conservative* $(\hat{\beta}_0 - \hat{\beta})^2$ is likely to be fatal to this line of exploration because the performance of $\hat{\beta}_c$ depends critically on the reliability of this estimation.

Evidently, there is a lot to be learned from the classic theory of admissibility before we can settle this matter, because this is squarely a problem of comparing estimators under the squared loss. Professor Berger has done much to build this field, so it is only fitting for me to present the problem of comparing $\hat{\beta}_c$ and $\hat{\beta}$ in general as a piece of cake to him on the occasion of his 60th birthday.

Happy Birthday, Jim, even if the cake turns out to be inadmissible!

Acknowledgments: The author thanks Alan Agresti, Joe Blitzstein, and Xianchao Xie for constructive comments, Bhramar Mukherjee and Nilanjan Chatterjee for their truly inspirational article, and NSF for partial financial support.

Chapter 4
Bayesian Model Selection and Hypothesis Tests

Model comparison remains an active research frontier in Bayesian analysis. The chapter introduces related specific research problems, including the selection of a number of components in a mixture model and the choice of a training sample size when using virtual simulated training samples. The chapter also discusses an intriguing general property that sets Bayesian testing apart from frequentist testing, by effectively rewarding honest choice of an alternative hypothesis. Cheating does not pay.

4.1 Performance of Bayesian Model Selection Criteria for Gaussian Mixture Models

Russell J. Steele and Adrian E. Raftery

It is a great pleasure to congratulate James O. Berger on his 60th birthday. Jim is one of the giants of the Bayesian renaissance of the past few decades, and one area on which he has had an overwhelming impact is Bayesian model selection. In this section we discuss Bayesian model selection for Gaussian mixture models.

Throughout his career, Jim has built a strong case for using Bayesian methods for model selection, particularly as opposed to standard frequentist methods based on p values (Berger and Delampady, 1987; Berger and Sellke, 1987; Barbieri and Berger, 2004). The case is especially strong for the problem of choosing the number of components in mixture models, where frequentist methods have difficulties. There is an elegant but rather complicated theory for frequentist testing of one mixture component versus two, or one versus more than one (Lindsay, 1995), but we do not know of a fully satisfactory frequentist theory for selecting the number of components more generally.

Here we compare some of the Bayesianly motivated or justifiable methods for choosing the number of components in a mixture. We consider posterior probabili-

M.-H. Chen et al. (eds.), *Frontiers of Statistical Decision Making and Bayesian Analysis*, DOI 10.1007/978-1-4419-6944-6_4,

ties for a widely used proper prior, BIC, ICL, DIC, and AIC. We also introduce an explicit unit-information prior for mixture models, analogous to the prior to which BIC corresponds in regular models.

We compare these criteria via a simulation study. The design of the simulation study is critical, as it is easy to design a simulation study to favor one criterion or another. As a result, we based the design on the scientific literature rather than just specifying values ourselves, as is often done. We extracted 43 published estimates of mixture model parameters from the literature across a range of disciplines, achieving broad coverage of the literature prior to 2000. Using cluster analysis after appropriate normalization, we identified six representative sets of parameter values for our study.

The results were perhaps surprising: BIC outperformed the other criteria, including posterior probabilities based on proper priors. This does confirm the informal experience of workers in the area, particularly those using mixture models for clustering, who have been using BIC widely for this purpose for the past decade.

In Section 4.1.1 we review the existing Bayesian model selection criteria for mixture models that we include in our comparison. In Section 4.1.2 we describe our new unit information prior for mixture models. In Section 4.1.3 we give results for a simulated example and a real data example. Then in Section 4.1.4 we describe our simulation study and give the results. In Section 4.1.5 we discuss some other methods and issues in this area.

4.1.1 Bayesian Model Selection for Mixture Models

4.1.1.1 Priors for Mixture Models

We consider the univariate Gaussian mixture model with G components where observations $y_1, \ldots, y_n$ are independently and identically distributed with the density

$$p(y_i|\mu, \sigma^2, \lambda) = \sum_{g=1}^{G} \lambda_g f(y_i|\mu_g, \sigma_g^2),$$

where f is the univariate normal density with mean and variance parameters μ_g and σ_g^2, $\lambda_i \in (0,1)$, and $\sum_{i=1}^{G} \lambda_i = 1$.

Initially the most used prior was a semi-conjugate specification for the parameters of the mixture model, which was conjugate conditional on the unknown mixture component memberships (West, 1992; Diebolt and Robert, 1994; Chib, 1995). This assumes a Gaussian prior for each of the μ_j with prior mean ξ_j and prior variance $\sigma_j^2 \tau_j$. The priors for the variances are scaled inverse χ^2 random variables, i.e., $\sigma_j^{-2} \sim \frac{1}{2\beta}\chi^2_{2\alpha}$, where α and β are fixed hyperparameters. With this prior, full conditional posterior distribution conditional on the unknown cluster memberships can be found in closed form. Nobile and Fearnside (2007) used a similar structure, but

they placed an additional level of hierarchical, uniform prior distributions on τ_j and β_j.

Another commonly used approach is the conditionally semi-conjugate prior of Richardson and Green (1997). This differs from the conditionally conjugate prior in that the priors for the component density means are assumed to be independent of the component density variance parameters. It has been used as for finite Gaussian mixture models (Robert, Rydèn, and Titterington, 2000; Stephens, 2000a), and also more recently for Gaussian hidden Markov Models (Spezia, 2009) and image segmentation (Ferreira da Silva, 2009).

Richardson and Green (1997), and subsequent authors, also assigned an additional hierarchical prior to the scaling constant of the prior for the variances, assuming that $\beta \sim \frac{1}{2h}\chi^2_{2g}$. The prior proposed by Stephens (2000a) differs slightly from that of Richardson and Green in that κ and ξ are also allowed to vary, namely $\xi \sim \text{Unif}[\infty,\infty]$, and $\kappa^{-1} \sim \frac{1}{l}\chi^2_l$ where $l = 0.0001$. Stephens suggested this prior because he found that the posterior for the number of components G was sensitive to the the prior on μ (and thus to the value of κ).

The choice of prior hyperparameters can have a big effect on the estimation of the mixture parameters (Jasra, Holmes, and Stephens, 2005). In general, the hyperparameters for the component density priors have been chosen to be the same for each component density (Frühwirth-Schnatter, 2006). Data-dependent choices of the hyperparameters have been proposed by Raftery (1996a), Wasserman (2000), and Richardson and Green (1997) to achieve weakly informative priors. Richardson and Green (1997) chose ξ to be the overall mean of the data, the prior variance of the component means to be proportional to the square of the range of the data, the prior distribution of the variances to have scale parameters proportional to the square of the range, and $\alpha = 2$. Richardson and Green (1997) used a uniform Dirichlet prior for the mixture proportions, λ_g, although this choice can cause difficulties in convergence of Markov Chain Monte Carlo algorithms (Plummer, 2008). In our reading of the literature, we have found that these choices have been regularly used in both methodological and applied work.

4.1.1.2 Criteria for Choosing the Number of Mixture Components

Fully Bayesian MAP estimate. The obvious Bayesian choice of the number of mixture components is the posterior mode, or maximum a posteriori (MAP) estimate of G. This can be evaluated by reversible jump MCMC (Richardson and Green, 1997), or by the marked point process method of Stephens (2000b). Here we use Stephen's hierarchical modification of Richardson and Green's prior; his paper provides some evidence that it performs similarly to theirs, but is less sensitive to prior specification. We also use his marked point process algorithm, as implemented in his software, available at http://www.stat.washington.edu/stephens/papers/software.tar.

BIC. The BIC (Schwarz, 1978) provides a widely used approximation to the integrated likelihood (integrated over the model parameters) for regular models. It was used for mixtures by Roeder and Wasserman (1997) and has been widely used for

mixture models since, particularly for clustering (Dasgupta and Raftery, 1998; Fraley and Raftery, 2002), with good results in practice. It is defined as

$$\mathrm{BIC}(G) = 2p(y|\hat{\tau}, G) - d\log(n),$$

where d is the number of free parameters in the mixture model. For regular models, BIC is derived as an approximation to twice the log integrated likelihood using the Laplace method (Tierney and Kadane, 1986), but the necessary regularity conditions do not hold for mixture models in general (Aitkin and Rubin, 1985). However, Roeder and Wasserman (1997) showed that BIC leads to to a consistent estimator of the mixture density, and Keribin (2000) showed that BIC is consistent for choosing the number of components in a mixture model.

DIC. The DIC (Spiegelhalter et al., 2002) is an AIC-like likelihood penalization criterion, where the number of *effective* model parameters is used instead of the actual number of free parameters in the model. One stated objective of the DIC for model comparison is to minimize predictive error of the selected model. It has the form

$$\mathrm{DIC}(G) = -2\log p(y|\hat{\tau}, G) + 2p_d,$$

where $\hat{\tau}$ is a "good" estimate of τ with respect to the posterior distribution of the data (often a posterior mean, median, or mode) and p_d can be written as:

$$p_D = E_{\tau|y}(\log p(y|\hat{\tau})) - \log p(y|\hat{\tau}).$$

One can thus estimate p_D (and possibly $\hat{\tau}$) via $\tau_t, t = 1, ..., T$ where the τ_t are draws from a posterior sampling algorithm. This gives the following empirical estimate of the effective number of parameters for a particular model to be used in the above DIC calculation:

$$\hat{p}_D = \frac{1}{T}\sum_{i=1}^{T} \log p(y|\tau_t) - \log p(y|\hat{\tau}).$$

The choice of $\hat{\tau}$ can have a large influence on the estimated number of parameters. Although Spiegelhalter et al. (2002) used the posterior mean in most cases, here we will use the largest posterior mode. The posterior mean (and the posterior median as well) can give poor results (e.g., negative numbers of parameters) because of the multimodality of the mixture density.

Celeux, et al. (2006) discussed several alternative choices for $\hat{\tau}$, although they did not come to a definitive conclusion as to which would perform the best for choosing the order of a mixture. Plummer (2008) proposed a modification of the DIC to adjust for the complications with respect to estimating the effective number of parameters derived through a view of the DIC as an approximate penalized loss function. McGrory and Titterington (2007) used a variational Bayes approach to derive a version of the DIC that they found to perform well for choosing the number of components.

ICL. Biernacki, Celeux, and Govaert (1998) suggested an information criterion based on the complete data likelihood. They noted that although integrating over

the parameters of the observed is difficult, the following integral

$$p(y,z|G) = \int p(y,z|\tau,G)p(\tau)d\tau$$

is sometimes available in closed form. Even if it is not, one can rewrite the integral as

$$p(y|z,G)p(z|G) = \int p(y|z,\tau,G)p(z|G,\tau)p(\tau)d\tau.$$

Biernacki, Celeux, and Govaert (1998) noted that Laplace approximations to $p(y|z,\tau, G)$ are often valid and that the remaining integral of $p(z|\tau,G)$ over $p(\tau)$ are often available in closed form, as long as $p(\tau) = p(\tau)p(\lambda)$ and the prior distribution of z is independent of τ. Therefore, they proposed approximating $\log(p(y|z,G))$ using the BIC approximation

$$\log(p(y|z,\hat{\tau}^*,G))) - \frac{d}{2}\log(n),$$

where $\hat{\tau}^* = \text{argmax}_\tau p(y|z,\tau,G)$, which will not necessarily be the same as $\hat{\tau}$ from maximizing the observed data likelihood.

This leads to approximating twice the negative integrated classification likelihood by

$$\text{ICL} = -2*\log(p(y|\hat{z}',\hat{\tau}^*,G))) + (d-(G-1))*\log(n) - 2*K(\hat{z}'),$$

where $K(z) = \int p(z|\lambda,G)p(\lambda|G)$ depends on the prior for λ, d is the number of total free parameters in the mixture model as before, and $\hat{z}'$ is the MAP estimate of z given $\hat{\tau}^*$, i.e.

$$\hat{z}'_{ij} = \begin{cases} 1 \text{ if } \text{argmax}_g \hat{z}^*_{ig} = j, \\ 0 \text{ otherwise.} \end{cases}$$

If one specifies a Dirichlet$(\alpha_1,\ldots,\alpha_G)$ distribution for the mixing parameters, then $K(z)$ can be obtained in closed form as

$$K(z) = \log(\Gamma(\frac{G}{2})) + \sum_{g=1}^{G}\log(\Gamma(n_g(z)+\alpha_g)) - \sum_{g=1}^{G}\log(\Gamma(\alpha_g)) - \log(\Gamma(n+\sum_{g=1}^{G}\alpha_g)),$$

where $n(z)_g$ is the number of observations assigned to the g-th group by the allocation matrix z.

Biernacki, Celeux, and Govaert (1998) noted that if the $n_g(z)$ are far from zero, an additional approximation for $K(z)$ can be made using a Stirling's approximation (dropping $O(1)$ terms) as

$$K(z) \approx n\sum_{g=1}^{G}\hat{\lambda}_g\log(\hat{\lambda}_g) - \frac{1}{2}(G-1)*\log(n)$$

which gives us the following BIC approximation to the integrated complete-data likelihood

$$\text{ICL–BIC} = -2 * \log(p(y, \hat{z}' | \hat{\tau}^*, G))) + d * \log(n).$$

Biernacki, Celeux, and Govaert (2000) presented just the ICL–BIC, with the additional suggestion to use $\hat{\tau}$ instead of $\hat{\tau}^*$. This is the form of ICL that we use in our experiments.

AIC. The best known of the information criteria used for determining the number of components is Akaike's Information Criterion (AIC). The AIC is calculated for mixtures as:

$$\text{AIC}(G) = -2 \log p(y | \hat{\tau}, G) + 2d,$$

where d is the number of free parameters in the mixture (e.g., for one-dimensional normal mixtures with unconstrained variances, $d = 3 * G - 1$). The theoretical justification for AIC is that choosing the minimum value of the AIC asymptotically minimizes the mean Kullback-Leibler information for discrimination between the proposed distribution and the true distribution, i.e., the model with the minimum value value of the AIC should be asymptotically closest in Kullback-Leibler distance to the true model. However, several studies (Koehler and Murphree, 1988; Celeux and Soromenho, 1996) have found that the AIC overestimates the number of components for mixtures, most likely due to violations of the regularity conditions required for the approximation to hold. Compared to BIC, the AIC penalizes models with larger numbers of parameters less, leading to the choice of more mixture components.

AIC also has a Bayesian interpretation, leading to the MAP estimate in regular models when the amount of the information in the prior is of the same order as the amount of information in the data (Akaike, 1983). This is a highly informative prior, and will not be plausible in most cases, so the Bayesian interpretation of AIC is questionable in many situations.

4.1.2 A Unit Information Prior for Mixture Models

Kass and Wasserman (1995) showed that for regular models, the BIC provides an approximation to Bayes factors (posterior odds for one model against another when prior odds are equal to 1, in turn equal to ratios of integrated likelihoods) under a certain prior for the parameters. This is the so-called unit information prior (UIP), equal to a multivariate normal prior distribution with mean equal to the maximum likelihood estimates for the parameters and covariance matrix equal to the inverse information matrix scaled by the number of observations, i.e. the observed information matrix for a single observation. Raftery (1995) extended this result to show that BIC approximates the integrated likelihood for a single model under the UIP.

However, these results do not hold for mixture models because the required regularity conditions are not satisfied. As a result, we propose here an analogy to the

UIP for mixture models. We subsequently use the resulting MAP of G as another Bayesian estimator of G and assess its performance.

Let z be a matrix of indicator variables, such that $z_{ij} = 1$ if observation i is sampled from component j and 0 otherwise. Then one can write:

$$p(y,z|\mu,\sigma^2,\lambda) = \prod_{i=1}^{n}\prod_{g=1}^{G} \lambda_g^{z_{ig}} f(y_i|\mu_g,\sigma_g^2)^{z_{ig}}.$$

Conditional on z, the complete-data information matrix for the parameters μ and σ^2 becomes block diagonal by component:

$$i_C(\mu,\sigma^2) = \begin{array}{c} \\ \mu_1 \\ \sigma_1^2 \\ \vdots \\ \mu_G \\ \sigma_G^2 \end{array} \begin{array}{c} \begin{array}{ccccc} \mu_1 & \sigma_1^2 & \cdots & \mu_G & \sigma_G^2 \end{array} \\ \begin{bmatrix} \frac{n(z)_1}{\sigma_1^2} & 0 & \cdots & 0 & 0 \\ 0 & \frac{n(z)_1}{2\sigma_1^4} & \cdots & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \cdots & \frac{n(z)_G}{\sigma_G^2} & 0 \\ 0 & 0 & \cdots & 0 & \frac{n(z)_G}{2\sigma_1^4} \end{bmatrix} \end{array},$$

where $n(z)_g$ is equal to the number of observations assigned to group g according to the matrix z.

We now propose a data-dependent prior for mixture models based on the observed information for the complete data likelihood and $n(z)_g = 1$ for all $g = 1,...,G$. We use the following conjugate form for the priors for the component parameters:

- $\lambda \sim$ Dirichlet $(\alpha_1,...,\alpha_G)$;
- $\mu_g \sim$ Normal $(\mu_g^0, \sigma_g^2/\kappa_g)$; and
- $\sigma_g^2 \sim \sigma^{2(0)}$ Inv-$\chi^2_{\nu_g}$.

Each of the priors would be conjugate for the likelihoods in question if the group memberships were known and, hence, $p(y,z)$ has an analytic form.

One approach for specifying values for the prior's hyperparameters is to set the prior covariance matrix for the prior parameters equal to the scaled inverse of the observed information matrix for a single observation. Kass and Wasserman (1995) suggested this approach, but in the context of multivariate normal distributions for the parameters, rather than for the conjugate priors above. The prior variance matrix for the parameters μ_g and σ_g^2 is

$$\mathrm{Var}_\pi(\mu_g,\sigma_g^2) = \begin{bmatrix} \frac{\sigma^{2(0)}}{\kappa_g(\nu_g-2)} & 0 \\ 0 & \frac{2(\sigma_g^{2(0)})^2}{(\nu_g-2)^2(\nu_g-4)} \end{bmatrix}.$$

Setting the prior variance matrix equal to $i_C^{-1}(\hat{\mu}_g,\hat{\sigma}_g^2)$, it is necessary to solve a system of two equations in three variables:

$$\hat{\sigma}_g^2 = \frac{\sigma^{2(0)}}{\kappa_g(\nu_g - 2)} \text{ and } 2(\hat{\sigma}_g^2)^2 = \frac{2(\sigma_g^{2(0)})^2}{(\nu_g - 2)^2(\nu_g - 4)},$$

which has the following set of solutions:

$$(\nu_g - 4) = \kappa_g^2 \text{ and } (\nu_g - 2)\hat{\sigma}_g^2 = \sigma_g^{2(0)}.$$

We choose $\nu_g = 5$ and $\kappa_g = 1$, yielding $\sigma_g^{2(0)} = 3\hat{\sigma}_g^2$, because $\nu_g = 5$ is the smallest integer number of degrees of freedom that guarantees a finite variance for the variance parameters. This also has the very appealing feature of yielding a marginal unit-information prior for the component means. Figure 4.1 shows plots of the priors for the component means for a simulated data set of 400 observations where the true $G = 3$ (a histogram of the data is shown in Figure 4.2).

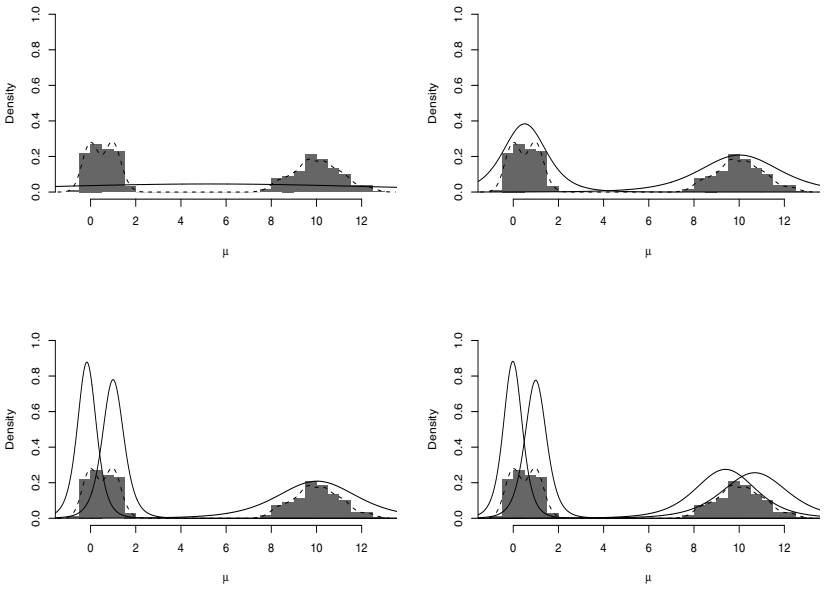


FIGURE 4.1. Plots of marginal priors for means for 1-4 components for the trimodal data set using the standard approximate unit information prior.

Note that we do not assume that the mean and variance parameters have the same prior distribution for each mixture component, and so we impose a prior labeling on the parameters of the mixture. One would need to take this into account when making inference about any of the parameters of the mixture model with the exception of G. In the absence of prior knowledge about the group labelings, the methods of

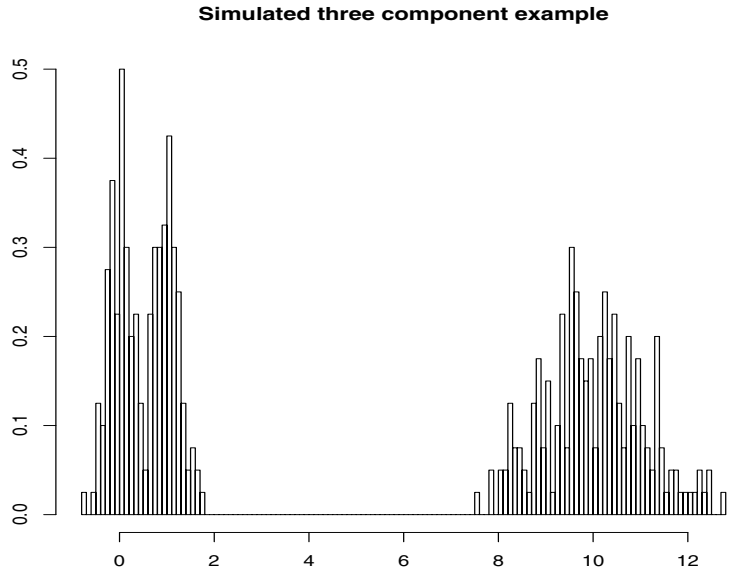


FIGURE 4.2. Histogram of 400 simulated data points.

Stephens (2000b) could be applied in order to make inference about all parameters of the distribution.

In the case of G, relabeling is not necessary, as the suggested prior yields posterior inference about G equivalent to the posterior inference using an exchangeable mixture prior with $G!$ components with each component corresponding to a different labeling for the components of the prior above. This is shown by the following theorem.

Theorem 4.1. *Posterior inference for the number of components based on a non-exchangeable prior distribution for the component parameters is equivalent to inference based on an exchangeable mixture prior that contains all $G!$ label-permuted versions of the non-exchangeable prior.*

Proof. Let $p_1(\tau)$ be the non-exchangeable prior distribution of interest. Let $p_s(\tau), s = 1,\ldots,G!$ be the the $G!$ label-permuted versions of the non-exchangeable prior. Let $\pi^*(\tau)$ be the exchangeable mixture of $p_s(\tau)$. Then $p(y|G) = \int p(y|\tau)p_i(\tau)d\tau$ is the same for $i = 1,...,G!$ because the likelihood is symmetric with respect to permutations of the labels. Thus, we have

$$p_1(y|G) = \frac{1}{G!}G! \int p(y|\tau, G)p_1(\tau|G)d\tau = \frac{1}{G!}\sum_{s=1}^{G!} \int p(y|\tau, G)p_s(\tau|G)d\tau$$
$$= \int p(y|\tau, G)\Big\{\frac{1}{G!}\sum_{s=1}^{G!} p_s(\tau|G)\Big\}d\tau = \int p(y|\tau, G)\pi^*(\tau)d\tau = p_*(y|G).$$

Because posterior inference about G depends on τ only through $p(y|G)$, any posterior inference about G using $p_1(\tau)$ as a prior will be equivalent to inference using $\pi^*(\tau)$ as a prior. □

Theorem 4.1 shows that the simpler non-exchangeable prior provides the same posterior inference about G as the computationally more expensive exchangeable mixture prior with $G!$ components. The exchangeable prior has the appealing characteristic that it does not assign substantial prior mass to values of the μ_g's far from the likelihood modes.

We compute the posterior probabilities of different values of G with this prior using the incremental mixture importance sampling (IMIS) method of Steele, Raftery, and Emond (2006).

4.1.3 Examples

We now compare the results of choosing the number of mixture components from the methods described above for one simulated example and one real data example.

4.1.3.1 Simulated Example

Figure 4.3 shows the posterior probabilities of different values of G for the trimodal dataset in Figure 4.2 using the different methods discussed above. In Figure 4.3, IL indicates estimated posterior probabilities from the unit information prior of Section 4.1.2, and Step-MPP indicates estimated posterior probabilities using the prior and methods of Stephens (2000a). For AIC, the posterior probabilities are derived from the Bayesian interpretation of AIC as minus twice the log integrated likelihood for a highly informative prior, as $p(G|y) \propto \exp(-\text{AIC}/2)$. DIC is computed using the unit information prior in Section 4.1.2 and is also viewed as an approximation to minus twice the log integrated likelihood. The prior for G was uniform over the integers $1, 2, \ldots, 7$ for each of the component density priors. The true number of components in this case is $G = 3$, and BIC and ICL both put almost all the posterior mass on this value. The other methods put some posterior mass on bigger values.

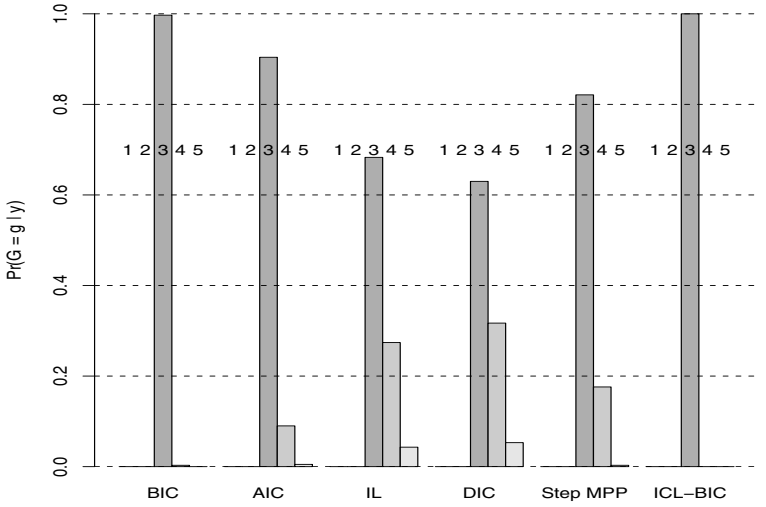


FIGURE 4.3. Posterior probabilities of G for trimodal data in Figure 4.2.

4.1.3.2 Galaxy Data

Figure 4.4 shows the posterior probabilities of G for the canonical galaxy velocity data set (Postman, Huchra, and Geller, 1986) analyzed by Roeder (1990) and several others since.

The methods give very different results. The fully Bayesian approach of Stephens (2000a) (which is similar to that of Richardson and Green, 1997) puts the most mass on the largest value considered, $G = 7$. AIC, DIC and the Bayesian approach using our unit information prior put most of the mass on $G = 6$ and $G = 7$. BIC shares the mass between $G = 3$ (25%) and $G = 6$ (75%), while ICL puts almost all the mass on $G = 3$.

We do not know the "correct" answer for this famous dataset, and so we cannot say which methods are "right" or "wrong." However, exploratory analysis of this dataset by Fraley and Raftery (2007) sheds some light. If one looks at the full dataset using standard goodness-of-fit assessment methods (the empirical cumulative distribution function and Kolmogorov-Smirnov tests), it is clear that it is not well fit by a single normal. This is in line with the fact that all methods give essentially zero posterior probability to $G = 1$.

There are two clear small "clusters," one at the left and one at the right of the dataset. When a mixture model with $G = 3$ is fit, these are clearly separated out, with one component corresponding to each of these small clusters, and the remaining

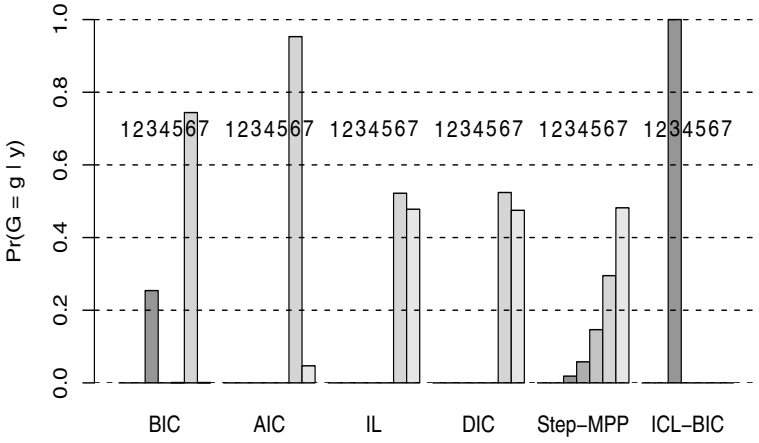


FIGURE 4.4. Posterior probabilities of G for galaxy example.

components corresponding to the majority of points in the middle (72 of the 82 points).

When one looks at this majority group of points in the middle, there are no strong deviations from normality apparent (Fraley and Raftery, 2007, Figure 4), and standard goodness of fit tests do not reject normality. For example, the P-value from a Kolmogorov-Smirnov test of normal is 0.254.

This suggests that no reasonable statistical analysis based on the data alone should categorically exclude $G = 3$ in favor of larger values of G. Yet, AIC, DIC, and the fully Bayesian analyses based on both priors considered do just that. ICL goes to the other extreme: it overwhelmingly favors $G = 3$ over all other possibilities. That leaves BIC, which gives mass to both $G = 3$ and $G = 6$, which seems scientifically reasonable.

We attach two caveats to these results. First, the DIC is based on the unit information prior in Section 4.1.2, and the results might be different for other priors, including the (very similar) priors of Richardson and Green (1997) and Stephens (2000a). Second, only BIC, the fully Bayesian methods and, arguably, AIC, can be interpreted as yielding posterior probabilities, but we put DIC and ICL on the same probabilistic evidence scale for comparison purposes.

4.1.4 Simulation Study

We now describe our simulation study to compare the different methods for choosing the number of components.

4.1.4.1 Study Design

The results of a simulation study for comparing methods can depend critically on the study design. Often the simulation study is designed by the researcher in a fairly arbitrary way. Given this, we tried to ground our study explicitly on the published experience with mixture models.

We searched the literature for published parameter estimates to be used as the basis for our design. In order to choose examples in a fair, comprehensive, but reasonable fashion, we looked at all papers published before July 1999 listed in the Science Citation Index/Web of Science that referenced one of three textbooks on mixture models:

- Everitt and Hand, Finite Mixture Distributions (1981);
- Titterington, Smith, and Makov, Statistical Analysis of Mixture Distributions (1985); and
- McLachlan and Basford, Mixture Models: Inference and Applications to Clustering (1987).

We also included a set of examples that were listed in the Titterington, Smith, and Makov book. In all, we looked at over 200 papers published between 1936 and 1999. Of these, we found 22 papers that listed 43 parameter estimates for two-component models. (There were very few reported estimates for mixture models with more than two components.)

The 43 two-component examples and references are listed in Tables 4 and 5 of Steele and Raftery (2009), along with the mixture parameter estimates. We standardized each example's parameters such that the group with the larger estimated component membership had mean 0 and variance 1. The smaller group's mean was then normalized to $|\mu_2 - \mu_1|$ and the variance to $\frac{\sigma_2^2}{\sigma_1^2}$. The standardized values are shown in Table 6 of Steele and Raftery (2009).

From an experimental design perspective, one would like the simulation study to "cover" the space of the mixture examples in the literature. Using the model-based clustering R package `mclust` (Fraley and Raftery, 1998), we fit a mixture of multivariate normal distributions to the literature parameter estimates, restricting the component densities to have equal, diagonal covariance matrices, using the BIC to choose the "optimal" number of components. Similar experimental design approaches for computer experiments can be found in texts on number theoretic methods (see for example Fang and Wang, 1994).

The BIC suggested 5 components, and we therefore used 5 settings of the mixture parameters for the two-component mixture model in our experiment. For balance we

TABLE 4.1. Parameters for the simulation study, as suggested by the 43 parameter estimates from the literature.

G=1: (Experiments 1-5) n=400, 300, 200, 150, 100 $\mu_1 = 0, \sigma_1 = 1$, for all experiments				
G=2: $\mu_1 = 0, \sigma_1 = 1$ for all experiments				
Experiment	μ_2	σ_2	λ_2	n
6	1.0	0.5	.25	400
7	2.0	2.0	.25	300
8	3.0	1.0	.25	200
9	4.0	2.5	.25	150
10	6.0	1.5	.25	100

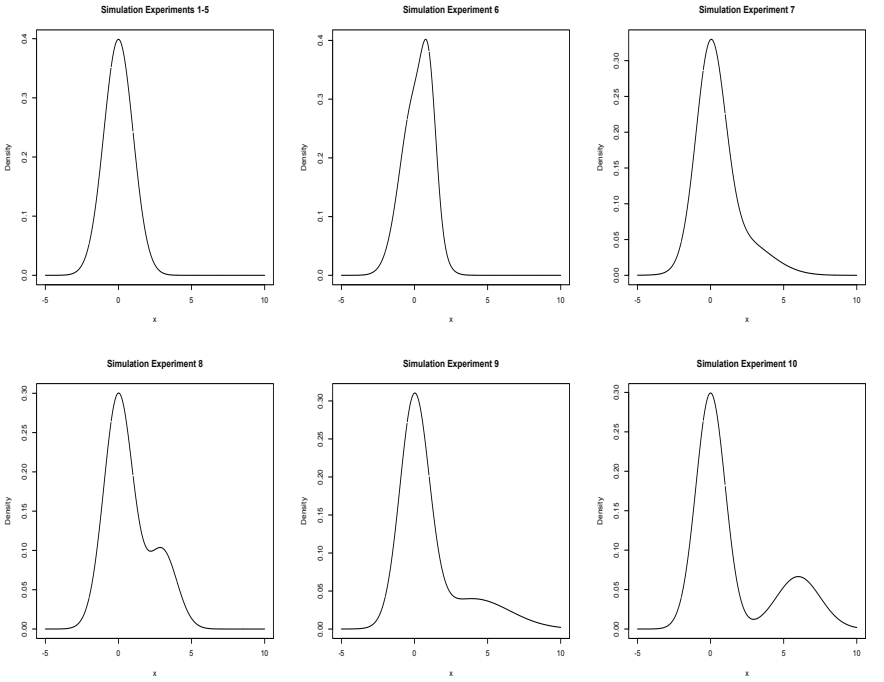


FIGURE 4.5. Simulation study: The densities used. The top left plot shows the one-component normal density which was the basis for experiments 1–5. The remaining plots show the two-component mixture of normals used for experiments 6–10.

also included 5 experiments with one component, and different sample sizes. The final simulation study design is shown in Table 4.1, and the mixture densities used are shown in Figure 4.5.

We generated 50 data sets for each of the 10 experiments. For each of the 500 simulated datasets, we then found the selected number of components for each of the six selection methods compared, as described in Section 4.1.3.

4.1.4.2 Results

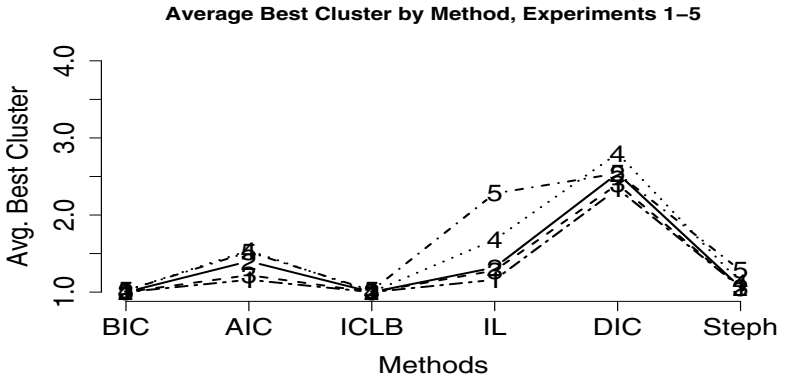
The average number of components selected by each method in each experiment is shown in Figure 4.6. In Figure 4.6, the experiments are listed in Table 4.1, ICLB refers to the ICL-BIC criterion, IL is the integrated from the unit information prior in Section 4.1.2, and Steph refers to the MAP estimate from Stephens's fully Bayesian method. For experiments 1–5 for which the true number of components was 1, BIC and ICL were very accurate, Stephens's fully Bayesian method was almost as accurate, while AIC, our new UIP, and, most strikingly, DIC, overestimated the number of components considerably.

For experiments 6–10, for which the true number of components was 2, BIC was highly accurate on average for experiments 7–10 but less so for experiment 6. Stephens's method was also accurate on average. AIC overestimated the number somewhat on average for experiments 9 and 10. The other methods, ICL, the UIP and DIC, were much more variable and inaccurate.

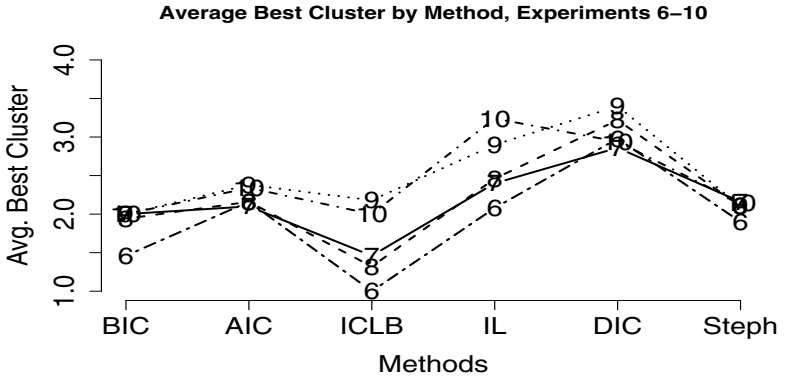
TABLE 4.2. Simulation study: the number of times each of the six model selection criteria chose the correct number of components for the ten experiments in Table 4.1.

Experiment	BIC	Stephens	AIC	ICL	UIP	DIC
1	50	49	45	50	44	20
2	50	48	38	50	39	17
3	50	50	42	50	40	22
4	49	48	34	50	30	14
5	49	46	33	49	19	16
6	23	29	35	0	40	20
7	50	42	46	19	34	23
8	47	45	45	16	33	14
9	50	41	37	39	22	10
10	50	43	39	50	7	20
Total	468	441	394	373	308	176
% Correct	94	88	79	75	62	35

The number of times each method chose the correct number of components is shown in Table 4.2. In Table 4.2, UIP refers to the MAP estimate from the fully Bayesian analysis with the unit information prior in Section 4.1.2. BIC performed best overall, achieving almost perfect accuracy for each experiment except experiment 6. This may be because experiment 6, while a two-component mixture, is very



(a)



(b)

FIGURE 4.6. Average selected number of mixture components by method and experiment. (a) Experiments 1–5, for which the true $G = 1$. (b) Experiments 6–10, for which the true $G = 2$.

close to a single Gaussian, and BIC chose $G = 1$ a relatively high proportion of the time. Stephens's method was second best overall, clearly outperforming the other methods. It also did worse for experiment 6 than for the other experiments.

DIC performed particularly poorly, uniformly across all experiments. This may be due to the fact that it used our unit information prior, but the fully Bayesian method using the UIP performed much better, so this is doubtful.

Mixture modeling is often done to estimate the density rather than to assess the number of components. We therefore compared the performance of the six methods for density estimation. To do this in a comparable way, we chose the value of G selected by each method and then found the MLE of the parameters for that value of G and the corresponding density, and computed its mean integrated squared error.

TABLE 4.3. The average Mean Integrated Squared Error (MISE) for the 10 experiments in Table 4.1. The values in the table are multiplied by 1000.

Experiment	BIC	Stephens	AIC	ICL	UIP	DIC
1	0.19	0.21	0.22	0.19	0.23	0.67
2	0.21	0.24	0.33	0.21	0.31	0.65
3	0.35	0.35	0.41	0.35	0.50	1.32
4	0.48	0.51	1.30	0.48	1.35	2.24
5	0.60	1.00	1.58	0.60	2.75	3.20
6	1.53	1.13	0.86	2.31	0.77	0.76
7	0.23	0.24	0.23	2.18	0.25	0.28
8	0.55	0.39	0.37	2.45	0.42	0.61
9	0.37	0.75	0.47	0.61	0.58	0.77
10	0.34	0.44	0.39	0.34	0.75	0.58
Mean	0.48	0.53	0.62	0.97	0.79	1.11

The results are shown in Table 4.3. Again, BIC did best, followed by the fully Bayesian Stephens's method, and DIC did worst by far.

4.1.5 Discussion

We have considered five established Bayesianly motivated methods for choosing the number of components in a mixture model, and introduced an additional one based on a new unit information prior for mixture models. We compared all six methods using a simulation study whose design was based on a broad survey of mixture model parameter estimates published in the literature before 2000. BIC performed best, quite decisively, with the fully Bayesian method of Stephens (2000a) (similar to that of Richardson and Green 1997) outperforming the other methods. AIC, ICL and DIC performed poorly. So did our own new proposed prior in this context, unfortunately.

There are other Bayesian approaches to choosing the number of components that we have not considered here. Steele (2002) proposed a modification of BIC, defined as follows:

$$\mathrm{BIC}_2 = 2\log(p(y|\hat{\tau},G) + 2\sum_{g=1}^{G}\log(\hat{n}_g + 0.0001) + (G-1)\log(n). \qquad (4.1.1)$$

BIC_2 penalizes the parameters of the mixture components by the logarithm of the estimated sample size of their component rather than of the entire sample size, which seems more in line with the derivation of BIC for regular models. However, BIC_2 performed less well than the raw BIC in Steele's (2002, chapter 3) simulation study, and so we did not include it here.

Fraley and Raftery (2007) proposed a regularized version of BIC. This addressed the problem with BIC that it depends on the maximum likelihood estimator, which

can sometimes be degenerate in mixture models, particularly if there are ties in the data. It replaced the MLE by the posterior mode with a very weakly informative prior, crafted so that the posterior mode hardly differs from the MLE except when the MLE is degenerate. This is implemented in the `mclust` R package. We did not include this in the current comparison, and we conjecture that it would perform similarly to BIC, but perhaps slightly better because of the regularization.

We have addressed only one aspect of the prior specification, namely the prior for the component density parameters. Analysts also have the flexibility to specify different priors on the mixing parameters (for example, a choice between the often used uniform Dirichlet and the Jeffrey's prior (Robert, 2001)), but this has been relatively unexplored in the mixture literature. Stephens (2000a) uses a Poisson prior on the number of components, and it would be interesting to examine how the interplay between the three priors (on the component densities, mixing parameters, and the number of components, respectively) affects inference and can be used by analysts in a constructive way.

We did not consider the class of Dirichlet process mixture priors (Escobar and West, 1995), mostly for reasons of brevity and relevance. The motivation behind the use of the Dirichlet process mixture approach to finite mixture models is mostly to model the underlying population density, rather than to make inference about the number of components. For further discussion of this issue, see Green and Richardson (2001) on the choice between parametric and non-parametric modeling for mixtures and the effect on choosing the number of components.

Gaussian mixture models are often used for clustering, particularly in more than one dimension. Choosing the number of mixture components is not necessarily the same as choosing the number of clusters. This is because a cluster may arise from a non-Gaussian distribution, which is itself approximated by a mixture of Gaussians. Methods for combining mixture components to form clusters have been proposed by Wang and Raftery (2002, Section 4.5), Tantrum, Murua, and Stuetzle (2003), Li (2005), Jörnsten and Keleş (2008), and Baudry et al. (2008).

Acknowledgments: Raftery's research was supported by NICHD grant HD-054511, NIH grant GM-084163 and NSF grant ATM-0724721. Steele's research was supported by a Discovery Grant from the National Resources and Engineering Council of Canada. The authors are grateful to Matthew Stephens for helpful discussions.

4.2 How Large Should the Training Sample Be?

Luis Pericchi

The original approach taken by Berger and Pericchi (1996a,b), the "*Empirical Intrinsic Bayes Factor*," was based on taking (several) real minimal (in size) training samples, and then forming the arithmetic average of Bayes factors. Minimal Training Samples were the most natural, since each training sample used for training

was lost for discrimination of models, and furthermore under several points of view *minimality* of training samples encapsulates the concept of *Unit Information Priors* (Kass and Wasserman, 1995), and *Matching Predictive Priors* (Berger and Pericchi, 2001). The Empirical Intrinsic approach besides solving a problem, that of finding a proper scaling of a Bayes factor, opened up a host of new possibilities for objective hypothesis testing and model selection. Among them: 1) the possibility of using longer than minimal real training samples; 2) the possibility of using Intrinsic Priors directly, without any real training sample (when Intrinsic Priors can be calculated); and 3) the possibility of using simulated training samples, of virtually *any* size smaller than the sample size $M \leq n$. The third possibility offers a "*wonderland*" or "*free lunch*" situation, on which the use of training samples does not "*waste*" real samples. The usual situation in *Cross-Validation* or in the framework of Chakrabarti and Ghosh (2007) is not in Wonderland, since they use real training samples, but Casella and Moreno (2009) for example, are in situation 3, that is in Wonderland. Now the question we address is: How can we decide an *optimal* training sample size $M*$ in Wondeland? For different purposes of the analysis, does the optimal size change?

For simple null hypotheses, we explore two perspectives.

1. The first is the **identification perspective** when the correct identification of the hypothesis is the central focus. We explore rules like "*The Type I Error Rule*" for selecting the (simulated) training sample size or *minimizing the sum of Type I and Type Error II*. We conclude in favor for **extremely** small training samples if not minimal. This validates the original suggestion by Berger and Pericchi (1996a) of using *minimal training samples*.
2. On the other hand, if **prediction of future observations** with square loss by model averaging is the primary goal, that is, the *prediction perspective*, then very large training samples may reduce the error of prediction. For purely prediction purposes, this opens up a window of exciting new methods.

These opposing answers leave us in a drama when the purpose of analysis is not absolutely either identification or prediction. We suggest then a compromise solution on which the errors of identification are kept small, assuring large sample consistency and the error of prediction is drastically diminished. We call it the *5% cubic root* rule on which the compromise training sample M^* is taken as a simple function of the sample size n: $M^* = Min[0.05n, n^{1/3}]$. More general null hypotheses will be addressed elsewhere.

4.2.1 General Methodology

In this section, for simplicity and ease of exposition, we assume that we have a point simple null hypothesis, that is,

$$H_0\colon \theta = \theta_0 \text{ versus } H_1\colon \theta \neq \theta_0.$$

The more general situation is a subject of current research (see Pericchi, 2009).

Our starting point is **The Intrinsic Prior Equation**, which for exchangeable observations (for the sake of simplicity) is (Berger and Pericchi, 1996a,b):

$$\pi^I(\theta) = \pi^N(\theta) \int f(x(l)|\theta) \frac{f(x(l)|\theta_0)}{m_1^N(x(l))} dx(l), \tag{4.2.1}$$

where $x(l)$ is a random training sample of size (length) M and

$$m_1^N(x(l)) = \int f(x(l)|\theta)\pi^N(\theta)d\theta.$$

Exponential Distribution Example. Assume the data comes from an exponential likelihood:

$$f(x_i|\beta) = \beta \exp(-\beta x_i),\ x_i > 0,\ \beta > 0.$$

In this illustration, the hypotheses are:

$$H_0\colon \beta = \beta_0 = 2 \text{ versus } H_1\colon \beta \neq \beta_0.$$

It turns out, using the moment generating function of an exponential distribution, that for a training sample of length M, in terms of the sufficient statistics $\bar{X}_M = \frac{1}{M}\sum_{i=1}^M x_i$,

$$f(\bar{X}_M|\beta) = \frac{(M\beta)^M}{\Gamma(M)} \bar{X}_M^{(M-1)} \exp(-M\beta\bar{X}_M). \tag{4.2.2}$$

The Jeffreys prior for β is $\pi^N(\beta) = 1/\beta$, and it follows that $m_1^n(\bar{X}_M) = \frac{1}{\bar{X}_M}$. Therefore, using expressions (4.2.1) and (4.2.2), the Intrinsic Prior (IPrior) for a training sample of length M is

$$\pi^I(\theta) = \pi^N(\theta) \int f(\bar{X}_M|\beta)[M\beta_0\bar{X}_M]^M \exp[-M\beta_0\bar{X}_M]d\bar{X}_M.$$

It turns out that this integral is

$$\pi^I(\theta) = \frac{\Gamma(2M)}{\Gamma(M)^2} \times \beta_0^M \times \frac{\beta^{(M-1)}}{(\beta_0+\beta)^{(2M)}}.$$

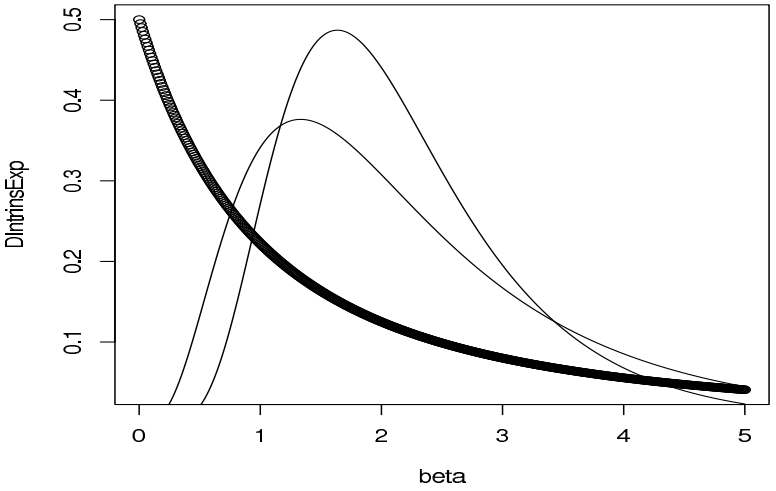
This is a very interesting distribution. In fact,

$$\beta/\beta_0 \sim Beta_2(M,M).$$

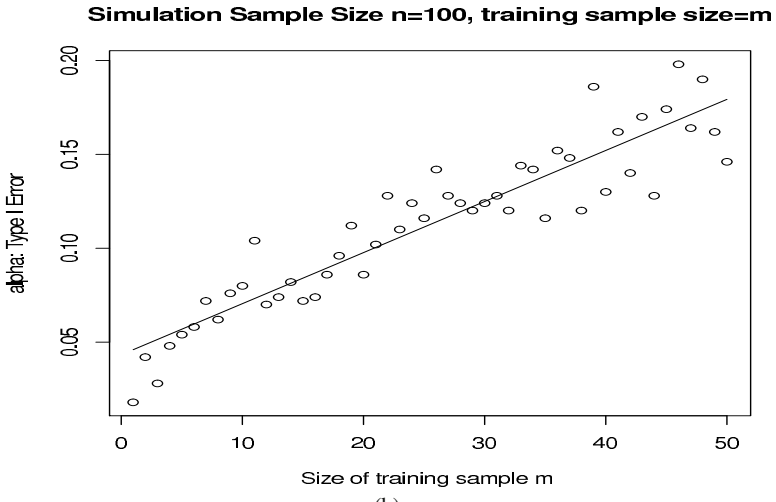
Definition: Beta Distribution of the Second Kind.

$$p(y|p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} \times \frac{y^{p-1}}{(1+y)^{(p+q)}},\ y > 0,$$

denoted by $y \sim Beta_2(p,q)$, and to generate samples from $Beta_2(p,q)$, we can use the following algorithm: (a) $z \sim Beta(p,q)$ and (b) $y = \frac{z}{1-z}$.



(a)



(b)

FIGURE 4.7. (a) Intrinsic priors for the exponential example with minimal training sample $M = 1$ (points) and $M = 5$, $M = 10$ and $H_0 = \beta_0 = 2$; and (b) Type I errors for the exponential example as a function of M.

In Figure 4.7(a), the IPrior is plotted for $M = 1$, minimal training sample, and also for $M = 5$ and 10. All three priors are sensible since apart from being proper, their medians are all equal to $\beta_0 = 2$, so it is balanced and the null hypothesis is in

the "center" of the prior. However, it can be argued that for $M > 1$, the prior is more concentrated around the null hypothesis, which is "more pleasant to the mind." Fine, but....now we have opened a "Pandora Box," how to assign M? One possibility is to consider the class of IPriors for $1 < M < n$, where n is the sample size, this is the approach taken by Casella and Moreno (2009). This is fine if there is robustness with respect to the decision. But this is the exception rather than the rule, since there is an enormous variation in that class of priors. Decision robustness can only be obtained when there is an overwhelming support for one of the hypotheses. It is the intermediate cases which are the most interesting, particularly when the data tend to favor the null hypothesis. This can not be addressed by the whole class of IPriors.

We explore different ways to assign M on a non-trivial set of training samples $1 < M < M_1$. In order to compute the Bayes factor in this example, we need to calculate:

$$m^I(\bar{x}) = \int \beta^n \exp(-\beta n\bar{x})\pi^I(\beta)d\beta.$$

Since this is an involved integral, we resort to the following general approximation (Berger and Pericchi, 1996a)

$$m^I(\bar{x}) = m^N(\bar{x})[\pi^I(\hat{\beta})/\pi^N(\hat{\beta}) + o(1/n)].$$

Using this approximation, it follows that the Bayes factor satisfies approximately the following expression:

$$B_{01}^M = \beta_0^n \exp(-\beta_0 n\bar{x}_n)(\bar{x}_n)^{(n+M)}(\beta_0 + \frac{1}{\bar{x}_n})^{(2M)} \frac{n^n\Gamma(M)^2}{\beta_0^M\Gamma(n)\Gamma(2M)}.$$

We may try to explore here a rule like **Type I Error Rule** as follows. We compute

$$Pr(B_{01}^M \leq 1|H_0 : \beta = \beta_0) = \alpha(M,n),$$

by simulation from H_0 and the range to be considered for M is such that $\alpha(M,n) \leq 0.05$.

We performed extensive simulations in Figure 4.7(b), and we conclude that Type I Error is increasing with M, and thus the optimal is the minimal training sample $M_0 = 1$, although small values of M above $M = 1$ lead to small type I error. An interesting *curiosity* is that a 5% type I error corresponds to a $M = 0.05n$. This *curiosity* will be confirmed in our next important example. Tentatively we consider the class of IPriors:

$$\text{Five PerCent Class: } \{\pi^I(\beta|M), M = Max[1, Min[0.05n, 5]]\}.$$

Remark: It is advisable to avoid letting M be a proportion of the sample size n, since then consistency will be lost when sampling from the null hypothesis, as will be clearly seen in the next example. By bounding the class by 5 as above, consistency, under the null hypothesis, is guaranteed. To continue the exploration of different routes to choose M in a non-trivial manner, we move to a normal example.

4.2.2 An Exact Calculation

We now explore in depth a case amenable to all calculations for a normal mean with known variance.

$$Y_i \sim N(\mu, \sigma_0^2), \text{ with hypotheses: } H_0\colon \mu = \mu_0 \text{ versus } H_1\colon \mu \neq \mu_0.$$

Using the intrinsic prior equation (4.2.1), for a training sample of size M we find

$$\pi^I(\mu) = N(\mu_0, 2\sigma_0^2/M). \tag{4.2.3}$$

Using this prior we get the marginal under H_1 given by

$$m_1^I(\bar{y}|M) = N(\mu_0, \sigma_0^2(\frac{M+2n}{nM})) \tag{4.2.4}$$

and using the intrinsic prior (4.2.3), the Bayes factor is given by

$$2\log(B_{01}) = \log(1+2n/M) - \frac{n(\bar{y}-\mu_0)^2}{\sigma_0^2}\frac{2n}{M+2n}. \tag{4.2.5}$$

Then, under H_0

$$\frac{n(\bar{y}-\mu_0)^2}{\sigma_0^2} \sim \chi_1^2,$$

and under H_1 according to the marginal density (4.2.4),

$$\frac{(\bar{y}-\mu_0)^2}{\sigma_0^2(\frac{M+2n}{nM})} \sim \chi_1^2.$$

4.2.2.1 The Identification Route

We begin setting the problem as an **Identification Problem**, and we compare procedures by (Bayesian versions) of Type I and Type II errors.

Definition: Bayesian Type I Error.

$$\alpha_B(M,n) = Pr[2\log B_{01} < 0|H_0].$$

In the example of this section, since under H_0, $\frac{n(\bar{y}-\mu_0)^2}{\sigma_0^2} \sim \chi_1^2$, it turns out that

$$\alpha_B(M,n) = Pr(\chi_1^2 > (1+f/2)\log(1+2/f)),$$

where $f = M/n$ is the fraction of the sample, which is the training sample, and, for example, $f = 0.05$ is that the training sample is 5% of the sample n.

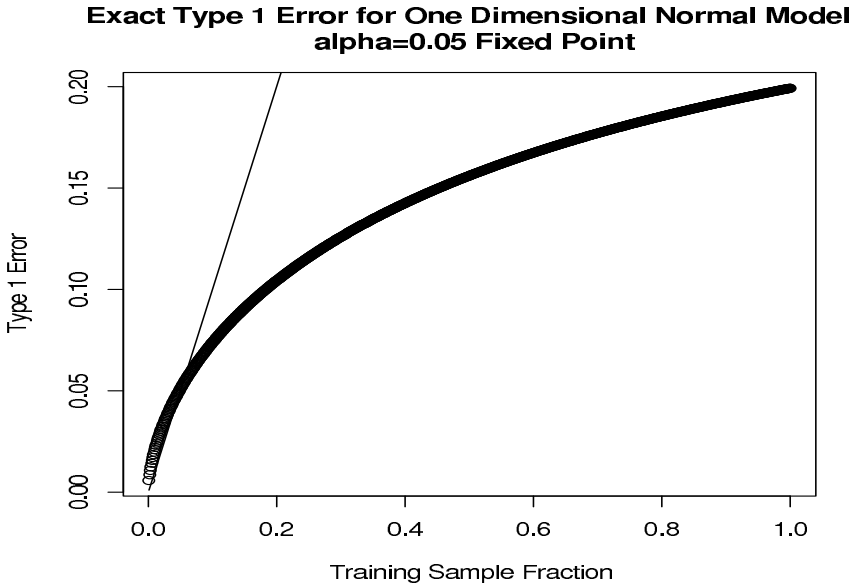


FIGURE 4.8. Bayesian Type I Error as function of M. 5percent proportion produces 5percent type I error.

Definition: Bayesian Type II Error.

$$\beta_B(M,n) = Pr(2\log(B_{01}) > 0|H_1),$$

where the probability is calculated from $m_1(\mathbf{y}|M)$. In the Normal mean example it turns out that

$$\beta_B(M,n) = Pr(\chi_1^2 < (f/2)\log(1+2/f)),$$

since under m_1 it is the case that $\frac{nM(\bar{y}-\mu_0)^2}{\sigma_0^2(M+2n)} \sim \chi_1^2$. The surprising fact is that *both* Type I and Type II Bayesian errors are increasing with f, see Figures 4.8-4.12. In fact, it is very interesting that 5 percent as training sample leads to a near 5 percent of Bayesian Type 1 error, a *curiosity* already noted in the previous example.

Conclusion. To reduce both types of error, we should reduce training samples, that is take Minimal Training Samples, in this case $M = 1$ is optimal *in that sense*. If a compromise has to be reached still the training sample sizes should be rather smaller than large. Alternatively, instead of a fraction we may like to use a power of sample size. The fact is that if we use **any** fraction $f > 0$ of sample size, then under H_0 the $2\log(B_{01})$ is bounded for all sample sizes, an so it is not consistent under H_0. This fact can be easily deduced for example from (4.2.5), for $M/n = f =$constant. For $f = 0.05$, Figure 4.9 tells us that a power of 0.7 approximates for a long range, so if this power is assumed the 5 percent rule is approximately obeyed. Sample size to the power of 0.7 will yield a consistent procedure, but with a high Bayesian Type II

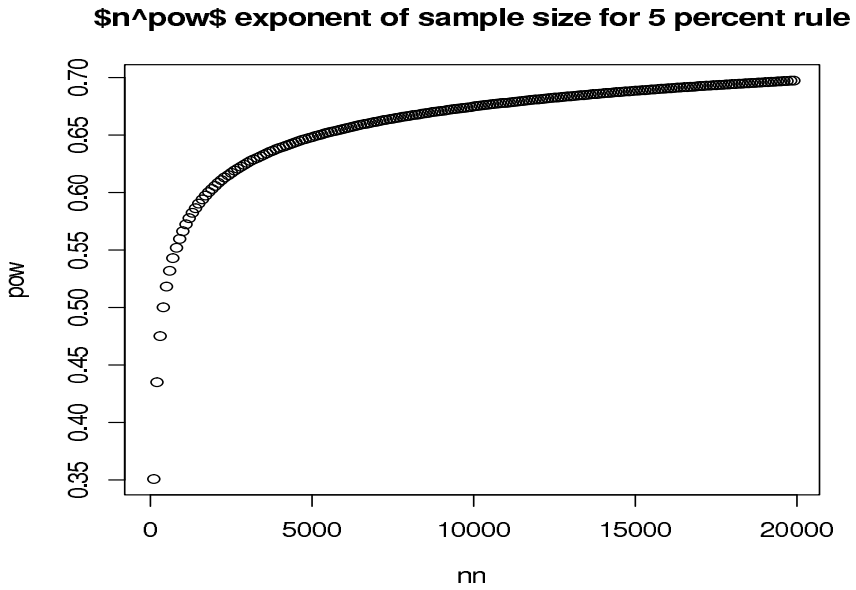


FIGURE 4.9. Power of M equivalent to 5 percent.

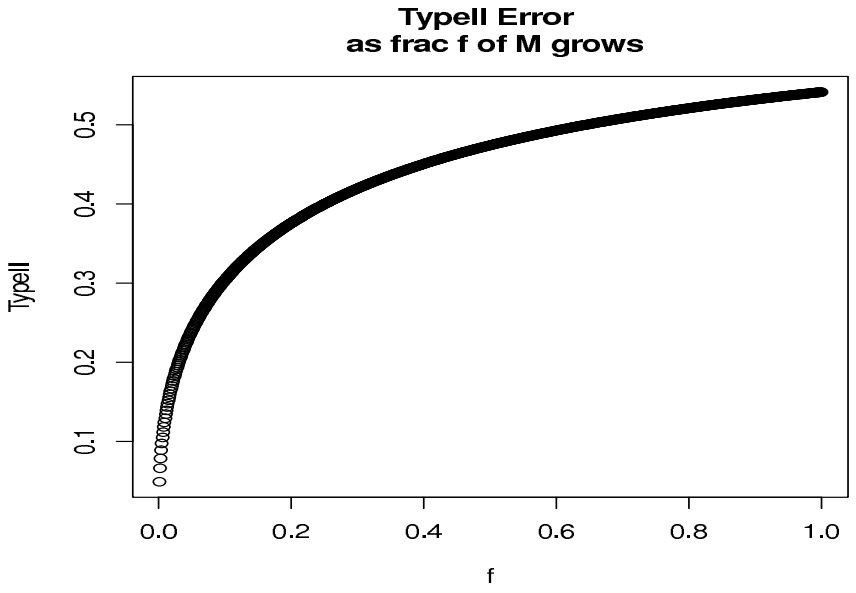


FIGURE 4.10. Bayesian Type II Error as function of M, growing steady.

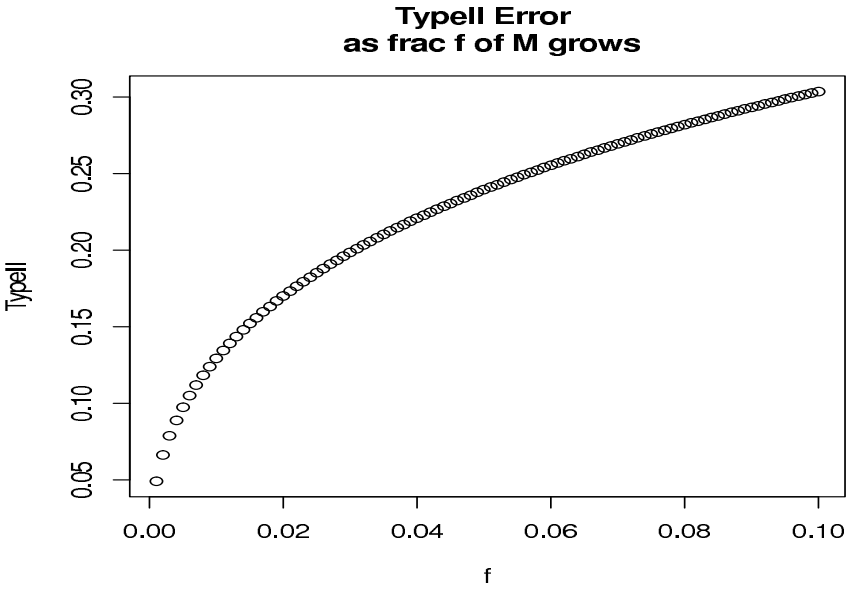


FIGURE 4.11. Bayesian Type II Error as function of M. 5percent proportion produces 25percent type II error.

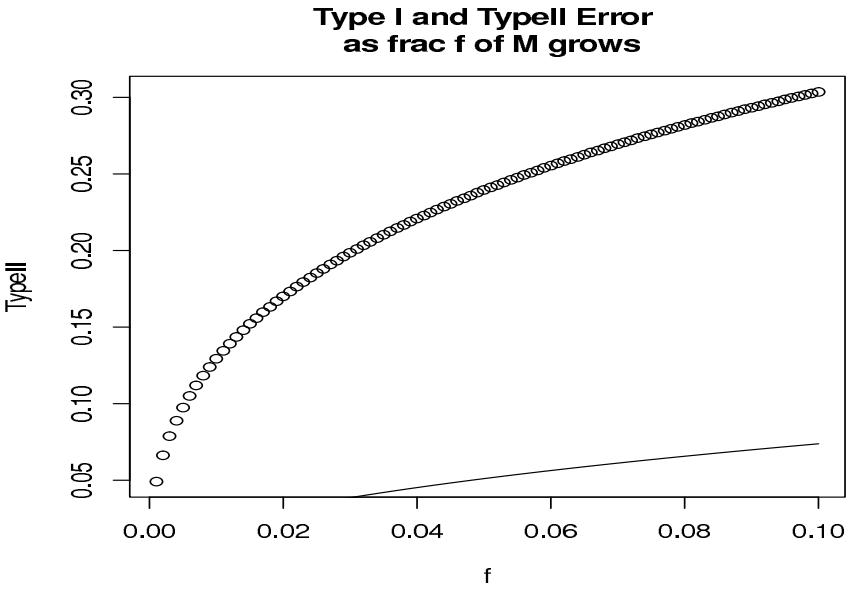


FIGURE 4.12. Bayesian Type I and Type II Error as function of M.

error of about $\beta_B = 0.25$, see Figure 4.10, and with a non-negligible Type I Error of 0.05 even for very large sample sizes. This seems to be unacceptably large. Thus, if we are going to envisage a "*power rule*" that power should be much smaller than 0.7. Next, we turn to a radically different criterion.

4.2.2.2 The Prediction Route

We now take a completely different route: We focus on the Bayesian Risk of predicting the observations, but **NOT** by one single model but by the **fully Bayesian route of Model Averaging**, which is the most widely justifiable procedure under a Bayesian point of view. For which training sample size M will be the Bayesian risk minimized? First of all notice that typically the variance of the log of the Bayes factor decreases with M.

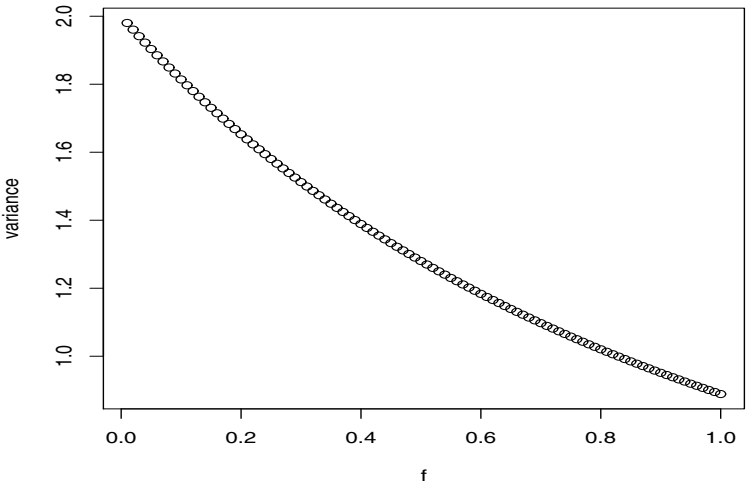


FIGURE 4.13. Variance of 2LogB01 as function of fraction $f = M/n$ to minimize variance increase fraction f.

Normal Example Continued. From the expression (4.2.5), we get that under H_0 the $\mathrm{Var}(2\log(B_{0,1}))$ is $2(\frac{2n}{M+2n})^2$ or $(\frac{2}{2+f})^2$. Under m_1 the factor turns to be $(\frac{2}{f})^2$ both decreasing with M as seen in Figure 4.13. This variance reduction may anticipate an effect in the reduction of error of prediction.

Prediction Criteria. We use square error loss, the most used (over-used?) loss function.

Definition: The Prediction Risk as a function of M.

$$\text{Risk}(M) = \int (\hat{y}^* - \bar{y})^2 p(\bar{y}|M) d\bar{y},$$

where

$$p(\bar{y}|M) = P(H_0|\mathbf{y},M) f(\bar{y}|\mu_0) + P(H_1|\mathbf{y},M) m_1(\bar{y}|M),$$

and

$$\begin{aligned}\hat{y}^* &= P(H_0|\mathbf{y},M) \int \bar{y} f(\bar{y}|\mu_0) d\bar{y} + P(H_1|\mathbf{y},M) \int \bar{y} m_1(\bar{y}|M) d\bar{y} \\ &= P(H_0|\mathbf{y},M)\hat{y}_0^* + P(H_1|\mathbf{y},M)\hat{y_1^*}.\end{aligned}$$

Lengthy algebra yields that

$$\begin{aligned}\text{Risk}(M) = (\hat{y}_0^* - \hat{y_1^*})^2 P(H_0|\mathbf{y},M)\, P(H_1|\mathbf{y},M) + \text{Var}_0(\bar{y}) P(H_0|\mathbf{y},M) \\ + \text{Var}_1(\bar{y}|M) P(H_1|\mathbf{y},M),\end{aligned}$$

where $\text{Var}_0(\bar{y}) = \sigma_0^2/n, \text{Var}_1(\bar{y}|M) = \sigma_0^2(1/n + 2/M)$.

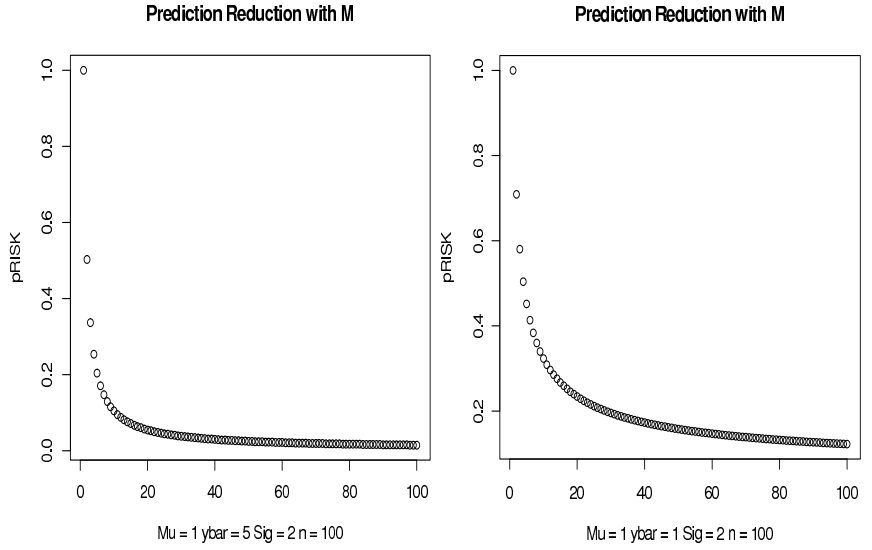


FIGURE 4.14. Effects on the relative risk of the training sample size M, showing rapid decrease with M.

Figure 4.14 shows the reduction of the risk as the training sample size grows. Note however an important phenomenon: the decrease in risk is quite fast. So allowing slightly bigger than minimal training samples permits a drastic reduction of the prediction risk. On the other hand, we see in Figure 4.15 that the change in the Probability of the null may be sizeable with M. Taking into consideration all these

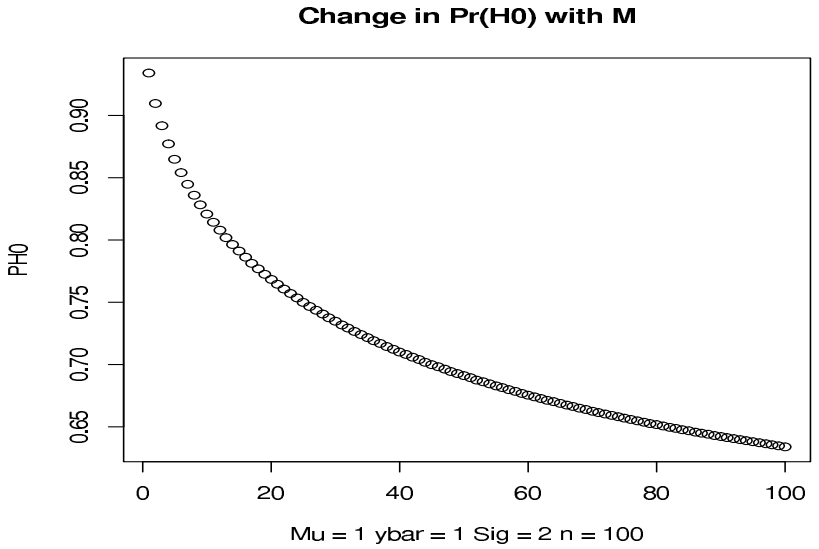


FIGURE 4.15. Effects on the probability of the alternative and the relative risk of the training sample size M.

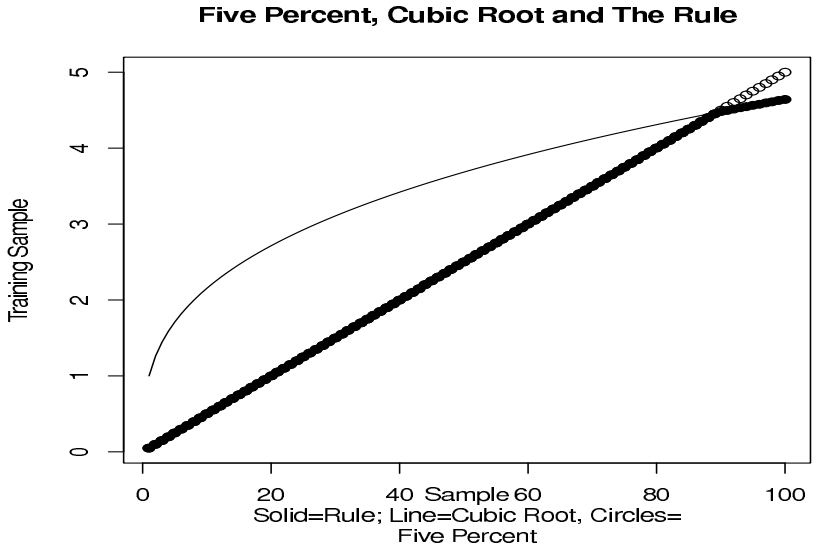


FIGURE 4.16. The five percent-cubic root rule as a function of sample size n.

facts we finally propose the rule:

$$M = min[0.05n; (n)^{1/3}].$$

The 5percent-cubic root rule is depicted in Figure 4.16.

4.2.3 Discussion of the FivePercent-Cubic-Root Rule

Limitations. It should first be noted that we considered situations on which the Null Hypothesis is simple and the alternative is one-dimensional. Generalizations are being worked out.

Versions. Apart from the *continuous* version, we may take *integer* versions, by taking the function "ceiling" in the rule above. There is nothing to prevent the use of fractional M though, like $M = 0.25$ for $n = 5$. The `ceiling` version would lead to $M = 1$. This seems fine, but smaller type I errors seems preferable.

Pure Cubic Root Rule? Setting the rule simply as: $M = (n)^{1/3}$, achieve most of the desiderata of a sensible procedure. Nevertheless the simple cubic root will allow higher values of type I error for very small sample sizes, so our preference for $M = Min[0.05n; (n)^{1/3}]$.

Different Thresholds. It may be argued that the rule depends on the threshold (on the log Bayes factor) of 0 in the definitions of Bayesian Type I and II Errors. However the value of zero is a natural one, and does not seem esential in our arguments which involve type II error. On the other hand, Type I error seems to be fairly flat with respect to the training sample size for thresholds which are smaller than one.

Tentative Nature of the Rule. Still the rule should be contrasted with practical experiences, for fine tuning. However a compromise between minimal and maximal training samples seems useful to be considered.

Acknowledgments: This work was supported in part by NIH Grant P20-RR016470.

4.3 A Conservative Property of Bayesian Hypothesis Tests

Valen E. Johnson

Statistical hypothesis tests can be conducted for a variety of purposes. As Marden (2000) notes, the two (or more) hypotheses under consideration may be on equal footing, or one hypothesis may represent a simplification of a larger model and interest may lie in determining whether the simpler model provides an adequate representation of the data. In classical statistical tests, the null hypothesis occasionally represents a theory that one believes is true, but more often it represents a straw

man that one hopes to reject. In this article, I restrict attention to the latter case and assume that a null hypothesis has been completely specified and that the goal of an investigation is to reject it. If a Bayesian test is used for this purpose, I demonstrate that the expected weight of evidence against the null hypothesis is maximized when the "true" alternative hypothesis has been specified. I discuss the implications of this fact from several perspectives, and demonstrate that it implies, among other things, that a Bayesian cannot skew the results of a hypothesis test in favor of the alternative hypothesis by specifying an overly optimistic (or pessimist) prior distribution on the parameters of interest.

4.3.1 An Inequality

In parametric settings, if $\mathbf{x}$ denotes data collected during a study and θ denotes a parameter of interest, then a classical test of a point null hypothesis is often stated as a test of

$$H_0\colon \theta = \theta_0 \qquad \text{versus} \qquad H_1\colon \theta \neq \theta_0. \tag{4.3.1}$$

The specification of Bayesian tests is more complex, requiring that (proper) prior densities be specified on θ under both the null and alternative hypotheses. A Bayesian hypothesis test corresponding to (4.3.1) might thus be written

$$H_0\colon \theta \sim \pi_0(\theta) \qquad \text{versus} \qquad H_1\colon \theta \sim \pi_1(\theta), \tag{4.3.2}$$

where $\pi_0(\theta)$ assigns unit mass to θ_0, and $\pi_1(\theta)$ represents any other probability density function. For purposes of this article, I assume that $\pi_0(\theta)$ is known and that there is no ambiguity regarding its functional form. It need not, however, represent a density function that assigns unit mass to a single point—any probability density function will do.

In contrast to the specification of $\pi_0(\theta)$, there is often controversy surrounding the specification of $\pi_1(\theta)$. For example, in testing whether two variables are correlated, the null hypothesis is usually specified as a statement that the correlation coefficient is 0. What to assume about the correlation coefficient under the alternative hypothesis is generally less clear.

To complete the specification of the hypothesis test indicated in (4.3.2), let $f(\mathbf{x}\,|\,\theta)$ denote the sampling density of $\mathbf{x}$ given θ. I assume that $\pi_1(\theta)$ is specified before $\mathbf{x}$ is observed.

In a Bayesian hypothesis test, the posterior odds in favor of the alternative hypothesis are obtained by multiplying the prior odds in favor of the alternative hypothesis by the Bayes factor. This relation can be expressed mathematically as

$$\frac{\mathbf{Pr}(H_1\,|\,\mathbf{x})}{\mathbf{Pr}(H_0\,|\,\mathbf{x})} = \frac{m_1(\mathbf{x})}{m_0(\mathbf{x})} \times \frac{\alpha}{1-\alpha}, \tag{4.3.3}$$

where α is the prior probability assigned to the alternative hypothesis and $\alpha/(1-\alpha)$ are the prior odds in favor of the alternative hypothesis. The quantities $m_1(\mathbf{x})$ and $m_0(\mathbf{x})$ denote the marginal densities of the data under the alternative and null hypotheses, respectively, and their ratio is called the *Bayes factor*. The marginal density of the data under the alternative hypothesis is defined to be

$$m_1(\mathbf{x}) = \int_{\Theta} f(\mathbf{x}\,|\,\theta)\pi_1(\theta)\,d\theta, \tag{4.3.4}$$

which is just a weighted average of the sampling density of the data, weighted with respect to the prior density on the unknown parameter vector. A similar expression provides the marginal density of the data under the null hypothesis. The Bayes factor, $m_1(\mathbf{x})/m_0(\mathbf{x})$, may thus be regarded as a ratio of integrated likelihood functions.

Often, the null and alternative hypotheses are assumed to be equally likely *a priori*, which means that $\alpha = 0.5$ and $\alpha/(1-\alpha) = 1$. In this case, the Bayes factor equals the posterior odds in favor of the alternative hypothesis.

The logarithm of the Bayes factor is called the *weight of evidence* and is used by Bayesians to summarize the result of a hypothesis test.

The expected weight of evidence in a hypothesis test has the following property. Suppose that the prior density $\pi_1(\theta)$ assumed for the parameter vector θ under the alternative hypothesis is "incorrect," and that the "true" prior density for θ is $\pi_t(\theta)$. The meanings of "incorrect" and "true" in this context are discussed below. Let $m_t(\mathbf{x})$ denote the marginal density of $\mathbf{x}$ obtained from (4.3.4) with $\pi_t(\theta)$ substituted for $\pi_1(\theta)$. Then from Gibbs' inequality, it follows that

$$\begin{aligned}\int_{\mathscr{X}} m_t(\mathbf{x}) \log\left[\frac{m_t(\mathbf{x})}{m_0(\mathbf{x})}\right] d\mathbf{x} - \int_{\mathscr{X}} m_t(\mathbf{x}) \log\left[\frac{m_1(\mathbf{x})}{m_0(\mathbf{x})}\right] d\mathbf{x} \\ = \int_{\mathscr{X}} m_t(\mathbf{x}) \log\left[\frac{m_t(\mathbf{x})}{m_1(\mathbf{x})}\right] d\mathbf{x} \\ \geq 0.\end{aligned}$$

That is,

$$\int_{\mathscr{X}} m_t(\mathbf{x}) \log\left[\frac{m_t(\mathbf{x})}{m_0(\mathbf{x})}\right] d\mathbf{x} \geq \int_{\mathscr{X}} m_t(\mathbf{x}) \log\left[\frac{m_1(\mathbf{x})}{m_0(\mathbf{x})}\right] d\mathbf{x}. \tag{4.3.5}$$

The last inequality states that the expected weight of evidence in favor of the alternative hypothesis is always greater if the true prior is used in its definition. This is a slightly different way of stating that the expected weight of evidence always favors the true hypothesis (e.g., Good, 1950), but the introduction of the device of a "true" prior is useful for examining the effects of misspecifying the prior density used to define the alternative hypothesis is (4.3.2).

4.3.2 Discussion

Equation (4.3.5) can be interpreted in a number of ways. From a frequentist perspective, the true value of a model parameter is assumed to be an unknown constant. The true prior density is then the density function that places unit mass on this value. The more mass that $\pi_1(\theta)$ assigns away from this value, the higher the penalty the Bayesian pays for not concentrating his prior on it. In other words, the posterior odds computed in favor of the alternative hypothesis using any prior that does not concentrate its mass on the true (but unknown) parameter value are very likely to be smaller than the posterior odds that would be achieved by using the correct prior.

The interpretation of (4.3.5) from a Bayesian perspective is more direct. When planning an experiment with the goal of rejecting a null hypothesis, it is in the investigator's best interest to define the alternative hypothesis using the prior distribution that most accurately represents available prior information.

It is worth noting that many objective Bayesian testing procedures use non-informative prior densities to define alternative hypotheses. Presumably this is done in an effort to avoid "biasing" test results in favor of the alternative hypothesis. In fact, however, (4.3.5) demonstrates that the use of such priors inevitably results in a decrease in the expected weight of evidence that would be collected in favor of a true alternative hypothesis. The prior density used to define the alternative hypothesis should be based on the most accurate information available regarding the value of the parameter of interest, *under the assumption that the null hypothesis is false.* That is, the prior density used to define the alternative hypothesis should represent the alternative hypothesis being tested, not a disperse version of the null hypothesis.

Equation (4.3.5) cannot be extended to obtain a similar inequality on posterior model probabilities. Nonetheless, for an accepted value of the prior odds, (4.3.5) might be interpreted to mean that the posterior probability assigned (by another Bayesian) to the alternative hypothesis represents (for you) a conservative estimate of the true posterior probability of the alternative hypothesis. In fact, a plausible rule of thumb is to assume that the true posterior probability of the alternative hypothesis will usually lie between the posterior probability reported by another Bayesian and the posterior probability obtained by setting the Bayes factor equal to the likelihood ratio statistic (i.e., by letting $\pi_1(\theta)$ concentrate its mass on the maximum likelihood estimate of θ).

As a final comment, it is worth contrasting Bayesian testing procedures to Bayesian inferential procedures used in testing contexts. Because Bayesian testing procedures require the specification of proper prior distributions on parameters of interest under both null and alternative hypotheses, it is common practice to use Bayesian inferential procedures in place of formal hypothesis tests even in what are inherently testing problems. For example, instead of testing whether a new medical treatment has a higher efficacy rate than a standard treatment, many scientists will instead report the posterior probability that the new treatment's efficacy rate exceeds the rate assumed for the standard treatment. Such posterior probabilities are often calculated using a non-informative prior density on the new therapy's efficacy rate. Equation (4.3.5), however, reveals an important distinction between Bayesian test-

ing procedures and Bayesian inferential procedures. A simple example illustrates this difference.

Suppose that an (unscrupulous) investigator wishes to reject a null hypothesis that a success probability p is equal to, say, 0.2. To this end, he imposes a prior distribution on p that concentrates all of its mass on the interval $(0.8, 1.0)$, and then collects a binomial observation x with denominator n and success probability p. Because the investigator assigned no prior mass outside of the interval $(0.8, 1.0)$, his posterior distribution on p will also concentrate all of its mass on the interval $(0.8, 1.0)$. The posterior probability that $p < 0.8$ is then 0, no matter what the value of x is.

The situation is fundamentally different when viewed as a statistical test. Suppose, for instance, that $x = 40$ and $n = 100$. Assuming that the prior odds between null and alternative hypotheses are 1.0 and using the same prior distribution on p as before, a simple calculation shows that the posterior probability assigned to the alternative hypothesis is less than 9×10^{-13}, which for practical purposes is 0. Thus, while the inferential procedure concludes that the probability that $p \geq 0.8$ is 1, the testing procedure concludes that the probability $p \geq 0.8$ is essentially 0.

Taking this example a bit further, suppose that the investigator had defined the alternative hypothesis by instead assuming that p had a uniform distribution on the unit interval. With this prior, the posterior probability assigned to the alternative hypothesis would have been 0.9998, or very close to 1. By attempting to bias the result of the test against the null hypothesis, the investigator increased the weight of evidence collected in its favor.

4.4 An Assessment of the Performance of Bayesian Model Averaging in the Linear Model

Ilya Lipkovich, Keying Ye, and Eric P. Smith

Bayesian model averaging (BMA) is an approach in modern applied statistics that provides data analysts with an efficient tool for discovering promising models and obtaining estimates of their posterior probabilities via Markov chain Monte Carlo (MCMC) (see Hoeting et al., 1999). These probabilities can be further used as weights for model averaged predictions and estimates of the parameters of interest. As a result, variance components due to model selection are estimated and accounted for, contrary to the practice of conventional data analysis (such as, for example, stepwise model selection). In addition, variable activation probabilities can be obtained for each variable of interest.

BMA (and the frequentist counterpart) is becoming an increasingly popular data analysis tool which allows the data analyst to account for uncertainty associated with the model selection process. An interesting application of BMA methodology in the context of linear regression is a method of simultaneous variable selection and

outlier detection presented in Hoeting, Raftery, and Madigan (1996) and Hoeting et al. (1999).

Prior to model averaging, variable selection has been recognized as "one of the most pervasive model selection problems in statistical applications" (George, 2000), and a lot of different methods emerged during the last 30 years, especially in the context of linear regression (see Miller, 2002; McQuarrie and Tsai, 1998). Many researchers focused on developing an appropriate model selection criterion assuming that few reasonable models are available. Such methods include PRESS (Allen, 1971), Mallows' C_p (Mallows, 1973), Akaike's AIC (Akaike, 1973), Schwarz's BIC (Schwarz, 1978), RIC (Foster and George, 1994), bootstrap model selection (Shao, 1996), and others. In reality, however, the researchers often have to choose a single or few best models from the enormous amount of potential models using techniques such as stepwise regression of Efroymson (1960) and its different variations, or, for example, the leaps-and-bounds algorithm of Furnival and Wilson (1974).

Typically researchers use both approaches, first trying to generate several best models for different numbers of variables, and then select the best dimensionality according to one of the criteria listed. Any combination of these approaches to model selection, however, do not seem to take into account the uncertainty associated with model selection and therefore in practice tend to produce overoptimistic and biased prediction intervals, as will be discussed later. In addition, the statistical validity of various variable selection and elimination techniques (stepwise and forward selection, backward elimination) is suspect. The computations are typically organized in "one variable at a time" fashion seemingly employing the statistical theory of comparing two nested hypotheses, however ignoring the fact that the true null distributions of the widely used "F statistics" (such as F-to-enter) are unknown and can be far from the assumed F distribution (see Miller, 2002).

The two elements of the model selection problem (model search and model selection criterion) are naturally integrated in model averaging. This overcomes the inherent deficiency of the deterministic model selection approach by combining (averaging) information on all or a subset of models when estimating parameters, making inferences, or predicting new observations, instead of using a single model.

The technical approach to BMA is relative straightforward and involves updating information across a set of possible models. Following Madigan and Raftery (1994), if Δ is the quantity of interest, such as a parameter of the regression model or a future observation, then its posterior distribution given data D and a set of K models is a mixture of posterior distributions (see also Leamer, 1978):

$$P(\Delta|D) = \sum_{i=1}^{K} P(\Delta|M_k, D)P(M_k|D),$$

the posterior probability for model M_k is given by

$$P(M_k|D) = \frac{P(D|M_k)P(M_k)}{\sum_{i=1}^{K} P(D|M_i)P(M_i)},$$

where

$$P(D|M_k) = \int P(D|\beta_k, M_k) f(\beta_k|M_k) d\beta_k$$

is the predictive distribution for model M_k.

The standard Bayesian approach to quantifying uncertainty is by incorporating it as an extra layer in the hierarchical model. In BMA we assume an extension of the Bayesian hierarchical model, as explained in George (1999), "By using individual model prior probabilities to describe model uncertainty, the class of models under consideration is replaced by a single large mixture model. Under this mixture model, a single model is drawn from the prior, the prior parameters are then drawn from the corresponding parameters' priors and finally the data is drawn from the identified model."

In this section, we will try to think about models in a broad context when appropriate since different researchers have applied BMA within quite different classes of models. In many cases however the model space will be reduced to the subsets of predictors in the linear regression model. The objective of this research is to use simulation to assess BMA. Using a basic multiple regression model, we evaluate the efficiency of BMA relative to a "true" model and a single selected model. We use simulation experiments to evaluate the performance of the Bayesian Model Averaging method by comparing averaged linear models against a single model. We compare performance over different subsets of models and evaluate the importance of correlation among predictors on efficiency.

4.4.1 Assessment of BMA Performance

One of the main arguments for using the BMA is based on its ability to improve our predictions, as measured by the out-of-sample prediction error. Summarizing the experience of several studies, Raftery (1995) state that "in most cases BMA improves predictive performance over the single best model by about the same amount as would be achieved by increasing the sample size by 4%."

Hoeting et al. (1999) present several examples of BMA applications that use out-of-sample validation methods to illustrate its predictive performance. The cross-validation was performed by splitting each data set into two parts: training set, D^T and prediction set, D^P. The training set is used for model selection and the second set for prediction. Using several models turned out better than using a single model in most of the cases.

Two measures of predictive ability were used, the coverage for a 90% predictive interval, measured by the proportion of observations of the second set falling within the 90% of the corresponding posterior prediction interval (see Hoeting et al., 1999). The second measure is the logarithmic scoring rule of Good (1952). Specifically they measure the predictive ability of a single model M as

$$-\sum_{\Delta \in D^P} \log\left\{P(\Delta|M, D^T)\right\},$$

and compare it with predictive ability of BMA as measured by

$$-\sum_{\Delta\in\mathscr{D}^P}\Big\{\log\Big[\sum_{k=1}^{K}P(\Delta|M_k,D^T)P(M_k|D^T)\Big]\Big\}.$$

The smaller the predictive log score for a given model or model average, the better the predictive performance.

An intuitive explanation of the superior performance of BMA was given in George (1999), who noted that BMA-based prediction can be viewed as an application of averaging of several approximately unbiased estimates with weights adaptively accounting for their different variance, hence better prediction. A more analytical argument was used in Madigan and Raftery (1994), which follows from the non-negativity of the Kullback-Leibler information divergence:

$$-E\Big[\log\Big\{\sum_{k=1}^{K}P(\Delta|M_k,D^T)P(M_k|D^T)\Big\}\Big]\leq -E[\log\{P(\Delta|M_j,D)\}]$$

for $j=1,\ldots,K$. See also Chickering and Heckerman (2000) who used Bayesian networks and found that the predictive ability of model averaging is better than using a single model.

4.4.2 A Simulation Study of BMA Performance

4.4.2.1 Out-of-Sample Performance of BMA

To assess the performance of BMA in the case of a multiple linear regression, we conducted a simulation study. The idea of simulation was to use some "true" model to generate the data and then compare the out-of-sample prediction for the averaged model to that of various candidate models (full model, top selected model, and true model). Of course, the usefulness of the results is limited because the simulation is based on the assumption that a "true" model exists, which may not be the case in a real-life situation. However, the simulation study can be used as a way of validating the BMA procedure.

The simulation procedure started at generating training and prediction data, denoted by D^T and D^P, respectively under the same generation scheme. The number of explanatory variables was taken to be 20 and the number of cases in both training and prediction data sets was set to 100, and 500. Both fixed and random design matrices (X) were considered in the simulation experiment. For the fixed design the same (generated from a multivariate normal distribution) values of explanatory variables for both training and prediction data sets were used; for the random design different X matrices for the two data sets were generated. The design matrix had one of three correlation patterns: high ($r=0.8$), medium ($r=0.5$), and uncorrelated. For this purpose, variables were divided into 4 groups (x_1-x_5), (x_6-x_{10}), ($x_{11}-x_{15}$),

$(x_{16} - x_{20})$, with equal correlations within each group and uncorrelated across the groups. All x's were standardized to have zero means and unit variances. The values for response variable, y's, were generated by applying the same coefficients for those variables in the same group, $\beta \in (0.1, 0.2, 0.3, 0.0)$ and adding normal errors with zero means and unit variances. Notice that the coefficients in the last group are set to zero, which means that the full model is not the true model. To ensure that different data sets are comparable (have similar R^2), the β coefficients were adjusted by the appropriate factor, c so as to make the R^2 for the simulated data to be 0.5 when the true coefficients are applied. This factor is data-dependent and it is computed as

$$c = \left\{ N^{-1} \sum_{i=1}^{N} \left(\sum_{j=1}^{20} \beta_j x_{ij} \right)^2 \right\}^{-1/2},$$

where N is the sample size, β's are the assumed model coefficients and the x's are simulated. This can be motivated as follows: for a regression model with random design, unit variance of standard error, and true coefficients, β, the expected R^2 is computed as

$$\tilde{R}^2 = \frac{E\left(\sum_{j=1}^{p} \beta_j x_j \right)^2}{1 + E\left(\sum_{j=1}^{p} \beta_j x_j \right)^2},$$

(see Breiman, 1996), and hence the correction factor needed to achieve a certain target R_t^2 can be estimated as

$$c = \frac{R_t}{\sqrt{N^{-1} \sum_{i=1}^{N} \left(\sum_{j=1}^{p} \beta_j x_{ij} \right)^2 (1 - R_t^2)}}.$$

Next we ran 20,000 iterations of the Markov chain Monte Carlo model composition (MC^3, see Madigan and York, 1995) algorithm on the training data and obtain a list of models and associated values of the Bayesian Information Criterion (BIC) (Schwarz, 1978). Then Occam's window was applied to reduce the model set by dropping the models whose BIC were worse by a certain amount than that of the model with highest BIC (see Raftery, 1995, p. 31 for guidelines about interpreting differences in BICs). For this simulation the size of the window was varied from 5 to 25 units of BIC.

In the next step, we estimated the regression coefficients for the model from the model set using training data (D^T). Prediction error measures, based on the differences between actual and fitted values were computed. Let

$$PE_M(A, B) = \sum_{i=1}^{N} \left(y_i^B - \sum_{j \in M} x_{ij}^B \hat{\beta}_j^A \right)^2, \tag{4.4.1}$$

be a measure of prediction error for model M fitted to data A and applied to predict y's in data B. This means that the coefficients for a model M are first estimated on data set A and further applied to the explanatory variables in the data B. This quantity can be computed for the top model (model with highest BIC found by our MC^3 algorithm), full model and the "BMA model." The fitted vales for the "BMA models" are the weighted averages of predictions obtained by using individual models, $\sum_{M\in\mathscr{M}}\left(\sum_{j\in M} x_{ij}^B\hat{\beta}_j^A\right)\hat{P}(M|D)$, where $\mathscr{M}$ is the set of active models after applying Occam's window. The model weights were computed using the BIC approximation to posterior model probabilities (see Raftery, 1995).

Once we have all of the above, for the top, full, BMA and the true models, we computed $PE_{top}(D^T,D^P)$, $PE_{full}(D^T,D^P)$, $PE_{BMA}(D^T,D^P)$, and $PE_{true}(D^T,D^P)$ respectively as defined in (4.4.1). These quantities are used to compute the relative efficiency of the averaged model when applied to the new observations. Of course, we hope that BMA will be efficient when applied not only to the data used for estimation but also to a new data. Finally, we express the relative efficiency of BMA with respect to the top, full, and true models as the ratios

$$\frac{PE_{top}(D^T,D^P)}{PE_{BMA}(D^T,D^P)},\ \frac{PE_{full}(D^T,D^P)}{PE_{BMA}(D^T,D^P)},\ \text{and}\ \frac{PE_{true}(D^T,D^P)}{PE_{BMA}(D^T,D^P)}.$$

For example, a relative efficiency of 1.07 means that using BMA is equivalent to a 7% increase in sample size.

TABLE 4.4. Summary for BMA out-of-sample performance (relative efficiency) against top, full and true models. Occam's window for computing BMA was set to 25 units. The ratios are averages over 200 iterations, standard errors are in parentheses.

		Relative Efficiency of BMA		
X-Design	Correlation among X's	PE_{top}/PE_{BMA}	PE_{full}/PE_{BMA}	PE_{true}/PE_{BMA}
Random	High (0.8)	1.054 (0.009)	1.057 (0.014)	0.979 (0.008)
	Medium (0.5)	1.083 (0.014)	1.007 (0.014)	0.909 (0.012)
	Low (0.0)	1.085 (0.008)	1.075 (0.017)	0.937 (0.015)
Fixed	High (0.8)	1.083 (0.037)	1.007 (0.014)	0.909 (0.012)
	Medium (0.5)	1.059 (0.024)	0.993 (0.009)	0.918 (0.007)
	Low (0.0)	1.071 (0.022)	1.028 (0.012)	0.953 (0.008)

The described procedure was repeated 200 times for each scenario, the results are summarized in the series of tables which contain the averages based on 200 replicates and the estimates of associated Monte Carlo standard errors in parentheses. Table 4.4 gives the relative efficiencies for BMA approach compared to the top, full and true models. For this part of the experiment, Occam's window was set to 25 units which, on average, translated into about 4,000 models captured. We see that BMA tends to outperform both full and top models in terms of the out-of-sample prediction.

Of course, BMA fails to outperform the true model. However, in real life we never know the true model, if such exists, and that is why we are using BMA in the first place. The fact that BMA outperforms the full model is of course due to a relatively small number of observations in the sample, $N = 100$. When we tried the same experiment with $N = 500$ (the results are not reported here), the advantage of BMA vanished. This is because with a large sample the coefficients of the irrelevant variables can be estimated accurately in the full model. In our results, there does not seem to be a difference between the correlated and independent designs, which seems to be counterintuitive. When the predictor variables are correlated, there is some overlap between them and, consequently, there is an overlap between different models so we would expect to see an increased efficiency of BMA. In other simulations, we tried to introduce some x's which were unrelated to y (the regression coefficients were set to zero), while they could maintain high correlations with other x's related to y via their non-zero coefficients, and therefore affect y indirectly. We wanted to see if BMA would prove more efficient in the situation of correlated design, picking on this extra information about y that could be provided by "irrelevant" x's. However, as in the present simulation, the efficiencies of BMA for both correlated and uncorrelated design matrices were about the same. Interestingly, BMA appeared more efficient compared to using the full model when the design is random.

TABLE 4.5. Summary of BMA out-of-sample performance when using different size of Occam's window (random design and high correlations, standard errors are in parentheses).

Occam's window size	Average number of models contained in model set	Relative Efficiency of BMA	
		PE_{top}/PE_{BMA}	PE_{full}/PE_{BMA}
5	106.60	1.033 (0.006)	1.039 (0.016)
10	1014.5	1.035 (0.007)	1.037 (0.011)
15	2586.2	1.048 (0.009)	1.067 (0.013)
20	3902.3	1.051 (0.010)	1.079 (0.013)
25	4551.2	1.054 (0.009)	1.057 (0.014)

Table 4.5 shows the simulation summaries for the case of random design, high correlations, and with different widths of Occam's window (the difference in BIC between the best and the worst model included in the model set). The second column contains the number of models captured with the Occam's window, averaged over 200 runs. As one can see, increasing the size of the window somewhat improves the performance of BMA. There is some evidence that after the number of models in the set exceeds some optimum level, the performance begins to decrease. In summary, our experiment supports the claim made in Raftery (1995, p. 147) that "taking account of model selection uncertainty yields better out-of-sample predictive performance than any one model that might reasonably have been selected. This is true whether one averages across all models or used Occam's window." As our study shows, averaging across only 100 models with top BIC gives about the same improvement in terms of predicting power, as using thousands of available models.

4.4.2.2 The Limits of BMA

Some authors (Hoeting et al., 1999) made a point that averaging across a large set of models is better than using a single model because a single model is always conditional on a single data set which may reflect the idiosyncrasy of this particular data that may not be seen in new replications. They also suggested that the "averaged" model would allow a researcher to overcome this limitation. It appears, however, that since the model weights computed in BMA are also based on a single data set, they would be more suitable to this particular data and therefore may not perform as well when applied to some new data. To show that the "averaged model" is giving a false sense of performance when applied to the same data that was used to select the set of models and determine their posterior weights, we designed a special procedure as follows.

First we construct the model weights by running the MC^3 procedure on the training data, as in the previous simulation experiment. Then we observe that if the average of all these models were used to predict the same training data, its apparently good performance could be explained by the fact that (i) the model coefficients were estimated using the same data set as the one used for prediction and (ii) the model weights were obtained using the same data set as the one used for prediction. To remove the first source of overfitting, (i), from our analysis and evaluate the role of the second source, (ii), we developed the following procedure. First we estimate each model in the model set using an independent data set (the prediction set, D^P) and then apply these coefficients to the original training set, using the same model weights as were obtained on that training data. The prediction error, $PE_{BMA}(D^P, D^T)$, has to be compared with our previous estimate of "out-of-sample" prediction error, $PE_{BMA}(D^T, D^P)$. Now we argue that if the former is systematically smaller than the latter, the only explanation would be that the model weights were somehow adapted to the idiosyncrasies of the training data and therefore may not generalize well to the new data. For example, it may happen that some variable which has a large coefficient in the "true" model did not come out very significant when estimated in the training sample. This would affect the estimate of its activation probability (sum of weights for the models where this variable was included) obtained from that same sample. Hence, the composition of model weights may be not optimal for predicting observations from another sample of the same population.

To put it more technically, we compute the "backward" prediction errors:

$$PE_{true}(D^P, D^T),\ PE_{full}(D^P, D^T), \text{ and } PE_{BMA}(D^P, D^T)$$

and use them to obtain the following ratios,

$$\frac{PE_{BMA}(D^T, D^P)}{PE_{BMA}(D^P, D^T)},\ \frac{PE_{top}(D^T, D^P)}{PE_{top}(D^P, D^T}, \text{ and } \frac{PE_{full}(D^T, D^P)}{PE_{full}(D^P, D^T)}.$$

Notice that the PEs in both numerator and denominator use different data sets for estimating model parameters and prediction, the only asymmetry being that the prediction error in the denominator is computed when the same data are used for model

selection and prediction (note that D^T is always used for model(s) selection). Therefore, ratios higher than 1.0 for BMA would indicate that it is "working better" when the same data are used for prediction and model selection. This difference should obviously vanish for the full or true models (since there is no model selection here), hence we expect the corresponding ratio to be centered around 1.0. Our hope is that BMA would also produce a ratio close to 1.0, or at least lower than that produced by the top model.

TABLE 4.6. Summary of BMA performance against the top, full, and true models when the same data are used for model selection and prediction, standard errors are in parentheses.

		Relative Efficiency of BMA		
X-Design	Correlation among X's	PE_{top}/PE_{BMA}	PE_{full}/PE_{BMA}	PE_{true}/PE_{BMA}
Random	High (0.8)	1.072 (0.010)	1.451 (0.027)	1.343 (0.020)
	Medium (0.5)	1.132 (0.013)	1.533 (0.033)	1.403 (0.026)
	Low (0.0)	1.030 (0.007)	1.129 (0.016)	1.006 (0.011)
Fixed	High (0.8)	1.132 (0.013)	1.533 (0.033)	1.403 (0.026)
	Medium (0.5)	1.125 (0.009)	1.463 (0.024)	1.355 (0.021)
	Low (0.0)	1.027 (0.005)	1.096 (0.014)	1.027 (0.010)

TABLE 4.7. Summary of over-optimism for BMA and the top model when predicting the same data that were used for model selection. Occam's window for BMA set to 25 BIC units. The ratios are averaged over 200 runs, standard errors are in parentheses.

		Measure of over-optimism due to model selection, ratio of $PE(D^T, D^P)/PE(D^P, D^T)$			
X-Design	Correlation among X's	BMA	Top Model	Full Model	True Model
Random	High (0.8)	1.432 (0.036)	1.394 (0.037)	1.033 (0.024)	1.031 (0.022)
	Medium (0.5)	1.593 (0.040)	1.512 (0.039)	1.042 (0.028)	1.026 (0.027)
	Low (0.0)	1.046 (0.026)	1.075 (0.024)	0.969 (0.018)	0.949 (0.020)
Fixed	High (0.8)	1.593 (0.040)	1.512 (0.039)	1.042 (0.028)	1.026 (0.027)
	Medium (0.5)	1.513 (0.038)	1.424 (0.037)	1.023 (0.019)	1.024 (0.022)
	Low (0.0)	1.056 (0.032)	1.083 (0.030)	0.969 (0.021)	0.962 (0.024)

The simulation results are summarized in Tables 4.6 and 4.7. Table 4.6 is analogous to the Table 4.4, showing the out-of-sample prediction error for BMA against prediction errors for the other methods. The only difference is that now the set of models used for BMA is determined with the same data set that is used for prediction. Similarly, we let the top model predict the data that were used in selecting this model. There should be no special advantage for the full and true models in switching the two data sets. The size of Occam's window was set to the highest value, 25, which on average corresponded to 4,600 models. Interestingly, the performance of BMA against that of the true and full models has improved dramatically when compared to the results shown in Table 4.4. This "improvement" is an artifact, since the "BMA model" is now put in the advantageous position. Surprisingly, the ratio

PE_{top}/PE_{BMA} is also higher now (for most of the scenarios), though one would expect that it is the top model that should receive the greatest advantage when using the using the same data for prediction and model selection.

Table 4.7 shows that, as we expected, the ratio for the full model is around unity and the ratio for the model with top BIC is higher when it is applied to the training set. Observe, however, that our measure of BMA over-optimism is still rather high, and in most of the cases is even higher than that for the top model. This means that our model weights based on the top BIC that were found in the training data and included in the Occam's window, may still contain selection bias and probably can be improved by using cross-validation or bootstrap in conjunction with the MC^3 approach. The summaries in Table 4.7 were computed for the size of Occam's window set to 25 units and similar ratios of over-performance (not reported here) were observed when using windows of smaller sizes (from 5 to 20). Therefore, increasing the size of Occam's window does not remove the effect of the over-fitting due to model selection.

4.4.3 Summary

Our simulation has confirmed earlier work that indicates an improvement in prediction error from BMA. The improvement is sample size dependent and is smaller for larger sample sizes. We did not find an effect due to correlation between the design variables on the BMA performance. The fact that there appear to be substantial overfitting in BMA due to model selection suggests that an improved procedure may be constructed by making the model selection process to be less driven by the training data. This can be accomplished by using resampling or subsampling of training data when navigating through model space and computing associated BICs and model weights.

Chapter 5
Bayesian Inference for Complex Computer Models

One of the big success stories of Bayesian inference is inference in large complex and highly structured models. A typical example is inference for computer models. Scientists use complex computer models to study the behavior of complex physical processes such as weather forecasting, disease dynamics, hydrology, traffic models, etc. Inference involves three related models, the true system, the complex simulation model and possibly a computationally more efficient emulation model. Appropriate propagation of uncertainties, good choice of emulation models, and calibration of parameters for the emulation model pose challenging inference problems reviewed in this chapter.

5.1 A Methodological Review of Computer Models

M.J. Bayarri

Statistical analyses combining computer models and field data pose new and important challenges for statisticians. Jim Berger has made important contributions to address many of these challenges, and has provided the field with very valuable insights. Space is too limited here to give an adequate summary of all his contributions in the area, so the focus will be narrow and very methodological, highlighting some few key methodological contributions that, even if covered in other papers, can go unnoticed by the complexity of the particular analysis at hand. Other methodological contributions are briefly mentioned. No effort is made to summarize other aspects of the analyses, such as important issues arising in MCMC implementations. References to very relevant work by other authors are avoided, not because they are not important or less influential, but because of the unique emphasis of this review (summarizing some of Jim Berger's contributions to the world of computer modeling) and the severe lack of space. The work discussed is mostly Bayesian, which is a particularly relevant methodology in this area because of the need to handle and

M.-H. Chen et al. (eds.), *Frontiers of Statistical Decision Making and Bayesian Analysis*, DOI 10.1007/978-1-4419-6944-6_5,

combine the disparate sources of uncertainties present, and to deal with confounding. Likelihood-based simplifications are often required, however. When possible, objective priors (proper and improper) are used, but the special characteristics of the problem might require extensive use of (quite) informative priors for some parameters.

5.1.1 Computer Models and Emulators

Numerical (computer) solvers of complex mathematical/physical models, known as *simulators* or *computer models*, are ubiquitous in all areas of application, science and everyday life; they are created to 'simulate' real processes. They have been used for multiple purposes (sensitivity analysis, optimization, understanding of the underlying physical processes, etc.), but their main role is increasingly becoming that of prediction at untried situations (extrapolation), when physical data are lacking or very scarce.

Analysis of computer model experiments involve both runs of the computer model at various inputs and field data. Most common computer models are *deterministic*, that is, if the computer model is run more than once at the same set of inputs, it produces the same output, and we will assume so in this section. Typical computer models have two types of inputs, denoted by $\mathbf{x}$ and $\mathbf{u}$. Inputs $\mathbf{x}$ occur in both the computer model and the field runs, whereas $\mathbf{u}$ are calibration/tuning parameters that are only needed to run the computer model. The inputs $\mathbf{x}$ are usually controllable by the experimenter, and are known for the field data (but see Section 5.1.5).

The outputs of computer models are usually extremely complex, but in application, the main interest is typically in a much simpler function, $y^M(\mathbf{x},\mathbf{u})$, which we assume (for the moment) to be scalar, and refer to it as 'the output' of the simulator.

Computer models never reproduce reality perfectly. This is explicitly modeled by introducing a discrepancy (or bias) function between the real process $y^R(\mathbf{x})$ and the model, so that

$$y^R(\mathbf{x}) = y^M(\mathbf{x},\mathbf{u}_*) + b_{\mathbf{u}_*}(\mathbf{x}), \tag{5.1.1}$$

where $\mathbf{u}_*$ is the (unknown) true value of a calibration parameter (or the 'best fit' of a tuning parameter), and $b_{\mathbf{u}_*}(\mathbf{x})$ is the discrepancy between the best model prediction and reality at input $\mathbf{x}$. Field data are assumed to be noisy observations of the real process, so that

$$y_j^F(\mathbf{x}) = y^R(\mathbf{x}) + \varepsilon_{xj}, \tag{5.1.2}$$

where $y_j^F(\mathbf{x})$ denotes the jth field replication at input $\mathbf{x}$, and ε_{xj} are measurement errors, $\varepsilon_{xj} \sim \mathrm{N}(0, 1/\lambda_F)$.

For fast simulators, that is, when $y^M(\mathbf{x},\mathbf{u})$ is readily available as often as required, Bayesian analysis proceeds by assessing priors for all the unknowns and obtaining simulations (via MCMC) from their posterior distributions given field data $\mathbf{y}^F$, the vector of all n_j replications $y_j^F(\mathbf{x})$ at inputs $\mathbf{x} \in D^F$. Bayesian analysis can thus si-

multaneously assess the accuracy of the computer model (learning about $b_{\mathbf{u}_*}(\mathbf{x})$) and provide predictions of the real process $y^R(\mathbf{x})$ incorporating all sources of uncertainty. Calibration (learning about $\mathbf{u}_*$) is a byproduct of this analysis (but see Section 5.1.3).

Many computer models however are very slow, some requiring hours or days to produce a single run $y^M(\mathbf{x},\mathbf{u})$. Analysis then proceeds by developing a fast surrogate — an *emulator* — to the expensive-to-run computer code. Gaussian process response-surface methodology (GASP) has been consistently effective for constructing such emulators. The idea is to recognize the computer $y^M(\cdot)$ as an effectively unknown function (known only at some few inputs) and assign it a Gaussian stochastic process prior distribution,

$$y^M(\cdot) \sim \mathrm{GP}\left(\mu_M, \frac{1}{\lambda_M}\, c_M(\cdot,\cdot)\right), \tag{5.1.3}$$

where μ_M, λ_M and $c_M(\cdot,\cdot)$ are the mean, precision, and correlation function that characterize the Gaussian process. The most commonly used correlation function for emulators is the exponential correlation function

$$c_M\left(y^M(\mathbf{x},\mathbf{u}), y^M(\tilde{\mathbf{x}},\tilde{\mathbf{u}})\right) = \exp\{-\sum_k \beta_k \mid x_k - \tilde{x}_k \mid^{\alpha_k}\} \times \exp\{-\sum_l \beta_l^* \mid u_l - \tilde{u}_l \mid^{\alpha_l^*}\}. \tag{5.1.4}$$

The separability of this correlation function (each factors is itself a correlation function in one-dimension) greatly simplifies computation and allows for scalability to high dimensional inputs while still producing a sensible approximation to the simulator.

The computer model is run at some designed set of inputs D^M yielding the computer model data $\mathbf{y}^M$, that is, the vector of runs $y^M(\mathbf{x},\mathbf{u})$ for $(\mathbf{x},\mathbf{u}) \in D^M$. At any (untried) input $(\mathbf{x},\mathbf{u}) \notin D^M$, the emulator predicts the corresponding output using the posterior predictive distribution of $y^M(\cdot)$, $\pi(y^M(\mathbf{x},\mathbf{u}) \mid \text{data})$, where 'data' refers to both field data $\mathbf{y}^F$ and computer model runs $\mathbf{y}^M$. The posterior predictive mean of $y^M(\cdot)$ passes through the simulator runs $\mathbf{y}^M$, and interpolates in between, a convenient property for emulators.

5.1.2 The Discrepancy (Bias) Function

There are several distinctive issues in the statistical analysis of computer model data. Perhaps the most important is the need to explicitly consider and adequately model the discrepancy term, $b_{\mathbf{u}_*}(\mathbf{x})$, when relating computer model to reality in (5.1.1). This term is *crucial* to avoid overtuning of $\mathbf{u}$, to appropriately incorporate uncertainty about computer model output, to assess model adequacy, and to allow improved use of simulator-based predictions (see Bayarri et al., 2007b). From now on, we drop the implicit dependence of b on $\mathbf{u}_*$ from the notation; we will also denote $\mathbf{u}_*$ simply by $\mathbf{u}$.

A somewhat common approach to calibration of computer models (estimation of $\mathbf{u}$) is to model the bias function (discrepancy, offset, sometimes called 'error in the model') as additive random error. In such analyses, it is assumed that the real process is related to the computer model by

$$y^R(\mathbf{x}) = y^M(\mathbf{x}, \mathbf{u}) + b_x, \tag{5.1.5}$$

where b_x is (non-observable) random noise; field data is still modeled as given by (5.1.2). This approach is generally *not* appropriate. The reasons why a necessarily simplified computer model does not match reality are complex; the assumption that random variation can adequately model this mismatch is unrealistic. It is useful to study some of these issues through a simple pedagogical example.

Toy Example. This example was produced by engineers to help them understand the 'nuts and bolts' of our methodology as applied to complex, real examples (see Bayarri et al., 2007b; and Liu, Bayarri, and Berger, 2009 for details of the analysis). It builds on the kinetics of chemical reaction $S_iH_4 \rightarrow S_i + 2H_2$. Let $y(t)$ denote concentration of S_iH_4 at time t. The entertained model is:

$$\frac{dy(t)}{dt} = -uy(t), \quad y(0) = y_0,$$

where y_0 is initial concentration, and u is an unknown rate (an unknown calibration parameter). Therefore, at inputs (u,t) the computer model output is

$$y^M(t,u) = y_0 \exp(-ut). \tag{5.1.6}$$

Suppose further that, in reality, a residual concentration c is left unreacted, so the *real* kinetics is

$$y^R(t,u) = (y_0 - c)\exp(-ut) + c.$$

Assume that $y_0 = 5.0$ is known to the experimenter, but not $y^R(\cdot)$, and hence neither u nor c, taken to be $u = 1.7$ and $c = 1.5$. Data are simulated at a grid of t-values, adding i.i.d. mean zero Gaussian error to the real process y^R at these values. Note that, in this example, $b(t) = 1.5\{1 - \exp(-1.7t)\}$.

Simulated field data and the 'best fit' computer model (that is, $y^M(t,\hat{u})$, for the MSE or MLE estimate $\hat{u} = 0.63$) are displayed in Figure 5.1, First notice that ignoring the bias term entirely (like using $\hat{u}$ to estimate u) produces a significant 'overtuning' of the calibration parameter u: the fit tries to 'make up' for the model inadequacy by over-shifting u to compensate, resulting in $\hat{u} = 0.63$, quite far from the 'true' value $u = 1.7$.

A similarly severe overtuning occurs when assuming that the discrepancy between reality and the computer model can be represented by random error, which is clearly revealed by a formal (Bayesian) analysis. The analysis is based on use of a uniform prior for u, over the known possible range $(0,3)$, and exponential priors for the precisions of the model error b_x and the field error; the priors are centered at moderate multiples of their respective MLEs (specifically at 5 times the MLEs). For

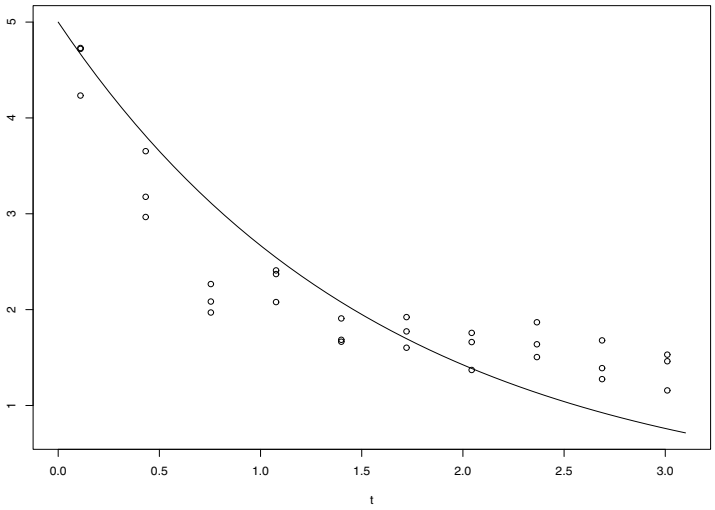


FIGURE 5.1. Maximum likelihood fit of model (5.1.6), and data for the toy example.

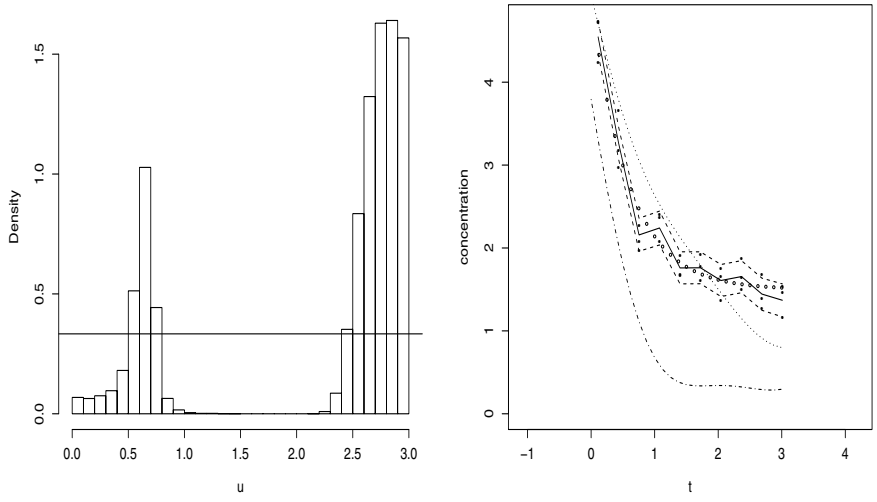


FIGURE 5.2. Toy example with the bias modeled as random error. Left Panel: Posterior distribution of u. Right Panel: The solid line is the mean prediction of reality, the dashed lines are 90% confidence bands. The dash-dotted and dotted lines correspond, respectively, to the strictly model prediction (no bias) associated to the posterior mean and to the least-squares estimate of u. The white circles are the true values of the real process, whereas the black ones represent the observed field data.

further details and discussion of these priors see Bayarri et al. (2005). The analysis yields entirely unreliable values of the calibration parameter u, and then compensates by producing very wide confidence bands for $y^R(.)$. This can clearly be seen in Figure 5.2; the left panel displays the posterior distribution of u, which curiously gives essentially no mass near the true value $u = 1.7$. The right panel shows the prediction of $y^R(\cdot)$ along with 90% confidence bands; not only does the predictive mean have a curiously rough shape, but the confidence bands are forced to be very wide to accommodate the discrepancy between the model and the field data.

An alternative standard statistical strategy for dealing with this situation would be to consider the residuals from model (5.1.5), to try to learn about the discrepancy between model and data. This has the difficulty that the overfitting gives 'false' residuals, so that it would be problematical to believe any structure found in the residuals.

To see why the modeling in (5.1.1) is superior to simply assuming that the bias is additive noise, let us return to the toy example, using GASP modeling for $b(\cdot)$.

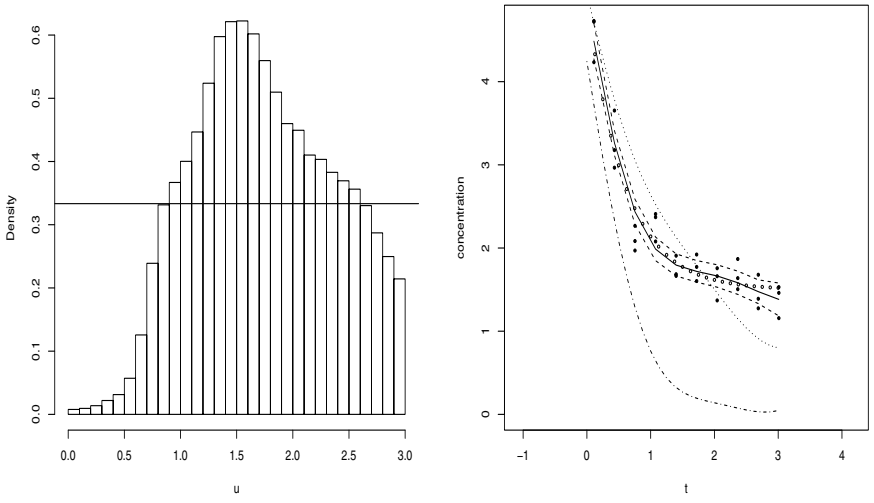


FIGURE 5.3. Toy example with GASP bias. Left Panel: Posterior distribution of u. Right Panel: The solid line is the mean prediction of reality, the dashed lines are 90% confidence bands. The dash-dotted and dotted lines correspond, respectively, to the strictly model prediction (no bias) associated to the posterior mean and to the least-squares estimate of u. The white circles are the true values of the real process, whereas the black ones represent the observed field data.

Toy Example (continued). We repeat the analysis of the Toy Example data but now using the modeling in (5.1.1), with a GASP prior for the bias with mean $\mu_b = 0$, precision λ_b and correlation function $c_b(b(t), b(\tilde{t})) = \exp\{-\beta_b(t-\tilde{t})^{1.9}\}$. The prior for

u and for the field and bias precisions are identical to the ones in the previous analysis; β_b is fixed at its MLE (for details, see Bayarri et al., 2005 and 2007b). Results are shown in Figure 5.3. In contrast to the previous analysis, the non-parametric modeling of the bias term dramatically improves both calibration and prediction of reality; the posterior distribution of u is much more sensible, and the confidence bands for the prediction of reality are much tighter (and still include reality).

Notice, however, that one does not really learn the value of the calibration parameter u; its posterior distribution is only modestly tighter than its prior distribution. This is not a flaw; there simply is not sufficient information available to learn precisely about u (see also Section 5.1.3). The point is that the analysis does not rule out the correct value of u, as happened with the overtuning when the bias was incorrectly modeled as additive noise, and instead allows the bias to adjust to the discrepancy between model and data. Avoiding the overtuning allows the prediction to be done more accurately, with tighter confidence bands.

Modeling the discrepancy *outside* the computer model, as in (5.1.1) (or using a multiplicative bias version, as in Bayarri et al. (2007a), is most common when the model is very expensive to run and is treated as a 'black box.' For simple deterministic models, in which access to the equations defining the model is possible and their solution is not too complex, the bias term can be taken inside the model itself so that it can be evolved according to the dynamics that govern the computer model.

5.1.3 Confounding of Tuning and Bias

Because of the nature of computer models, it is usually best to allow $b(\cdot)$ to be modeled noparametrically, as opposed to assuming that it has a specific parametric form or (worst) simply treating it as random error. This causes a major difficulty, however: there is then a severe confounding between $\mathbf{u}$ and the bias function; they are not identifiable. To see this, suppose that one observes a huge amount of field data so that reality $y^R(\mathbf{x})$ becomes effectively known. Suppose also that we can view $y^M(\mathbf{x}, \mathbf{u})$ as a completely known function. Then — even in this most favorable situation — for each $\mathbf{u}$, there is a $b(\mathbf{x}) = y^R(\mathbf{x}) - y^M(\mathbf{x}, \mathbf{u})$ that gives back reality $y^R(\mathbf{x})$; they cannot separately be identified. Note that this happens because the support of the prior for $b(\mathbf{x})$ is *any* (appropriately smooth) function but, again, this flexibility is usually required for appropriate modeling.

This non-identifiability between $\mathbf{u}$ and the bias was not recognized in the original literature in this area, so its profound implications for the analysis of computer models has only been highlighted in recent years. Bayesian analysis is still tenable, of course, but the following considerations are a result of this understanding:

- A fairly informative prior on $\mathbf{u}$ is highly desirable, and is often available when $\mathbf{u}$ has physical meaning (or physical limitations). In contrast, it is highly unlikely to have enough information about $b(\cdot)$ to produce a tight prior; common strategies to 'tighten' the prior on $b(\cdot)$ are to 'encourage' the bias to be 0 (by taking the

GASP prior to be $\mu_b = 0$) and to make it smooth (by taking all components of α_b to be 2 or 1.9 in the GASP correlation function).

- Because of the confounding, separate inference on $\mathbf{u}$ and $b(\mathbf{x})$ is usually very sensitive to the priors. In fact, since the bias function is a priori quite uncertain, one can not typically learn much about $\mathbf{u}$ by only comparing computer model and field data. Careful Bayesian analysis clearly reveals this issue; small changes in the highly uncertain prior for $b(\cdot)$ usually have an appreciable impact on the posterior distribution of $\mathbf{u}$. If calibration (learning about $\mathbf{u}$) is a primary goal, it is crucial to incorporate good prior information about $\mathbf{u}$.
- Sometimes it is of specific interest to modelers to learn about the bias function and how it behaves in different parts of the input space. For this purpose, a good prior on $\mathbf{u}$ would be extremely useful. In any case, modelers could be presented with the marginal posterior of $b(\cdot)$ along with conditional posteriors for the bias functions corresponding to specific values $\mathbf{u}$ of interest. Useful information as to the nature of the bias and its inherent uncertainty can still be obtained from the analysis.
- Perhaps surprisingly, prediction (and attached uncertainties) remain very stable over different priors specifications; $\mathbf{u}$ and the bias compensate for each other in a fashion that leaves prediction quite stable. For the typical real goal, which is prediction beyond the range of the data, it is crucial to thus use this joint posterior distribution of $\mathbf{u}$ and the bias (as in Bayarri et al., 2007a).

Some of the consequences of the confounding between bias and calibration parameters seem unsettling, in particular the difficulty of learning about the calibration parameters. It should be kept in mind, however, that this is just the reality of the situation; avoiding the confounding by overly restrictive modeling of the bias function (the most restrictive being ignoring it or treating it as random error) is likely to produce severe over-tuning of $\mathbf{u}$ rendering the results even less useful. Note, also, that it is unclear how one would approach this crucial issue of confounding from a non-Bayesian perspective.

5.1.4 Modularization

Modularization refers to partially keeping 'good modules' (or components) of an overall model separate from uncertain ones in Bayesian learning. Modularization is discussed at length in Liu, Bayarri, and Berger (2009), where it is methodologically motivated as a reasonable 'quick and dirty' method for preventing a misspecified module from "contaminating" correctly specified modules. The motivation is to improve modeling in very complex situations (as in computer models) in which use of more standard statistical model validation strategies are not possible. Note that not any modularization is deemed reasonable; indeed, it is argued in Liu, Bayarri, and Berger (2009), that modularization ideally should be justified from a modeling perspective, and not simply be viewed as a trick to improve computation (e.g., mixing in MCMC).

The idea of modularization is sketched next in a simple example, followed by indicating how and why it can usefully be applied to analysis of computer models. For further details, a thorough discussion of the method, and references to related ideas, see Liu, Bayarri, and Berger (2009).

Random Effects Example. Consider a simple random effects model in which we have n independent observations on each of N groups:

$$\begin{aligned} y_{ij} \mid b_i &= b_i + \varepsilon_{ij}, \qquad j = 1, \ldots n; \quad i = 1 \ldots N, \\ \varepsilon_{ij} \mid \sigma_i^2 &\sim N(0, \sigma_i^2), \qquad b_i \mid \tau^2 \sim N(0, \tau^2). \end{aligned} \tag{5.1.7}$$

Assume here that our 'suspect' module is the distribution of the random effects, about which we are not very confident, while the 'good' module is the distribution of the observables y_{ij}. We also assume that the number of groups N is very large, while the number of replications n is very small in comparison, but large enough for reasonable estimation of the σ_i^2. This is an extremely simplified version of a situation encountered in Bayarri et al. (2007a), in which functional data (both computer model and field data) were represented in terms of a large number of basis elements, with the modeling (5.1.1) and (5.1.2) applied to each coefficient, and the coefficients of the bias functions were modeled hierarchically.

Let $\sigma^2, \bar{\mathbf{y}}, \mathbf{s}^2$ denote N-vectors with components respectively $\sigma_i^2, \bar{y}_i$ and $s_i^2 = \sum_{j=1}^n (y_{ij} - \bar{y}_i)^2/n$. With objective priors $\pi(\sigma_i^2) \propto (\sigma_i^2)^{-1}$ and $\pi(\tau^2 \mid \sigma^2) \propto (\tau^2 + \bar{\sigma}^2/n)^{-1}$, the marginal posterior for τ^2 and σ^2 is

$$\begin{aligned} \pi(\tau^2, \sigma^2 \mid \bar{\mathbf{y}}, \mathbf{s}^2) \propto \frac{1}{\tau^2 + \bar{\sigma}^2/n} \prod_{i=1}^N \Big[(\sigma_i^2)^{-\frac{n+1}{2}} \exp\Big(- \frac{n s_i^2}{2\sigma_i^2} \Big) \\ \times \frac{1}{(\tau^2 + \sigma_i^2/n)^{1/2}} \exp\Big\{ - \frac{\bar{y}_i^2}{2(\tau^2 + \sigma_i^2/n)} \Big\} \Big]. \end{aligned} \tag{5.1.8}$$

Suppose now that one of the random effects, say b_k, happens to be very large while the others are small. (In this simple setting, this would suggest a violation of the normality assumption for the b_i, and point to the need for modeling with, say, heavier tails but, in the world of complex computer models, neither the cause nor the improved modeling are usually clear.) One might imagine that the unusually large b_k would cause τ^2 to be large, resulting in little shrinkage; then, while the hierarchical analysis would be 'wasted,' no real harm would result. However, a careful inspection of the posterior (5.1.8) reveals that (because N is large) the Bayesian analysis will, instead, force σ_k^2/n to be large while keeping τ^2 small. The consequence of having σ_k^2/n large and τ^2 small will result in a dramatic — and very incorrect — shrinkage of b_k from the large $\bar{y}_k$ towards 0 (see Liu, Bayarri, and Berger, 2009).

The modular approach in this example is simply to insist that the posterior for σ^2 be based only on the replicate observations, which contain almost all of the real information about σ_k^2 (and which would say that σ_k^2 is not large), effectively replacing $\pi(\sigma^2 \mid \bar{\mathbf{y}}, \mathbf{s}^2)$ by $\pi(\sigma^2 \mid \mathbf{s}^2)$. Note that the conditional posterior distributions for τ^2 and $\mathbf{b}$ are unchanged in terms of their mathematical expressions, but will

change very considerably in terms of their location: with σ_k^2 no longer being able to accommodate the outlier, τ^2 will become large, and the posterior for b_k will remain near the large $\bar{y}_k$ (and not 0).

In this simple example, it is easy to point to the culprit of the deceiving analysis: a normal distribution for the random effects is not good when a very big b_k is possible. The solution is equally simple: use a heavy tail distribution instead. The appeal of modularization is that it can be much easier to identify a useful restriction of the Bayesian analysis, than to develop a better model for suspect modules. In the previous example, deciding that the σ_i^2 will be determined only from the replications is much simpler than attempting to infer a good model for the random effects (which are not directly observed). In more complicated scenarios, as with complex computer models, this difference in difficulty is much more pronounced, and (smart) modularization can be the only feasible way to protect the information from 'good' modules from being unduly altered by uncertain modules.

Computer models have three modules: the computer model itself, the field data, and the bias or discrepancy term. Modularization has been explored and implemented for each of them (for details see Liu, Bayarri, and Berger, 2009):

- In the computer model. A most useful modularization consists in using only computer model runs data (and no field data) in learning about the correlation parameters in the GASP approximation. This is a very intuitive simplification, since usually the computer model runs have been carefully designed and there are enough of them to approximate the output of interest to a reasonable accuracy. This modularization is indeed explicitly or implicitly done by many researchers in practice. In both our previous analyses of the Toy Example data, the computer model was assumed to be unknown with a GASP prior as in (5.1.3) and (5.1.4), with the usual objective priors for μ_M, and λ_M, exponential priors for the β's centered at 10 times their respective MLE's, and the α's been fixed at their MLE's. Whenever the analysis required posterior draws from these parameters, they were generated from their posterior distribution given only the computer model runs.
- In the field data. When there are enough replications, it might be advisable to learn about field error based only on field residuals. This is the situation in the previous random effects model, and has been shown to prevent severe underestimation of important components of bias (see Bayarri et al., 2007a).
- In the bias function. In challenging situations with over parameterization and random inputs, Liu et al. (2008) proposes learning about correlation parameters of the bias function based on discrepancies. Several possible modularization schemes are entertained and critically compared.

Appropriate modularization in the computer models world is generally beneficial; indeed, the unavoidable confounding and uncertainties present increase the danger of a poor module wrongly 'contaminating' a good one. Modularization of influential hyperparameters (such as the mean and precision of the GASP for the bias function) is generally not advisable, however.

5.1.5 Additional Issues

Statisticians who have not been involved with computer models are often curious as to why computer model validation/checking can not be dealt with in the same way as one deals with standard statistical model validation/checking. This is an interesting question and is explicitly addressed in Bayarri et al. (2007a, 2007b).

Although we have taken the output of interest from the computer model $y^M(\mathbf{x}, \mathbf{u})$ to be a scalar in this review, it is usually a function, say of time t. If the function is smooth, one can recover it from relatively few t-values, so that t can then simply be considered another input in the analysis; simplifying assumptions on the correlation function are usually needed to efficiently handle the resulting correlation matrices (Bayarri et al., 2009b; Liu et al., 2008).

If the output function is not smooth, a common approach is to represent the unknown model and bias functions in terms of common basis elements and apply the methodology to each of the (unknown) coefficients; a hierarchical structure for the coefficients of the bias is often required (see Bayarri et al., 2007a and references there).

Often, the inputs contain both categorical and numerical components, so an overall GASP analysis is not appropriate. A possibility is to have a GASP over the numerical components for each value of the categorical inputs which are then related hierarchically (as in Bayarri et al., 2009b).

Sometimes the inputs $\mathbf{x}$ which enter both field data and computer model are not known precisely in the field: maybe only a 'nominal' value (that is, subject to error) is known (Bayarri et al., 2007a), or they vary from replication to replication (Liu et al., 2008), or maybe the field data are themselves random occurrences, at random inputs $\mathbf{x}$, and these distributions have to be ascertained based on relevant experiments (Bayarri et al., 2009a).

Computer models are increasingly being used for extrapolation past the range of the data; see Bayarri et al. (2007a) for such an analysis and the delicate methodological issues involved. In particular, they are often used to help assess the risk of catastrophic events; in Bayarri et al. (2009a) a combination of statistical modeling for extremes, utilization of data related to the geophysical process and computer model simulations allows for such an extrapolation.

In this section we have concentrated on *deterministic* computer models. There are large classes of *stochastic* computer models. One such class adds randomness to the differential equations defining the model; we do not comment on those here. Network models are a different class of stochastic models in which individual *agents* move (and interact) in a pre-defined network according to probability distributions and deterministic rules. In Molina, Bayarri, and Berger (2005) and Bayarri et al. (2004), one such simulator was used to model Chicago downtown traffic in rush hours. These models pose new methodological issues. First, ignoring uncertainty in the inputs not only results in severe underestimation of the prediction errors of measures of congestion, but can actually seriously bias the predictions themselves (Bayarri et al., 2004). Bayesian analysis takes care of and combines uncertainties in stochastic networks, but it is not easy. Implementation typically requires the use of

probabilistic networks to partially reproduce the movement of agents in the network (although not their interaction). Also, direct data on the system often consists of observing some measures of congestions in some of the links. This results in highly constrained parameter spaces, which are extremely difficult to handle; a solution is proposed in Molina, Bayarri, and Berger (2005). The technique should be applicable to many other discrete networks, such as telecommunications networks and certain agent-based models.

5.1.6 Summary

This very brief review highlights some key methodological contributions of Jim Berger to the analysis of computer model data; these often go unnoticed, hidden in part by the complexity of the applications themselves. They are however crucial in choosing and interpreting the appropriate analysis for computer models. Particular emphasis has been placed on (i) demonstrating the *requirement* of including a flexible discrepancy term in the analysis; (ii) highlighting the resulting unavoidable non-identifiability issue and its consequences; and (iii) discussing a partial solution to the frequent problem that one has varying confidence in different components of the overall analysis model and wants the analysis to be robust to ill-understood components. These issues arise in virtually all computer model analyses. Other important methodological issues were also briefly mentioned. Key references to previous work on these issues can be found in the references.

Acknowledgments: The invaluable help of Danilo Lopes in producing the Figures for this section is gratefully acknowledged. This research was supported in part by the National Science Foundation (GRANTS DMS-0757527) and by the Spanish Ministry of Education and Science (Grant MTM2007-61554). A large part of this research was conducted under the auspices and support of the Statistical and Applied Mathematical Sciences Institute (SAMSI) 2006-7 research program on the Development, Assessment and Utilization of Complex Computer Models. The author is indebted to Jim Berger for the many stimulating conversations, and for the invaluable insights into this fascinating but difficult world of computer models.

5.2 Computer Model Calibration with Multivariate Spatial Output

K. Sham Bhat, Murali Haran, and Marlos Goes

Complex computer models are widely used by scientists to understand and predict the behavior of complex physical processes. Examples of applications include climate science, weather forecasting, disease dynamics, and hydrology. Inference on

these complex systems often combines information from simulations of the complex computer model with field data collected from experiments or observations on the real physical system. The computer model simulations are frequently very computationally expensive, with each simulation taking minutes, days, or even weeks to complete, which makes Monte Carlo-based approaches to inference infeasible. Computer model emulation is a powerful approach pioneered by Sacks et al. (1989) to approximate the expensive computer model by a Gaussian process. Emulation allows approximate output at any parameter setting to be obtained from a computationally tractable Gaussian process fit to the output from the computer model at several parameter settings. This approach can then be used in a larger framework that includes a model for physical observations in order to do computer model calibration. Computer model calibration finds the value of the computer model parameters or 'inputs' most compatible with the observations of the process. Here we follow the general framework described in Kennedy and O'Hagan (2001) and further developed by many others (cf. Bayarri et al., 2007a; Sansó, Forest, and Zantedeschi, 2008).

Increasingly, computer model output is multivariate (cf. Bayarri et al., 2007a; Higdon et al., 2008). Of particular interest are models where the output is in the form of multivariate spatial data. We consider as a case study the problem of inferring the value of a climate parameter based on climate model output and physical observations that are in the form of multivariate spatial data sets. This problem is motivated by the goal of assessing the risks of future climate change. Specifically, we focus on the problem of learning about the climate parameter 'background ocean vertical diffusivity' (K_v), which determines the strength of the heat and salt diffusion in the ocean component of the climate model, and is a key parameter in climate model predictions of the Atlantic Meridional Overturning Circulation (AMOC). The AMOC, part of the global ocean circulation system, plays an important role in global climate. A weakening or possible collapse of the AMOC can potentially result in major temperature and precipitation changes and a shift in terrestrial ecosystems. AMOC predictions may be obtained from climate models, which include several parameterizations in order to mimic real physical processes. Of the model parameters, K_v is particularly important for predictions of AMOC. Reducing the uncertainty about the value of K_v will also reduce the uncertainty of other key model parameters like climate sensitivity (Forest et al., 2002). While the value of the parameter K_v may not resemble the observed ocean diffusivity, because it is intended to represent several mechanisms that generate turbulent mixing in the ocean, ocean tracers can provide information about large scale ocean patterns. Such information can be used to infer K_v in the model, since observed tracers are strongly affected by this parameter. For example, larger observed values of the tracer Δ^{14}C in the deep ocean suggest a higher intensity of vertical mixing. These tracer data are in the form of spatial fields. Hence, the computer model calibration problem here involves climate parameter inference based on multivariate spatial data. In this section, we consider inference based on three oceanic tracers, all in the form of relatively small one-dimensional spatial fields. We present a simple framework for combining information from multiple spatial fields from model simulations and physical observations in the context

of inferring the climate parameter K_v. We study the impact of including a Gaussian process model for the discrepancy between the model and the true system. In addition, we study the impact of model assumptions by holding out model output at a particular parameter setting and treating noisy versions of this output as 'real data.' We consider two statistical models, one that combines observation error and model discrepancy into a single independent error term, the second where model discrepancy is modeled separately using a Gaussian process. We are particularly interested in studying the impact of the discrepancy term on climate parameter inference. We also then examine the effect of estimating emulator spatial variance in a Bayesian framework versus using a plug-in approach.

The rest of this section is organized as follows. In Section 5.2.1, we discuss our approach for calibration with spatial output. We build upon this framework to perform parameter inference with multiple spatial fields in Section 5.2.2, paying special attention to model discrepancy and emulator variances. In Section 5.2.3, we describe our case study, discussing both the data set and modeling and implementation details. We describe the results of our study in Section 5.2.4 and conclude with a summary and discussion in Section 5.2.5.

5.2.1 Computer Model Calibration with Spatial Output

In this section, we describe our model for inferring calibration parameters from the observations and model output of a single spatial field. We use the two-stage approach described below for model calibration. We will also discuss the importance of various modeling assumptions.

In the first stage of our approach, we emulate the computer model by fitting a Gaussian process to the spatial computer model output. In the second stage, we connect the calibration parameters to the observations using the emulator, while allowing for other sources of uncertainty, such as model discrepancy and observation error. This allows us to use a Bayesian approach to obtain a posterior distribution for the parameters. Our approach of splitting inference into two stages has several advantages over fitting a single model in one inferential step including separating the parts of the statistical model that are known to be correct from the parts of the model that are questionable, improved diagnostics, and computational advantages (see Bayarri et al., 2007b; Liu, Bayarri, and Berger, 2009; Rougier, 2008a).

We begin with some notation. Let $Z(\mathbf{s})$ be the observation of the spatial field at location $\mathbf{s}$, where $\mathbf{s} \in D$ with $D \in \mathbb{R}^d$. For simplicity, and given the case study in Section 5.2.3, we assume that $d = 1$, i.e., we have a one-dimensional spatial field. Let θ be the calibration or model parameter of interest; our framework may easily be expanded to allow for vectors of parameters. $Y(\mathbf{s}, \theta)$ denotes the computer model output at the location $\mathbf{s}$, and at the calibration parameter setting θ. In general, the spatial data from the computer model grid may or may not coincide with the locations of the observations. The objective here is to infer a posterior distribution of θ given the observed data and computer model output.

Let $\mathbf{Y} = (Y_{11}, \cdots, Y_{n1}, Y_{12}, \cdots, Y_{n2}, \cdots, Y_{1p}, \cdots, Y_{np})'$, obtained by stacking computer model output at all calibration parameter settings, denote the computer model output for a single spatial field. Y_{ik} corresponds to the model output for location $\mathbf{s}_i$ and calibration parameter setting θ_k, n is the number of model output locations, and p is the number of calibration parameter settings. Similarly, $\mathbf{Z} = (Z_1, \cdots, Z_N)'$ are the observations for the spatial field, where N is the total number of observations.

5.2.1.1 Computer Model Emulation

We model the computer model output $\mathbf{Y}$ using a Gaussian process:

$$\mathbf{Y} \mid \beta, \theta, \xi_m \sim N(\mu_\beta(\theta), \Sigma_M(\xi_m)),$$

where we assume a linear mean function, $\mu_\beta(\theta) = X\beta$, with X a covariate matrix of dimension $np \times b$, where there are $(b-1)$ covariates. The covariates we use are location and the calibration parameter. ξ_m is a vector of covariance parameters that specify the covariance matrix $\Sigma_M(\xi_m)$ and β is a vector of regression coefficients. We use a Gaussian covariance function as described below:

$$(\Sigma_M)_{ij}(\phi, \kappa) = \zeta I(i = j) + \kappa \exp\left(-\frac{\|\mathbf{s}_i - \mathbf{s}_i\|^2}{\phi_s^2} - \frac{|\theta_i - \theta_j|^2}{\phi_c^2}\right), \tag{5.2.1}$$

where $\phi = (\phi_s, \phi_c)$, $\kappa, \zeta, \phi_s, \phi_c > 0$. The covariance function is separable over space and calibration parameters, although a nonseparable covariance could be chosen if appropriate (see Gneiting, 2002). Note that this function can be easily adapted to models for multiple calibration parameters, as well as multiple spatial dimensions.

Let the maximum likelihood estimate of (ξ_m, β) be $(\hat{\xi}_m, \hat{\beta})$. Let $\mathbf{S}$ be the set of locations where the observations were collected. Following the standard kriging framework (Cressie, 1993; Stein, 1999), the multinormal predictive distribution for the computer model output at a new θ at $\mathbf{S}$ is obtained by substituting $(\hat{\xi}_m, \hat{\beta})$ in place of (ξ_m, β) and conditioning on $\mathbf{Y}$. We denote the random variable with this predictive distribution by $\eta(\mathbf{Y}, \theta)$ in the second stage of our inference below.

5.2.1.2 Computer Model Parameter Inference

In order to infer θ based on the observations $\mathbf{Z}$, we need a probability model connecting θ and $\mathbf{Z}$. The predictive distribution from Section 5.2.1.1 provides a model for computer model output at any θ and any set of new locations. We now model the observations $\mathbf{Z}$ as realizations from a stochastic process obtained by accounting for additional error to the computer model emulator from Section 5.2.1.1. Our model for the observations $\mathbf{Z}$ is therefore

$$\mathbf{Z} = \eta(\mathbf{Y}, \theta) + \delta(\mathbf{S}) + \varepsilon,$$

where $\eta(\mathbf{Y},\theta)$ is as described in Section 5.2.1.1, $\varepsilon \sim N(0,\psi I)$, and $\varepsilon = (\varepsilon_1,\cdots,\varepsilon_N)'$ is the observation error with $\psi > 0$ as the observation error variance. The model discrepancy, $\delta(\mathbf{S})$, is modeled as a zero-mean Gaussian process. Hence, $\delta(\mathbf{S}) \sim N(\mathbf{0},\Sigma_d(\xi_d))$, where ξ_d is a vector of covariance parameters that specify the covariance matrix $\Sigma_d(\xi_d)$. We have in essence 'inferred a likelihood' for use in our Bayesian framework, since for any fixed $\mathbf{Z}$, we can obtain a value of the likelihood for any value of θ. We will discuss the merits of including a model discrepancy term in Section 5.2.2.3.

We may allow the emulator spatial variance scale parameter from the first stage, κ, to vary, rather than plugging in the MLE $\hat{\kappa}$. We can now perform inference on θ, ψ, κ, and ξ_d by specifying a prior for these parameters. Using Markov chain Monte Carlo (MCMC), we can estimate a posterior distribution for θ. It should be noted that the computational complexity of the matrix operations involved in the second stage of our approach is solely dependent on N, the size of $\mathbf{Z}$, and not $M = np$, where M is the size of the ensemble of model output $\mathbf{Y}$. We will discuss prior selection for θ, ψ, κ, and ξ_d in Section 5.2.3.2.

5.2.2 Calibration with Multivariate Spatial Output

In this section, we discuss how our approach can be used to combine information from multiple spatial fields. We use a separable covariance model (see for instance, Banerjee, Carlin, and Gelfand, 2004) to model the relationship of the computer model output from the three spatial fields. The similar shape of the empirical variograms of the model output from three spatial fields in our case study in Section 5.2.3 justify the use of a separable covariance model. We extend our notation to allow for multiple spatial fields. Let $\mathbf{Y}_1 = (Y_{11}\cdots Y_{1np})'$, $\mathbf{Y}_2 = (Y_{21}\cdots Y_{2np})'$, and $\mathbf{Y}_3 = (Y_{31}\cdots Y_{3np})'$ denote the computer model output for three spatial fields. Similarly, $\mathbf{Z}_1 = (Z_{11}\cdots Z_{1N})'$, $\mathbf{Z}_2 = (Z_{21}\cdots Y_{2N})'$, and $\mathbf{Z}_3 = (Z_{31}\cdots Y_{3N})'$ are the observations for the same three spatial fields. For convenience, we write $\mathbf{Y} = (Y_{11},Y_{21},Y_{31},Y_{12}\cdots Y_{1np},Y_{2np},Y_{3np})'$, and $\mathbf{Z} = (\mathbf{Z}_1\ \mathbf{Z}_2\ \mathbf{Z}_3)$.

5.2.2.1 Stage 1: Emulation for Multivariate Data

The computer model output for the spatial fields are modeled using using a separable cross-covariance function as described below:

$$\begin{aligned}\mathbf{Y} \mid \beta,\theta,\xi_m &\sim N(\mu_\beta(\theta),\Sigma_M(\xi_m)),\\ \mu_\beta &= (\mu_{\beta_1},\mu_{\beta_2},\mu_{\beta_3})',\\ \Sigma_M(\xi_m) &= P(\zeta)+H(\phi)\otimes T(\kappa,\rho),\end{aligned}$$

where μ_{β_i} is a function of the calibration parameters, and β_1,β_2,β_3 are the coefficient vectors for $\mathbf{Y}_1,\mathbf{Y}_2,\mathbf{Y}_3$ respectively. We note that we are implicitly assuming

a linear relationship among the spatial fields. For an approach based on a flexible hierarchical model allowing for non-linear relationships among spatial fields, see Bhat et al. (2009). A computationally inexpensive approach that avoids computer model emulation and hence utilizes several simplifying assumptions is described in Goes et al. (2009).

We now assume $\mu_{\beta_i}(\theta) = X\beta_i$ (for i=1,2) where X is the covariate matrix of dimension $M \times b$, with covariates (depth and calibration parameters) as specified in Section 5.2.1.1. ξ_m is a vector of covariance parameters that specify the cross-covariance matrix $\Sigma_M(\xi_m)$. $H(\phi)$ explains spatial dependence, while $T(\kappa,\rho)$ is interpreted as the cross-covariance between spatial fields. $P(\zeta)$ is a matrix that describes microscale variance of the process. The covariance matrices are defined as follows:

$$H_{ij}(\phi) = \exp\Big(-\frac{\|\mathbf{s}_i - \mathbf{s}_i\|^2}{\phi_s} - \frac{|\theta_i - \theta_j|^2}{\phi_c}\Big),$$

$$T = \begin{bmatrix} \kappa_1 & \rho_{12}\sqrt{\kappa_1\kappa_2} & \rho_{13}\sqrt{\kappa_1\kappa_3} \\ \rho_{12}\sqrt{\kappa_1\kappa_2} & \kappa_2 & \rho_{23}\sqrt{\kappa_2\kappa_3} \\ \rho_{13}\sqrt{\kappa_1\kappa_3} & \rho_{23}\sqrt{\kappa_2\kappa_3} & \kappa_3 \end{bmatrix}, \quad P = \begin{bmatrix} \zeta_1\mathbf{I}_N & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \zeta_2\mathbf{I}_N & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \zeta_3\mathbf{I}_N \end{bmatrix},$$

where $\phi = (\phi_s, \phi_c)$, and $\kappa_i, \zeta_i, \phi_s, \phi_c > 0$, $-1 \le \rho_{ij} \le 1$. We reduce parameters and ensure that $\Sigma_{\mathbf{Y}}$ is positive definite and symmetric by letting $\rho_{ii} = 1$ and $\rho_{ij} = \rho_{ji}$. We estimate MLEs for the following parameters using the computer model output: $\mathbf{Y}$: $\zeta_1, \zeta_2, \zeta_3, \kappa_1, \kappa_2, \kappa_3, \phi_s, \phi_c$. In principle, $\beta_1, \beta_2, \beta_3$ and ρ may be estimated using maximum likelihood, but in the case study in Section 5.2.3, we estimate $\beta_1, \beta_2, \beta_3$ using least squares regression and ρ using empirical sample correlations. As in Section 5.2.1.1, we obtain a multinormal predictive distribution $\eta(\mathbf{Y}, \theta)$ each θ at $\mathbf{S}$ by plugging in the MLEs and conditioning on $\mathbf{Y}$. For ease of computation, we order the model output by depth and calibration parameter, and we write $\mathbf{Y} = (Y_{11}, Y_{21}, Y_{31}, Y_{12} \cdots Y_{1np}, Y_{2np}, Y_{3np})'$.

5.2.2.2 Stage 2: Inference for Multiple Spatial Fields

We write the model for the observed data as follows:

$$\mathbf{Z} = \eta(\mathbf{Y}, \theta) + \delta(\mathbf{S}) + \varepsilon,$$

where $\eta(\mathbf{Y}, \theta)$ is as described earlier in Section 5.2.2.1, and $\varepsilon = (\varepsilon_{11}, \cdots, \varepsilon_{N1}, \varepsilon_{12}, \cdots, \varepsilon_{N2}, \varepsilon_{13}, \cdots, \varepsilon_{N3})'$ is the observation error. We assume that $\varepsilon \sim N(\mathbf{0}, \Sigma_\varepsilon)$ with

$$\Sigma_\varepsilon = \begin{bmatrix} \psi_1\mathbf{I}_N & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \psi_2\mathbf{I}_N & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \psi_3\mathbf{I}_N \end{bmatrix},$$

where $\psi_1, \psi_2, \psi_3 > 0$ are the observation error variances for the three spatial fields. The model error or discrepancy $\delta(\mathbf{S})$ is modeled as a vector of three indepen-

dent zero-mean Gaussian processes. The model for discrepancy is given in Section 5.2.2.3, and includes covariance parameters ξ_{d1}, ξ_{d2}, and ξ_{d3}.

Using the observations $\mathbf{Z}$, we obtain the posterior distribution of θ, ψ_1, ψ_2, ψ_3, ξ_{d1}, ξ_{d2}, and ξ_{d3} using MCMC as discussed in Section 5.2.1.2. We may also allow the emulator spatial variance parameters from the first stage, κ_1, κ_2, and κ_3 to be reestimated, rather than using plug-in MLEs. This can be done using by estimating the matrix T using an inverse Wishart prior. More details about prior selection are discussed in Section 5.2.3.2.

5.2.2.3 Model Discrepancy

An important concern in the process of calibration is whether the model is an adequate representation of the true phenomena in the system. When this is not the case, there is a need to consider ways to incorporate the difference between the computer model and reality. The latter is usually referred to as model discrepancy.

A framework to account for model discrepancy is introduced in Kennedy and O'Hagan (2001), and strong arguments in favor of inclusion of a model discrepancy term in any calibration approach is made in Bayarri et al. (2007b). Specifically, the argument is made that neglecting to account for the model discrepancy results in overfitting, resulting in potentially biased and incorrect inference of the calibration parameters. A test is introduced for whether model discrepancy is needed in Bayarri et al. (2009b), which almost always results in rejecting the hypothesis that the model represents the truth. However, O'Hagan (2009) suggests that the inclusion of a model discrepancy does not always result in a less biased estimates of calibration parameters, rather more biased estimates are possible. A further difficulty in including a model discrepancy term in our statistical model is the high dependence between the calibration parameter and model discrepancy term (Liu, Bayarri, and Berger, 2009). Previous work has shown that attempting to separate observation error and model error can impose nontrivial computation and conceptual problems (Kennedy and O'Hagan, 2001; Sansó, Forest, and Zantedeschi, 2008). An approach that combines observation error and model error into a single term, rather than estimate them separately is described in Sansó, Forest, and Zantedeschi (2008). Even this approach requires substantial compromises in computing techniques in order to fit the model. In our case study, we consider the two different approaches to incorporate the error into our statistical model as follows:

Approach 1: No model discrepancy term (model discrepancy and observation error combined).

$$\mathbf{Z} = \eta(\mathbf{Y}, \theta) + \varepsilon,$$

where $\eta(\mathbf{Y}, \theta)$ is the predictive distribution as described earlier in Section 5.2.2.1, and $\varepsilon = (\varepsilon_{11}, \cdots, \varepsilon_{N1}, \varepsilon_{12}, \cdots, \varepsilon_{N2}, \varepsilon_{13}, \cdots, \varepsilon_{N3})'$ is the observation error, and $\varepsilon \sim N(\mathbf{0}, \Sigma_{\mathcal{E}})$, with

$$\Sigma_{\mathcal{E}} = \begin{bmatrix} \psi_1 \mathbf{I}_N & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \psi_2 \mathbf{I}_N & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \psi_3 \mathbf{I}_N \end{bmatrix},$$

where $\psi_1, \psi_2, \psi_3 > 0$ are the observation error variances for the three spatial fields. In the univariate case, $\varepsilon \sim N(0, \psi I)$, where $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_N)'$ is the observation error with $\psi > 0$ as the observation error variance.

Approach 2: Model discrepancy modeled as a zero-mean Gaussian process.

$$\mathbf{Z} = \eta(\mathbf{Y}, \theta) + \delta(\mathbf{S}) + \varepsilon,$$

where $\eta(\mathbf{Y}, \theta)$ and ε are the same as in Approach 1 above. The model error or discrepancy $\delta(\mathbf{S})$ is modeled as a vector of three independent zero mean Gaussian processes below:

$$\delta(\mathbf{S}) \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \Sigma_{d1}(\xi_{d1}) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \Sigma_{d2}(\xi_{d2}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \Sigma_{d3}(\xi_{d3}) \end{bmatrix} \right),$$

where ξ_{dk}=$((\phi_{dk})_i, \kappa_{dk})$ covariance matrix $\Sigma_{dk}(\xi_{dk})$ is as follows:

$$(\Sigma_{dk})_{ij}(\phi_{dk}, \kappa_{dk}) = \kappa_{dk} \exp\left(-\frac{\|\mathbf{s}_i - \mathbf{s}_j\|^2}{\phi_{dk}^2} \right), \quad \kappa_{dk}, (\phi_{dk}) > 0. \tag{5.2.2}$$

In the univariate case, $\delta(\mathbf{S}) \sim N(\mathbf{0}, \Sigma_D(\xi_d))$, where the covariance matrix $\Sigma_D(\xi_d)$ is the same form as equation (5.2.2). While it can be argued that the model discrepancy should not be assumed to have zero mean, in practice it may be too hard to identify a non-zero mean. The Gaussian process is flexible enough to correct for an incorrect mean structure. Further, additional parameters to model the mean of the model discrepancy would be confounded with the climate calibration parameter.

5.2.2.4 Estimation of Emulator Spatial Variance Parameters

Bayarri et al. (2007b) discusses the issue of estimating emulator spatial parameters in a modularization framework; specifically the question of whether to estimate these parameters in a full Bayesian approach as opposed to a plug-in MLE approach. Bayarri et al. (2007b) argues that while a full Bayesian approach would be more informative because uncertainty in the emulator parameters is taken into account, such uncertainties are often small compared to the uncertainties due to the model discrepancy resulting in little difference in the final results. Further, using a full Bayesian approach often leads to a significant increase in computation time. The full Bayesian approach also results in identifiability issues. For example, attempting to estimate the microscale variation (emulator nugget) in the second stage is difficult because it is clearly confounded with the observation error variance. We have therefore used a plug-in approach so far.

We now study the estimation of the emulator spatial variance in a Bayesian framework in the second stage for Model 2, where a model discrepancy term is included. For the univariate case, this consists of estimating κ, in the multivariate case, we need to estimate the cross-covariance matrix $T(\kappa, \rho)$ in the second stage.

5.2.3 Application to Climate Parameter Inference

5.2.3.1 Ocean Tracer Data

In this study, we focus on three tracers that have previously been shown to be informative about K_v in ocean models: Δ^{14}C, trichlorofluoromethane (CFC11), and ocean temperature (T) (cf. Schmittner et al., 2009). ^{14}C (radiocarbon) is a radioactive isotope of carbon, which may be produced naturally and by detonation of thermonuclear devices. ^{14}C and CFC11 enter the oceans from the atmosphere by air-sea gas exchange and are transported from the ocean by advection, diffusion, and to a lesser degree by biological processes (McCarthy, Bower, and Jesson, 1977; Key et al., 2004.)

Δ^{14}C, CFC11, and ocean temperature (T) measurements were collected for all oceanic basins in the 1990s, with locations denoted by a latitude, longitude, and depth. The data have been controlled for quality and gridded by Key et al. (2004). We use the observations from the data synthesis project by Key et al. (2004), which are then aggregated globally (i.e., aggregated over latitude and longitudes), resulting in a data set with $N = 13$ depths. In addition, model output at $p = 10$ different values of K_v, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, and 0.5 cm^2/s, on a grid of locations of latitude, longitude, and ocean depth were evaluated from University of Victoria (UVic) Earth System Climate Model as described in Schmittner et al. (2009). The model output were also aggregated globally, providing a 'depth profile' representing an average between 1990-2000 (cf. Goes et al., 2009). The total number of depths in the model output is $n = 13$, resulting in M=np=130 model output values per tracer. Depths below 3000 m are excluded to minimize problems due to sparse sampling (Key et al., 2004) and model artifacts (Schmittner et al., 2009).

To perform statistical inference on climate parameters, we need to establish a relationship between the observations and the climate parameters. We accomplish this by using an earth system model, which simulates the complex phenomenon of the atmosphere and the oceans under specific input parameter settings to obtain output. The climate models are complex computer codes representing the solution to a large set of differential equations that approximate physical, chemical, and biological processes (Weaver et al., 2001). These climate models often take weeks to months to execute for any given calibration parameter setting, making it very computationally expensive to obtain output at a large number of parameter settings. This provides a compelling argument for using emulation.

5.2.3.2 Implementation Details

In this section, we discuss some of the details of the application of our approach to the ocean tracer data. We verify the emulator using a leave one out cross-validation approach, where we leave out the model output for one calibration parameter (Rougier, 2008b) and predict at all locations for that calibration parameter setting. Plots for the model output and predictions using this cross-validation approach for K_v=0.2 and 0.4 are shown in Figure 5.4. The predictions for the removed locations using cross-validation appear to be visually similar to the original model output (Figure 5.4).

In the second stage, we use MCMC to obtain the posterior distributions of θ. We use a Lognormal (-1.55, 0.59) on θ which reflects the geoscientists' prior uncertainty about K_v based on previous research (Bhat et al., 2009). We use a wide inverse gamma prior for the observation error and model discrepancy variances, specifically $\psi_1 \sim IG(2,10)$ and $\kappa_{d1} \sim IG(2,1000)$ for Δ^{14}C, $\psi_2 \sim IG(2,0.1)$ and $\kappa_{d2} \sim IG(2,0.6)$ for CFC11, and $\psi_3 \sim IG(2,0.1)$ and $\kappa_{d3} \sim IG(2,15)$ for T. We use wide uniform priors for the model discrepancy range parameter. For the emulator spatial variances, we use $\kappa_1 \sim IG(5,24000)$, $\kappa_2 \sim IG(5,2.4)$, and $\kappa_3 \sim IG(5,60)$. When combining multiple spatial fields we can instead place an inverse Wishart prior on the cross covariance matrix T, specifically, $T \sim IW(10, 8T_{MLE})$. Here T_{MLE} is matrix for T obtained by plugging in the MLE from the first stage. The other parameters for the Inverse Wishart were determined using formulae from Anderson (2003) to ensure that the distribution is centered around T_{MLE} and variances of individual matrix elements are relatively small. Specifically, the variances of individual matrix elements decrease as the first parameter of the Inverse Wishart distribution is increased. These priors were obtained after an exploratory analysis of the data suggested the approximate scale of these parameters. While we understand that one needs to be careful about using the data in any way to determine priors, our priors are fairly wide with infinite variance (except for the emulator spatial variance terms), and are not strongly informative.

To ensure convergence of our MCMC based estimates in the second stage, we obtained Monte Carlo standard errors for the posterior mean estimates of θ and other parameters computed by consistent batch means (Jones et al., 2006; Flegal, Haran, and Jones, 2008). The posterior mean estimates of θ had MCMC standard errors below 10^{-4} for both the univariate and bivariate approaches. The MCMC standard errors for the other parameters were less than 10^{-3} for both the univariate and bivariate approaches.

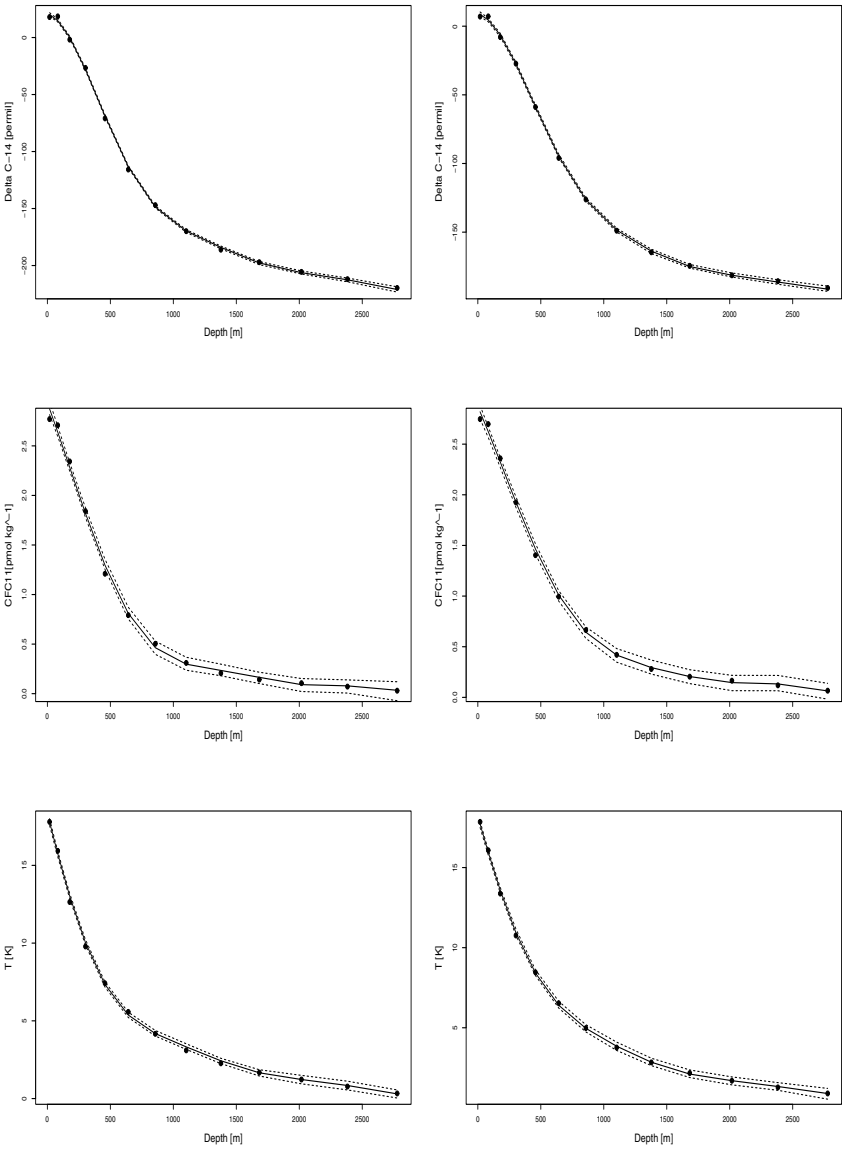


FIGURE 5.4. Cross-validation plots of predictions at K_v=0.2 and 0.4 with model output at K_v value held out. Black dots: model output, solid black lines: predictions, dotted black lines: 95% confidence regions. Left: K_v=0.2, Right: $K_v = 0.4$. Top row: Δ^{14}C, Middle row: CFC11, Bottom row: T.

5.2.4 Results

5.2.4.1 Ocean Tracer Data

In this section we present the results from our analyses using the tracers Δ^{14}C, CFC11, and T. While there is substantial overlap among the posterior distributions of K_v (with model discrepancy is included) obtained by using Δ^{14}C, CFC11, and T separately and then jointly, there are also clear differences (Figure 5.5). We calculated credible regions using the Highest Posterior Density (HPD) method (Chen, Shao, and Ibrahim, 2000). The 90% credible region for K_v using the single tracer Δ^{14}C is between 0.057 and 0.352 cm^2/s, the 90% credible region for K_v using the single tracer CFC11 is between 0.170 and 0.407 cm^2/s, the 90% credible region for K_v using the single tracer T is between 0.156 and 0.420 cm^2/s, and the 90% credible region for K_v using the tracers jointly is between 0.164 and 0.313 cm^2/s. Combining the information from all three tracers results in a sharper posterior distribution of K_v when model discrepancy is included (Figure 5.5).

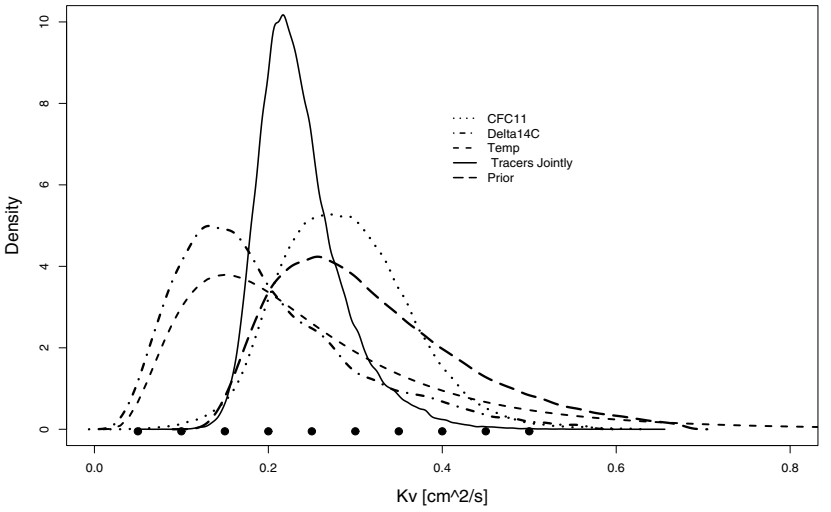


FIGURE 5.5. Log Normal Prior (dotted black line) and posterior of K_v with model discrepancy term included using (i) CFC11 tracer (dotted black line), (ii) Δ^{14}C tracer (dotted-dashed black line), (iii) T tracer (dotted-dashed black line), and (iv) all three tracers jointly (solid black line).

Inclusion of the model discrepancy term appears to shift the posterior probability distribution to the left when we combine the three tracers (Figure 5.6) and when we use the CFC and T tracers individually (Figure 5.7), suggesting that an approach without taking model discrepancy into account results in a bias. The 90% credi-

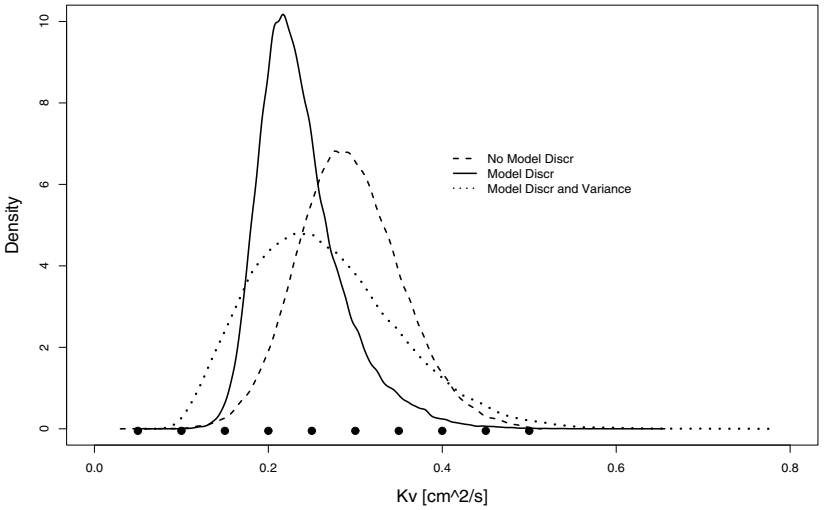
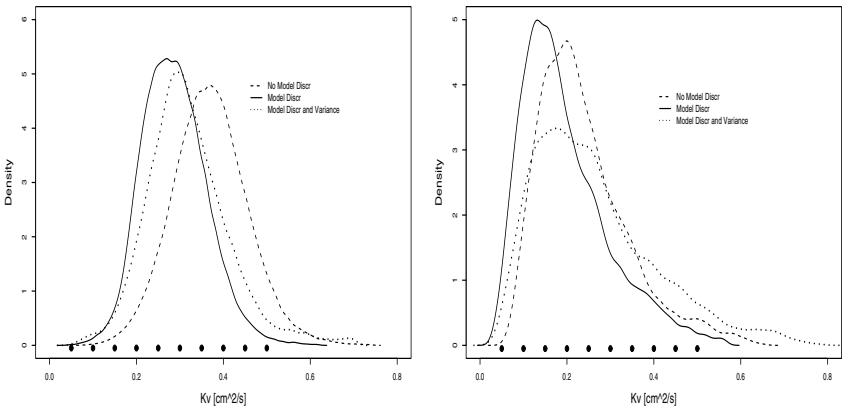


FIGURE 5.6. Posterior for K_v for all three tracers jointly: (i) excluding model error (dotted black line), (ii) including model error (solid black line), and (iii) including model error and estimation of emulator spatial variances (dotted-dashed black line).

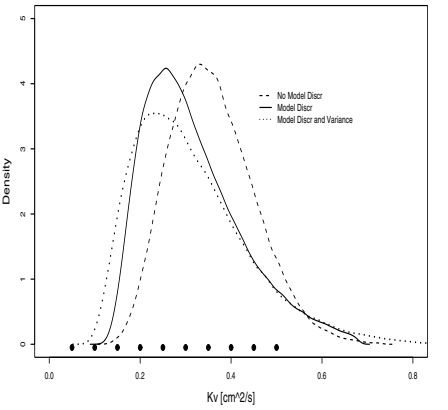
ble region for K_v using the tracers jointly is between 0.194 and 0.390 cm^2/s when model discrepancy *is not* included, between 0.164 and 0.313 cm^2/s when model discrepancy *is* included, and between 0.130 and 0.395 cm^2/s, when emulator spatial variance is also estimated in a fully Bayesian approach. There is little difference when the tracer $\Delta^{14}C$ is used, however, between the approach including model discrepancy and the approach that does not do so. Further, it appears that estimating emulator spatial variance in a fully Bayesian approach results in wider posterior probability distributions, likely due to the additional uncertainty contributed by the emulator. It appears that the posterior distribution for K_v for the tracers jointly is clearly sharper than for the tracers individually for all of the approaches (Figures 5.6 and 5.7).

5.2.4.2 Simulation Study

We investigated the impact of including a model discrepancy term as described in Section 5.2.2.3. Our goal in this study is to determine whether the inclusion of model discrepancy and estimation of emulator spatial variance actually results in better inference of the calibration parameter under different error situations. We hold out a calibration parameter setting, say K_v=0.35, and treat the model output (for all the tracers) for that parameter setting as the observations. We then apply our two stage approach as described earlier to the remainder of the model output and the synthetic



(a) Δ^{14}C (b) CFC11



(c) T

FIGURE 5.7. Posterior for K_v for the three tracers: (i) excluding model error (dotted black line), (ii) including model error (solid black line), and (iii) including model error and estimation of emulator spatial variances (dotted-dashed black line).

observations for all three modeling approaches; exclusion of the model discrepancy term, inclusion of the model discrepancy term, and inclusion of the model discrepancy term plus estimation of the emulator spatial variance. This procedure is executed for all three scenarios below:

Scenario 1: No error. In this scenario, we simply define the observations as the model output at the held out calibration parameter setting. That is, $\mathbf{Z}_k^*(\mathbf{s}_i) = \mathbf{Y}_k(\mathbf{s}_i, \theta^*)$ for $i = 1, \cdots, N$, where θ^* is the held out calibration parameter setting and $k = 1, 2, 3$ denotes the tracer of interest.

Scenario 2: Independent and identically distributed (i.i.d.) error. In this scenario, we add $N(0, \sigma_k^2)$ to the model output at the calibration parameter setting for each location for tracer i. Specifically, $\mathbf{Z}_k^*(\mathbf{s}_i) = \mathbf{Y}_k(\mathbf{s}_i, \theta^*) + \varepsilon_k^*$, where $\varepsilon_{ki}^* \sim N(0, \sigma_k^2)$. Since the scale of the three tracers are different, we must select different values for σ_k^2.

Scenario 3: Model discrepancy plus i.i.d. error. In this scenario, we add a $GP(\mu_k, \Sigma_k)$ to the observations in Scenario 2. Specifically, $\mathbf{Z}_k^*(\mathbf{s}_i) = \mathbf{Y}_k(\mathbf{s}_i, \theta^*) + \delta_k^* + \varepsilon_k^*$, where $\delta_k^* \sim GP(\mu_k, \Sigma_k)$. Again since the scale of the three tracers are different, so are the parameters of the Gaussian processes.

The results of this experiment suggest that adding a model discrepancy term results in more accurate inference and less overfitting for all three scenarios (Figure 5.8(a)-(c)). Estimation of the emulator spatial variance term results in much wider posterior probability distributions for all three scenarios (Figure 5.8(a)-(c)). In Scenario 1, both approaches of excluding and including the model discrepancy term result in having a posterior distribution centered near the held out parameter K_v=0.35. However the posterior distribution is sharper and slightly more accurate when the model parameter is included (Figure 5.8(a)). Estimation of the emulator spatial variance term results in a wider posterior probability distribution, but correctly centered around the held out parameter K_v=0.35 (Figure 5.8(a)). In Scenario 2, excluding the model discrepancy term results in a bias to the left (smaller values of K_v), while including the model discrepancy results in more accurate inference and a sharper posterior for the calibration parameter (Figure 5.8(b)). Estimating the emulator spatial variance results in a slightly biased and wider posterior for the calibration parameter (Figure 5.8(b)). In Scenario 3, excluding the model discrepancy term results in a clearly biased distribution that has a wide bimodal posterior, while including the model discrepancy term results in a posterior distribution that has much less bias, is unimodal, and is sharper (Figure 5.8(c)). Estimating the emulator spatial variance results in a wider posterior distribution that is biased slightly to the left (Figure 5.8(c)). It is important to stress that obtaining such results required much experimentation in determining instructive parameters for σ_k and Σ_k. Simulating observations with too much noise would clearly result in the signal being too weak, and thus the inability to obtain reasonable inference about the calibration parameter, while adding too little error would result in virtually the same inference as in the case with no added error.

Inspired by the suggestion from O'Hagan (2009) that the inclusion of a model discrepancy term may actually result in more biased estimates of the calibration parameter, we attempted to find a situation where the inclusion of the model discrepancy term 'makes the situation worse.' To do so we added a function proportional to 1/depth or 1/depth^2 as 'error' to the model output at K_v=0.35, and we obtained a situation where the inference was more biased and less accurate by including model discrepancy than without model discrepancy (see Figure 5.8(d)). Hence, in some situations, adding model discrepancy may make inference about calibration parameters worse.

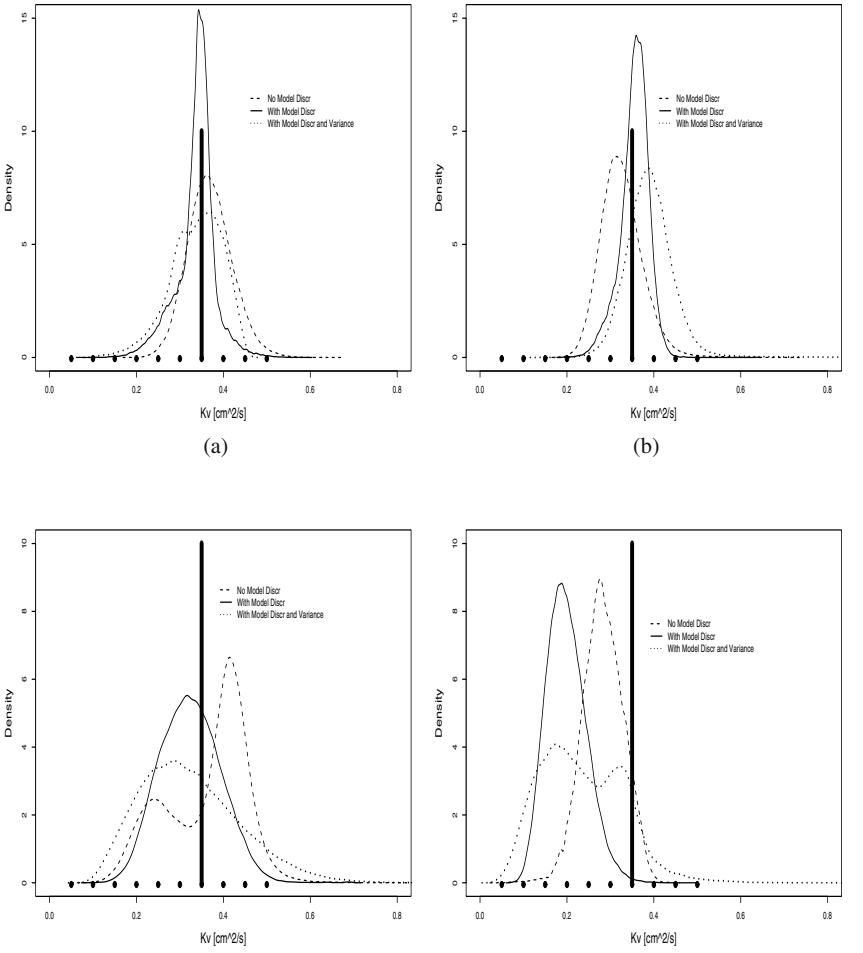


FIGURE 5.8. Posterior for K_v for the three tracers jointly: (i) excluding model error (dotted black line), (ii) including model error (solid black line), and (iii) including model error and estimation of emulator spatial variances (dotted-dashed black line) for simulation experiments. Top left: 'Observations' as model output at K_v=0.35 with no error, Top right: simulated iid error added, Bottom left: simulated model discrepancy. Bottom right: error function proportional to 1/depth^2 added. True parameter value of K_v=0.35 denoted by thick black line.

5.2.5 Summary

We develop and apply an approach for inferring calibration parameters by combining information from observations and climate model output for multiple tracers while taking into account multiple sources of uncertainty. We find that, as one would expect, combining information from multiple spatial fields results in tighter posterior distributions for the climate model parameter. We studied the impact of modeling the model discrepancy and observation error. Based on our study, we find that it is important to include a model discrepancy term, and modeling the discrepancy via a zero mean Gaussian process seems to be the safest approach to guard against bias and overfitting. These results corroborate, in the spatial output setting, the conclusions of Bayarri et al. (2007b). We note, however, that when the computer model is a poor representation of reality, the resulting inference may be more biased when model discrepancy is included. Our study suggests that estimating the emulator spatial variance in a fully Bayesian framework appears to simply reflect the uncertainty from the prior distribution of the emulator spatial variance to the posterior distribution of the calibration parameter. Hence, we recommend using a plug-in estimate of the emulator spatial variance unless there is clear prior information for these parameters.

A possible issue with calibration in general is the known confounding between the calibration parameters and the model discrepancy parameters. We also note that the climate parameter inference obtained here is based on heavily aggregated data, which neglects local spatial effects and small-scale behavior across the ocean, and uses a simple covariance function. Hence, computationally tractable approaches, for large datasets, such as those explored in Bhat et al. (2009), may provide more scientifically rigorous conclusions than those reported here.

Acknowledgments: We thank Andreas Schmittner for providing us the output of the published runs in Schmittner et al. (2009). The authors are grateful to Susie Bayarri, Jim Berger, and others in the 'Interaction of Deterministic And Stochastic Models' working group in the Statistical and Applied Mathematical Sciences (SAMSI) research program on Space-time Analysis in Environmental Mapping, Epidemiology and Climate Change for helpful discussions. We acknowledge support from the National Science Foundation (NSF-HSD) and the U.S. Geological Survey (USGS). Opinions, findings, and conclusions expressed in this work are those of the authors alone, and do not necessarily reflect the views of the NSF or USGS.

Chapter 6
Bayesian Nonparametrics and Semi-parametrics

One of the fastest growing research areas in Bayesian inference is the study of prior probability models for random distributions, also known as nonparametric Bayesian models. While the literature goes back to the 1970s, nonparametric Bayes remained a highly specialized field until the 1990s when new computational methods facilitated the use of such models for actual data analysis. This eventually led to a barrage of new nonparametric Bayesian literature over the last 10 years. In this chapter we highlight some of the current research challenges in nonparametric Bayes.

6.1 Bayesian Nonparametric Goodness of Fit Tests

Surya T. Tokdar, Arijit Chakrabarti, and Jayanta K. Ghosh

In goodness of fit problems we test a parametric null against a nonparametric alternative. For example, we may wish to test the accuracy of an algorithm that is supposed to generate standard normal variates, from n independent observations $X_{1:n} = (X_1, X_2, \cdots, X_n)$ produced by it. We will test for goodness of fit to standard normal against a rather general class of alternatives, which will typically be nonparametric. More generally, it is common to test the normality assumption of a given data set before processing it further. In this case one would test the null hypothesis that the true density is $N(\mu, \sigma^2)$ against a rich nonparametric class of densities. The Bayesian will either put a prior on the null and alternative and calculate the Bayes factor or put an overarching prior on both null and alternative and compute the posterior probability of the null.

The Bayesian literature on these testing problems is still rather meagre, unlike the case of nonparametric estimation on which elegant theory, algorithms, and applications exist; see the review papers by Müller and Quintana (2004) and Choudhuri, Ghosal, and Roy (2005), more references are provided in the following sections. In this section we survey what has been done by way of Bayesian nonparametric

M.-H. Chen et al. (eds.), *Frontiers of Statistical Decision Making and Bayesian Analysis*, DOI 10.1007/978-1-4419-6944-6_6,

testing of goodness of fit problems involving both simple and composite parametric nulls and nonparametric classes of alternatives. We take up the papers in roughly the same sequence they have been published. We also include some unpublished material being written for submission to a journal.

Our survey begins with a review of Rubin and Sethuraman (1965) in Section 6.1.1. Rubin and Sethuraman (1965) use neither a Bayes factor, nor the posterior probability of the null. Instead they start with a classical goodness of fit test and evaluate it asymptotically from a Bayesian point of view.

We next consider in Section 6.1.2 a paper by Dass and Lee (2004) presenting a theoretical treatment of consistency of posterior for testing a simple null hypothesis $H_0 : f = f_0$, where f_0 is a fully specified density like $N(0,1)$ and the alternative is a nonparametric class of densities. Invoking Doob's theorem on consistency (see for example Ghosh and Ramamoorthi, 2003), Dass and Lee (2004) prove that if the prior assigns a positive prior probability to the null and the null is actually true, then the posterior probability of the null will tend to one and hence the Bayesian will accept the null with probability tending to one as sample size n tends to infinity. Dass and Lee (2004) also present consistency theorems for the alternative. The use of Doob's theorem reduces a hard problem to one where we have an answer almost without any technical work.

If the null is composite as in the case of $H_0 : f = N(\mu, \sigma^2)$, $-\infty < \mu < \infty$, $\sigma^2 > 0$, Doob's theorem cannot be used to settle the question of consistency under the null. This case has to be treated by other methods, due to Ghosal, Lember, and van der Vaart (2008) and McVinish, Rousseau, and Mengersen (2009). We discuss these in Section 6.1.3.

Possibly the most important paper for our survey is Berger and Guglielmi (2001) who consider a composite null as described above and embed it into an alternative model constructed with a Polya Tree prior. They argue how such an embedding lets the null parameters carry their identity over to the alternative and discuss choice of objective prior distributions for these parameters. They show how to calculate the Bayes Factor efficiently. Most importantly, they also examine in detail how the test accepts and rejects the null for different data sets and then discuss carefully if such a decision is indeed consistent with our intuitive notions of normality or lack of it for a finite data set. We survey the test and the above facts in Section 6.1.4. We also include in this section a brief discussion of earlier attempts at Bayesian nonparametric testings based on Dirichlet processes and a recent, yet unpublished work that uses Dirichlet mixture priors.

We end this introduction by drawing attention to a closely related problem of Bayesian cluster analysis via Dirichlet mixtures, as developed in Bhattacharya, Ghosh, and Samanta (2009). We may treat each clustering of parameters as related to a partitioning of the X_i's in such a way that all X_i's in the partition are iid $N(\mu_i, \sigma_i^2)$. In particular the case of a single cluster containing all X_i's would correspond to the null hypothesis. Other clusterings will correspond to other models that are part of the alternative.

6.1.1 An Early Application of Bayesian Ideas in Goodness of Fit Problems

One of the earliest applications of Bayesian ideas to a nonparametric goodness of fit problem is due to Rubin and Sethuraman (1965). They don't use a nonparametric prior, in fact their paper predates that of Ferguson's (1973) seminal contribution by nearly a decade. We include it because it is a pioneering Bayesian contribution to a goodness of fit problem with a nonparametric flavor. We will avoid the involved notations in the original paper but simply summarize some relevant facts in our own notation.

Suppose $X_1, \ldots, X_n$ are iid with a normal distribution. Rubin and Sethuraman (1965, Section 3) test

$$H_0\text{: true density is } N(0,1) \text{ against } H_1\text{: true density is } N(\mu, \sigma^2),$$

where (μ, σ^2) is arbitrary, and assume that the critical region is given by large values of the Kolmogorov-Smirnov test statistic. They determine the cut-off point by varying critical regions of this kind and finding the one which minimizes the Bayes risk with respect to their assumed product of loss function and prior probability of (μ, σ^2) under the alternative. The "optimal" threshold is of the form $a\sqrt{\frac{\log n}{n}}$, where a (possibly dependent on n) is essentially determined by the product of loss function and prior. They also determine the region in the parameter space for high acceptance of the null and observe that when $\mu \approx 0$ and $\sigma \approx 1$, this region is approximately the region where the Kolmogorov-Smirnov distance between $N(0,1)$ and $N(\mu, \sigma^2)$ is less than the cut-off point. It is unclear if this test has any other good nonparametric property.

There is an interesting discussion in the last paragraph of Section 2 of Rubin and Sethuraman (1965) about the effect of an "incorrect" use of the constant a, instead of the optimal one for the critical region, which seems to be applicable beyond the example treated in that section. They show that type I error can be highly sensitive to even a small deviation from the optimal a while type II error is not.

6.1.2 Testing a Point Null versus Non-parametric Alternatives

In this section we focus on the issue of consistency in Bayesian testing of a point null hypothesis versus a nonparametric alternative. Operationally a Bayes test is performed through the evaluation of a Bayes factor and consistency (or lack of it) of a test is determined by the behavior of the same as the sample size grows. Our discussion is based on Dass and Lee (2004) and we will usually adhere to the notations adopted by them. We start with the basic definitions first before discussing the main results in their paper.

Consider a complete separable metric space $\mathscr{X}$ with the corresponding Borel sigma field $\mathscr{A}$. Let μ be a σ-finite measure on $(\mathscr{X},\mathscr{A})$ and $\mathscr{F}$ the space of all probability densities with respect to μ with support $\mathscr{X}$. Let P_f be the probability measure having density f. Up to an equivalence class (whereby two densities f and g belong to the same equivalence class if and only if $f = g$ *a.e.* μ), the correspondence between f and P_f is unique. In what follows $\mathscr{X}^\infty$ and P_f^∞ will denote usual infinite products of $\mathscr{X}$ and P_f respectively.

Let X be a random variable taking values in $\mathscr{X}$ with a distribution having density $f \in \mathscr{F}$. We are interested in testing $H_0 : f = f_0$ versus $H_1 : f \neq f_0$, where f_0 is completely specified. In the Bayesian approach to this testing problem, one first puts prior probabilities $p_0 > 0$ and $p_1 = (1-p_0)$ for H_0 and H_1 to be true and also puts a nonparametric prior π_1 on the space H_1. This amounts to defining an overall prior $\pi^*(f) = p_0 I_{f_0}(f) + (1-p_0)\pi_1(f)$ on $H_0 \cup H_1$. The Bayes factor based on sample $X_{1:n} = (X_1, X_2, \cdots, X_n)$, where X_i's are independent copies of X, is defined as

$$\mathrm{BF}_{01}(X_{1:n}) = \frac{\prod_{i=1}^n f_0(x_i)}{\int_{H_1} \prod_{i=1}^n f(x_i)\pi_1(df)}.$$

Taking the usual choice of $p_0 = 1/2$, the Bayes factor exactly equals the posterior odds ratio of H_0 with respect to H_1, given by $\pi^*(H_0 \mid X_{1:n})/\pi^*(H_1 \mid X_{1:n})$. Thus the Bayes factor is a measure of the relative odds of H_0 in comparison to H_1 in the light of the observed data. As the data size grows quite large, ideally the Bayes factor should be such that either H_0 or H_1 is favored very strongly (through their posterior odds ratio), depending on the true data generating density. Thus the Bayes factor is defined to be consistent if

$$\lim_{n\to\infty} \mathrm{BF}_{01}(X_{1:n}) = \infty, \ \ a.s.\ P_{f_0}^\infty, \text{ and}$$
$$\lim_{n\to\infty} \mathrm{BF}_{01}(X_{1:n}) = 0 \ \ a.s.\ P_f^\infty \text{ for any } f \neq f_0.$$

Trivially, consistency of Bayes factor implies and is implied by posterior consistency of the testing procedure.

We are now in a position to state the main results about consistency of Bayes factors in Dass and Lee (2004).

Theorem 6.1. $\lim_{n\to\infty} \mathrm{BF}_{01}(X_{1:n}) = \infty, \ \ a.s.\ P_{f_0}^\infty$.

Theorem 6.2. *Let* $\Theta = \{f \in H_1 : \mathrm{BF}_{01}(X_{1:n}) \to 0\ a.s.\ P_f^\infty\}$. *Then* $\pi_1(\Theta) = 1$.

Before stating the last result we need a definition. Given $f \in \mathscr{F}$, define $K_\varepsilon(f) = \{g \in \mathscr{F} : K(f,g) < \varepsilon\}$ for $\varepsilon > 0$, where $K(f,g)$ is the Kullback-Leibler divergence between f and g. Say that f is in the Kullback-Leibler support of a prior on $\mathscr{F}$ if the prior puts positive probability to the neighborhood $K_\varepsilon(f)$ of f, for each $\varepsilon > 0$.

Theorem 6.3. *Suppose* $f \in H_1$ *and* f *is in the Kullback-Leibler support of* π_1. *Then,* $\lim_{n\to\infty} \mathrm{BF}_{01}(X_{1:n}) = 0 \ \ a.s.\ P_f^\infty$.

Remark 6.1. Theorem 6.1 establishes consistency of Bayes factor under the null for any arbitrary nonparametric prior on H_1, while Theorem 6.2 says that the con-

sistency holds under the alternative for a class of densities that has π_1 probability 1. But Theorem 6.2 is not of much practical use since given a non-null density, one is not sure whether consistency holds or not if the true data were coming from this density. Theorem 6.3 specifies an explicit condition that takes care of this lacuna and can tell us definitively whether the Bayes factor is consistent for any *given* non-null density. For certain nonparametric priors π_1, sufficient conditions ensuring that a density indeed belongs to its Kullback-Leibler support are available in the literature (Ghosal, Ghosh, and Samanta, 1995; Ghosal, Ghosh, and Ramamoorthi, 1999ab; Barron, Schervish, and Wasserman, 1999; Petrone and Wasserman, 2002; Tokdar, 2006; Tokdar and Ghosh, 2007; Choi and Ramamoorthi, 2008; Wu and Ghosal, 2008ab). See also the next section for a general sketch of proof of this result.

We will not give the details of the proofs of the above results, but just sketch the main idea of the proof of Theorem 6.1. This is based on a clever application of Doob's theorem (Ghosh and Ramamoorthi, 2003, pp. 22-24) about posterior consistency. Doob's theorem says that the posterior is consistent for almost all f's under the prior π^* on $H_0 \cup H_1$. Since the prior probability of the (simple) null is positive, this implies that conditionally on the null being true, the posterior probability of the null will tend to 1 for almost all sequences under the null density f_0. This, in turn, implies that the Bayes factor $\mathrm{BF}_{01}(X_{1:n}) \to \infty$ a.s. $P^{\infty}_{f_0}$, since $\mathrm{BF}_{01}(X_{1:n}) = \frac{\pi^*(\{f_0\}|\tilde{x}_n)}{1-\pi^*(\{f_0\}|\tilde{x}_n)}$.

6.1.3 Posterior Consistency for a Composite Goodness of Fit Test

When the null hypothesis is composite, i.e., $H_0 = \{f \in \mathscr{F}_0\}$, where $\mathscr{F}_0$ is not singleton, the study of posterior consistency gets a lot more involved than the elegant treatment of Dass and Lee (2004) discussed in the previous section. We shall denote a Bayesian testing procedure for the composite case by the triplet (p_0, π_0, π_1) with the understanding: $\Pr(H_0) = p_0 = 1 - \Pr(H_1)$, $(f \mid H_0) \sim \pi_0$ and $(f \mid H_1) \sim \pi_1$. We call (p_0, π_0, π_1) consistent if $\Pr(H_0 \mid X_{1:n}) \to 1$ when the null is true, and $\Pr(H_0 \mid X_{1:n}) \to 0$ otherwise. As noted in the previous section, consistency of the testing procedure is equivalent to requiring that the Bayes factor

$$\mathrm{BF}_{01}(X_{1:n}) = \frac{\int \prod_{i=1}^n f(X_i) d\pi_0(f)}{\int \prod_{i=1}^n f(X_i) d\pi_1(f)}$$

of the null hypothesis to the alternative converges asymptotically to infinity when the null is true, and to zero otherwise.

For testing a parametric null against nonparametric alternatives, it is relatively easy to establish $\mathrm{BF}_{01}(X_{1:n}) \to 0$ when X_i's arise from a density f which is not a member of the null parametric family. A common requirement is that density estimation based on π_1 is consistent at the true f in some topology whereas the closure of $\mathscr{F}_0$, determined by the same topology, leaves out f. For example, if f belongs to the Kullback-Leibler support of π_1, then density estimation based on

π_1 is weakly consistent at f, due to Schwartz's theorem (see Schwartz, 1965 or Chapter 4.4 of Ghosh and Ramamoorthi, 2003). For any such f that is outside the weak closure of $\mathscr{F}_0$, $\mathrm{BF}_{01}(X_{1:n}) \to 0$ asymptotically. This result can be formally proved by arguing that the composite prior $\pi^* = 0.5\pi_0 + 0.5\pi_1$ has f in its Kullback-Leibler support and hence $\mathscr{F}_0$, lying entirely outside a weak neighborhood of f, receives vanishing mass from the posterior distribution $\pi^*(\cdot \mid X_{1:n})$. Consequently, $\mathrm{BF}_{01}(X_{1:n}) = \pi^*(\mathscr{F}_0|X_{1:n})/[1-\pi^*(\mathscr{F}_0|X_{1:n})] \to 0$.

It is much more difficult to prove $\mathrm{BF}_{01}(X_{1:n}) \to \infty$ when $f \in \mathscr{F}_0$. This is because the usual choices of π_1 contain $\mathscr{F}_0$ in their support and can recover an $f \in \mathscr{F}_0$ from data nearly as efficiently as the parametric model itself. See, for example, the almost-parametric rate of convergence of Dirichlet mixture of normal priors discussed in Ghosal and van der Vaart (2001). The special case of a simple null hypothesis $H_0 : f = f_0$ as discussed in Section 6.1.2, eschews this difficulty because no estimation is required for the null model. For the composite null case, one needs a careful comparison of how π_1 concentrates around an $f \in \mathscr{F}_0$ in comparison to π_0. Both Ghosal, Lember, and van der Vaart (2008) and McVinish, Rousseau, and Mengersen (2009) discuss formal ways to make this comparison based on neighborhoods of $f \in \mathscr{F}_0$, shrinking at the same rate at which π_1 can recover an f in a density estimation setting. Ghosal, Lember, and van der Vaart (2008) discuss goodness of fit tests as a special case of the more general problem of adaptive selection of nested models. For this review, we summarize the more direct treatment of McVinish, Rousseau, and Mengersen (2009).

To fix notations, let $\mathscr{F}_0 = \{f_\theta : \theta \in \Theta\}$ where Θ is a d-dimensional Euclidean subspace. Fix an $f^* = f_{\theta^*}$ for some $\theta^* \in \Theta$ and let P^* denote the product measure of $X_{1:\infty}$ under this density. Assume that for some universal constants $c > 0, C > 0$ the following conditions hold.

A1. Density estimation based on π_1 is consistent at f_{θ^*} with rate ε_n (see Ghosal, Ghosh, and van der Vaart, 2000 and Shen and Wasserman, 2001). That is, there exists a metric $d(\cdot,\cdot)$ on the densities and positive numbers $\varepsilon_n \downarrow 0$ such that

$$\pi_1(\{f : d(f, f_{\theta^*}) < \varepsilon_n\} \mid X_{1:n}) \to 1$$

in P^* probability.

A2. $\pi_0(K_n := \{\theta : K(f^*, f_\theta) < cn^{-1}, V(f^*, f_\theta) < cn^{-1}\}) \geq Cn^{-d/2}$, where for any two densities p and q, $K(p,q) = \int p \log(p/q)$ and $V(p,q) = \int p(\log(p/q))^2$. One could replace K_n with a simpler definition: $K_n = \{\theta : \int f^*(\log(f^*/f_\theta))_+ < cn^{-1}\}$, where $x_+ = \max(x, 0)$.

A3. $\pi_1(A_n := \{f : d(f, f_{\theta^*}) < \varepsilon_n\}) = o(n^{-d/2})$.

A2 is a natural support condition which holds for many standard parametric models. Conditions A1 and A3 maintain a fine balance with respect to the rate ε_n. Slowing down this rate favors A1 but may lead to a violation of A3, speeding up has the opposite effect. The main result of McVinish, Rousseau, and Mengersen (2009, Theorem 1) and its elegant proof offered by these authors are reproduced below.

Theorem 6.4. $[\mathrm{BF}_{01}(X_{1:n})]^{-1} \to 0$ *in* P^* *probability.*

Proof. Express the inverse of the Bayes factor as

$$\begin{aligned}[\mathrm{BF}_{01}(X_{1:n})]^{-1} &= \frac{\int_{A_n} \prod_{i=1}^n \frac{f(X_i)}{f^*(X_i)} \pi_1(df)}{\int_\Theta \prod_{i=1}^n \frac{f_\theta(X_i)}{f^*(X_i)} \pi_0(d\theta)} \frac{1}{\pi_1(A_n \mid X_{1:n})} \\ &\le \frac{n^{d/2} \int_{A_n} \prod_{i=1}^n \frac{f(X_i)}{f^*(X_i)} \pi_1(df)}{n^{d/2} \int_{K_n} \prod_{i=1}^n \frac{f_\theta(X_i)}{f^*(X_i)} \pi_0(d\theta)} \frac{1}{\pi_1(A_n \mid X_{1:n})}\end{aligned}$$

From A1, $[\pi_1(A_n \mid X_{1:n})]^{-1} = O_{P^*}(1)$. Define $L_n(\theta) = \frac{1}{n} \sum_{i=1}^n \log \frac{f^*(X_i)}{f_\theta(X_i)}$, $\theta \in \Theta$ and take π_{0,K_n} to be the restriction of π_0 to K_n. Then for small $\delta > 0$,

$$\begin{aligned}&P^*\left(n^{d/2} \int_{K_n} \prod_{i=1}^n \frac{f_\theta(X_i)}{f^*(X_i)} \pi_0(d\theta) \le \delta \right) \\ &\le P^*\left(\int L_n(\theta) \pi_{0,K_n}(d\theta) \ge \frac{1}{n} \log \frac{\pi_0(K_n)}{\delta n^{-d/2}} \right) \\ &\le P^*\left(\int [L_n(\theta) - K(f^*, f_\theta)] \pi_{0,K_n}(d\theta) \ge \frac{1}{n} \left[\log \frac{\pi_0(K_n)}{\delta n^{-d/2}} - c \right] \right) \\ &\le \frac{n \int V(f^*, f_\theta) \pi_{0,K_n}(d\theta)}{(\log \pi_0(K_n) + (d/2) \log n - \log \delta - c)^2} \\ &\le \frac{c}{(\log(C/\delta) - c)^2},\end{aligned}$$

where the first inequality follows from Jensen's inequality, the third from Chebyshev's inequality, and the second and the fourth from A2. Therefore,

$$[n^{d/2} \int_{K_n} \prod_{i=1}^n \frac{f_\theta(X_i)}{f^*(X_i)} \pi_0(d\theta)]^{-1} = O_{P^*}(1).$$

An application of Markov's inequality shows

$$P^*\left(n^{d/2} \int_{A_n} \prod_{i=1}^n \frac{f(X_i)}{f^*(X_i)} \pi_1(d\theta) > \delta \right) \le \delta^{-1} n^{d/2} \pi_1(A_n)$$

and hence, by A3, $n^{d/2} \int_{A_n} \prod_{i=1}^n \frac{f(X_i)}{f^*(X_i)} \pi_1(d\theta) = o_{P^*}(1)$. Combining these we get $\mathrm{BF}_{01}(X_{1:n})^{-1} = o_{P^*}(1)$.

A number of recent papers give sharp rates of convergence for popular nonparametric priors on densities, including the Dirichlet mixture priors (Ghosal and van der Vaart, 2007) and the logistic Gaussian process priors (van der Vaart and van Zanten, 2008). However, a rigorous demonstration of A3 is currently available for only a handful of cases, such as the Dirichlet mixture of Bernstein polynomials, log spline densities (Ghosal, Lember, and van der Vaart, 2008), and mixtures of triangular densities (McVinish, Rousseau, and Mengersen, 2009). It is not yet

clear whether A3 is a natural condition that is going to be automatically satisfied by many nonparametric priors, as it requires the prior distribution not to sit too tightly around elements of $\mathscr{F}_0$. Some insight into A3 may be gained by considering $a_1 n^{-1/2} \leq \varepsilon_n \leq a_2 n^{-1/2+\text{Be}a}$ for some Be$a \in (0, 1/2)$, $a_1, a_2 > 0$. In this case one can replace A3 with

A3′. $\pi_1(\{f : d(f, f^*) < \varepsilon\}) = O(\varepsilon^D)$ for some $D > d/(1/2 - \beta)$.

The above gives a sharp lower bound on the effective dimensionality of the nonparametric prior π_1 around f^*. Deeper investigations of the popular choices of π_1 will reveal whether such conditions hold, possibly for well chosen values of hyperparameters. McVinish, Rousseau, and Mengersen (2009), however, argue the necessity of A1-A3 via an example where a violation of these conditions leads to inconsistency of the testing procedure.

If it turns out that use of popular nonparametric priors like Dirichlet mixtures of normals may lead to inconsistency in goodness of fit tests, one should explore the possibility of introducing an indifference zone to separate the null and the alternative. This would make the consistency problem easier to resolve but specification of the indifference zone cannot be done without some subjective input from the user.

6.1.4 Bayesian Goodness of Fit Tests

The earliest papers, for example Florens, Richard, and Rolin (1996) and Carota and Parmigiani (1996), use a mixture of Dirichlet process prior (Antoniak, 1974; Ghosh and Ramamoorthi, 2003, p 113) on the alternative hypotheses. Letting θ denote the parameter underlying the null model, these authors use the following hierarchical formulation of the alternative: the alternative distribution given θ follows a Dirichlet process (DP) prior with precision constant α_θ and base measure $\bar{\alpha}_\theta$ and θ is modeled by a parametric prior that may match with the prior used on the null model. As the authors of these papers and also Berger and Guglielmi (2001) point out, this approach is flawed as the prior on the alternative sits on the discrete distributions, whereas in goodness of fit problems, we want it to sit on densities. Both these papers run into serious difficulties caused by this. A more promising paper, also mentioned and discussed in more detail by Berger and Guglielmi (2001), is Verdinelli and Wasserman (1998) which assigns a mixture of Gaussian processes as the prior on the alternative. The prior is constructed quite ingeniously and allows calculation of the Bayes factor.

Why are the priors on the alternatives mixtures? This is done so that the composite null like $N(\mu, \sigma^2)$ can have a natural embedding, i.e., nesting in some sense in the alternative space. For example, Florens, Richard, and Rolin (1996) ensure that for every θ, the conditional predictive distribution of a single observation is the same under both hypotheses. In the same vein, but avoiding a prior like the mixture of Dirichlet or the ingenious construction used by Verdinelli and Wasserman (1998),

Berger and Guglielmi (2001) take resort to a mixture of Polya tree priors (Lavine, 1992, 1994; Mauldin, Sudderth, and Williams, 1992).

Suppose the parametric null is $\theta \in \Theta$ and the distribution of X corresponding to θ is μ_θ. Berger and Guglielmi (2001) require

$$\mathbb{E}P_\theta = \mu_\theta, \quad \text{for all } \theta \in \Theta,$$

where P_θ is the random distribution of X under the nonparametric alternative and θ on the left is a hyperparameter of the Polya tree prior under the alternative.

There are many advantages of the Polya tree priors. A Polya tree process defines a random measure by assigning random conditional probabilities to sets within a nested dyadic partition of an infinite depth. For each dyadic set its random conditional probability given the parent in the previous partition is assigned a beta distribution with two shape parameters, namely, α's, which function like the two parameters of the Dirichlet process. They specify the mean and precision. Unlike the Dirichlet process, the use of infinitely many precision constants offers a greater control over the smoothness of the random measure defined by a Polya tree process. Indeed, by choosing a common value r_m for the precision constants at depth m of the partition, such that $\sum_m r_m^{-1/2} < \infty$, one ensures that the Polya tree sits on densities and obtains posterior consistency for estimating a large class densities (Ghosh and Ramamoorthi, 2003, pp. 186, 190; Walker, 2003, 2004; see also an earlier weaker result in Barron, Schervish, and Wasserman, 1999). Berger and Guglielmi (2001) discuss two constructions of a Polya tree prior, one where only the partition depends on θ and another where only the beta shape parameters depend on θ. Either choice embeds as prior mean the conditional predictive density under the null model with parameter θ. Berger and Guglielmi (2001) use the second approach in their examples, as a fixed partition offers significant computational advantages, as detailed in their Section 5. This flexible embedding of the null model within the alternative appears to be a key motivation for Berger and Guglielmi (2001) to use mixtures of Polya tree priors for the alternative model. We quote

> ...as we were striving for a nonsubjective approach, we wished to utilize (typically improper) noninformative prior distributions for θ in the models. Although θ occurs under both H_0 and H_1 , it could have very different meanings under each and could, therefore, require different noninformative priors. We felt that use of the mixture Polya tree process could solve this problem, in the sense that we could use results from Berger and Pericchi (1996c) and Berger, Pericchi, and Varshavsky (1998) to show that the usual noninformative priors for θ are appropriately calibrated between H_0 and H_1 when the alternative is a suitable mixture of Polya tree processes.

We refer the reader to Berger and Guglielmi (2001) for the details of their subtle justification for using the same noninformative prior for θ under both null and alternative.

Berger and Guglielmi (2001) also introduce an additional scaling factor h for r_m free of the partition depth m. They use different values of h to see the robustness of the Bayes factor with respect to h. They minimize over h to make a conservative choice of the Bayes Factor in their Table 1. This is another innovative contribution of

Berger and Guglielmi (2001), for which the reader needs to go back to the original paper to appreciate its full significance.

Berger and Guglielmi (2001) examine the results of the test for three examples. We consider only the first one with data being 100 stress rupture life times of Kevlar vessels, taken from Andrews and Herzberg (1985, p 183). In the log scale the data are claimed to be $N(\mu, \sigma^2)$ under the null.

Table 1 of Berger and Guglielmi (2001) gives the minimum of the Bayes factor with respect to the constant h for the two different choices of the Polya tree mixture mentioned above. The values of the minimized Bayes factor are very close to each other (in the range 0.0072 to 0.00180) and to the value of the Bayes factor based on the prior of Verdinelli and Wasserman (1998). This is a pleasant fact that provides support to both the Bayes factor tests and shows that the introduction of h seems to ameliorate the problem of the proper choice of r_m by reporting the minimized Bayes factor.

All these nice properties of the Polya tree priors notwithstanding, the random densities chosen by the prior have discontinuities on the countable dense set formed by the boundaries of the sets in the partition. This is at least one reason why there has been a lot of interest in priors sitting on smooth densities like Gaussian process priors, vide Tokdar and Ghosh (2007), Tokdar (2007), Lenk (1988, 1991) and the Dirichlet process mixture of normals, vide Lo (1984), Ferguson (1983), Escobar and West (1995), Müller, Erkanli, and West (1996), MacEachern and Müller (1998, 2000) and MacEachern (1998). Recently, for testing H_0: $f = N(\mu, \sigma^2)$ for some μ, σ^2, Tokdar (2009) has proposed the following alternative model: H_1: $f = \int N(\phi, \rho\sigma^2) P(d\phi, d\rho)$ where the mixing distribution P is modeled as $p(P \mid \mu, \sigma^2) = \mathrm{DP}(\alpha, G_0(\cdot \mid \mu, \sigma^2))$ with

$$G_0(\phi, \rho \mid \mu, \sigma^2) = N(\phi \mid \mu, (1-\rho)\sigma^2)\mathrm{Be}(\rho \mid \omega_1, \omega_2).$$

This alternative ensures $\mathbb{E}[f \mid \mu, \sigma^2, H_1] = N(\mu, \sigma^2)$ and thus provides an embedding similar to the one in Berger and Guglielmi (2001). Tokdar (2009) also discuss efficient computation of the Bayes factor via a variation of the sequential importance sampling algorithm of Liu (1996).

6.2 Species Sampling Model and Its Application to Bayesian Statistics

Jaeyong Lee

Since its introduction, the Dirichlet process (Ferguson, 1973) has been a central model in Bayesian nonparametric statistics. From the beginning, two important properties of the Dirichlet process have been known: discreteness and marginalization property of the Dirichlet process. The discreteness is often considered problematic in statistical modeling, for most statistical models deal with continous densities.

An easy remedy for this is the use of mixtures of Dirichlet processes, a convolution of a Dirichlet process with a continuous density (Lo, 1984; Escobar and West, 1995; Müller, Erkanli, and West, 1996). It is ironic that the problematic feature and its remedy turn out a great success in Bayesian nonparametric modeling, because it has found vast applications covering unexpected areas. For recent applications of Dirichlet mixtures, see Teh et al. (2006) and Fox, Sudderth, and Willsky (2007), for examples.

The marginalization property of the Dirichlet process, discussed in Blackwell and MacQueen (1973), is that a random sample from a random distribution with a Dirichlet process prior forms a Polya urn sequence. The Polya urn sequence has a predictive distribution of a very simple form and has been the main computational tool for implementation of posterior inference in Dirichlet mixture models. The computational effort for posterior simulation in Dirichlet mixture models reduces to essentially the same as posterior simulation for a related parametric model. Posterior computation can be done without sampling random probability measure unlike other nonparametric priors (Lavine, 1992; Lee, 2007).

The marginalization property of the Dirichlet process has also found important applications in population genetics, ecology, number theory, and combinatorics. See Kingman (1975), Aldous (1985), Pitman (1996, 2003) and references therein. In this connection, a new class of random probability measures, called species sampling models (SSM), has emerged. The SSM is the directing (or de Finetti) measure of the species sampling sequence (SSS) which is an exchangeable sequence of random variables with a certain form of predictive distribution. SSM and SSS are generalizations of the Dirichlet process and the Polya urn sequence, respectively. This literature has been largely neglected by the Bayesian community until recently. Only recently inference with SSM's has become an active research area in Bayesian statistics. On one side, the SSM is defined by a predictive distribution, and thus potentially shares the simplicity of posterior simulation with the Dirichlet process. On the other side, the SSM provides a large family of predictive distributions, allowing data analysts greater flexibility. Research on applications of SSM's to statistical inference includes Ishwaran and James (2003), Lijoi, Mena, and Prünster (2005), Navarrete, Quintana, and Müller (2008), Lijoi, Prünster, and Walker (2008), and Jang, Lee, and Lee (2010), to name just a few.

In this section, we review the basic theory of the SSM and applications. In Section 6.2.1, we introduce the exchangeable random partition and the SSS. In Section 6.2.2, we discuss three methods to construct the SSS, a generalization of the Chinese restaurant process, Poisson-Kingman partitions, and Gibbs partitions. In Section 6.2.3, we discuss the application of SSM's mixture models to statistical inference and some asymptotic results. The section concludes with a discussion.

6.2.1 Basic Theory

A SSM is characterized by the distribution of the exchangeable random partition and a diffuse probability measure. In this section, we review the basic theory of the SSM, and introduce the exchangeable random partition and SSS.

6.2.1.1 Exchangeable Random Partitions

We first introduce basic concepts of the exchangeable random partition. This is an important concept in the theory of SSM's. For a positive integer n, let $[n] = \{1,2,\ldots,n\}$. Any ordered finite sequence $\mathbf{n} = (n_1,n_2,\ldots,n_k)$ is called a composition of n if

$$n_i \geq 1,\ 1 \leq i \leq k,\ \text{and}\ \sum_{i=1}^{k} n_i = n \tag{6.2.1}$$

and an unordered finite sequence $\{n_1,n_2,\ldots,n_k\}$ with the same property (6.2.1) is called a partition of n. For any partition $\{A_1,A_2,\ldots,A_k\}$ and permutation σ of $[n]$, let $\sigma(\{A_1,A_2,\ldots,A_k\}) = \{\sigma(A_1),\sigma(A_2),\ldots,\sigma(A_k)\}$ where $\sigma(A) = \{\sigma(a) : a \in A\}$. A random partition $\Pi_n = \{A_1,A_2,\ldots,A_k\}$ of $[n]$ is called exchangeable if Π_n and $\sigma(\Pi_n)$ has the same distribution for any permutation σ of $[n]$, i.e., for any partition $\{A_1,A_2,\ldots,A_k\}$ and permutation σ of $[n]$,

$$P(\Pi_n = \{A_1,A_2,\ldots,A_k\}) = P(\sigma(\Pi_n) = \{A_1,A_2,\ldots,A_k\}).$$

It is not hard to see that Π_n is an exchangeable random partition of $[n]$ if and only if for any partition $\{A_1,A_2,\ldots,A_k\}$ of $[n]$,

$$P(\Pi_n = \{A_1,A_2,\ldots,A_k\}) = p(|A_1|,|A_2|,\ldots,|A_k|) \tag{6.2.2}$$

for some symmetric function p defined on $\mathscr{C}_n$, where $|A|$ is the cardinality of a set A and $\mathscr{C}_n$ is the set of all compositions of n. The function p in (6.2.2) is called an exchangeable partition probability function (EPPF) of Π_n.

The distribution of an exchangeable random partition can be characterized by a distribution of exchangeable random variables. Let $(x_1,x_2,\ldots,x_n)$ be an ordered list of n elements. Define an equivalence relation on $[n]$ from $(x_1,x_2,\ldots,x_n)$ as follows:

$$\text{for any } i,j \in [n],\ i \sim j \text{ if and only if } x_i = x_j. \tag{6.2.3}$$

The equivalence classes of the equivalence relation (6.2.3) induces a partition of $[n]$. Let $\Pi(x_1,\ldots,x_n)$ be the partition of $[n]$ defined by equivalence classes of (6.2.3). If $(X_1,X_2,\ldots,X_n)$ is an exchangeable sequence of random variables, then $\Pi(X_1,X_2,\ldots,X_n)$ is an exchangeable random partition. The following theorem asserts that in fact any exchangeable random partition can be obtained by exchangeable random variables.

Theorem 6.5. *Let Π_n be an exchangeable random partition of $[n]$ and $\pi_n = (N_{n,1}^{\downarrow}, N_{n,2}^{\downarrow}, \ldots, N_{n,k}^{\downarrow})$ be a partition of n defined by decreasing arrangement of block sizes of Π_n. Then,*

$$\Pi_n | \pi_n \stackrel{d}{=} \Pi(X_1, X_2, \ldots, X_n) | \pi_n,$$

where $(X_1, X_2, \ldots, X_n)|\pi_n \stackrel{d}{=} (x_{\sigma(1)}, x_{\sigma(2)}, \ldots, x_{\sigma(n)})$, σ is a uniform random permutation of $[n]$ and $(x_1, x_2, \ldots, x_n)$ is the sequence of $N_{n,1}^{\downarrow}$ 1's, $N_{n,2}^{\downarrow}$ 2's,..., $N_{n,k}^{\downarrow}$ k's.

We now consider exchangeable random partition defined on the set of positive integers $\mathbb{N}$. Suppose Π_n is an exchangeable random partition of $[n]$. For $1 \le m \le n$, let $\Pi_{m,n}$ be the restriction of Π_n to $[m]$, i.e., the partition of $[m]$ obtained from Π_n by removing $\{m+1, \ldots, n\}$. If Π_n is an exchangeable random partition of $[n]$, then $\Pi_{m,n}$ is an exchangeable random partition of $[m]$, $1 \le m \le n$. A sequence of random partition $\Pi_\infty = (\Pi_n)_{n \ge 1}$ is called an infinite exchangeable random partition (or exchangeable random partition on $\mathbb{N}$), if

(i) Π_n is an exchangeable random partition of $[n]$ for all n; and
(ii) $\Pi_m = \Pi_{m,n}$ a.s. for all $1 \le m \le n < \infty$.

The notion of the EPPF can be also extended to $\mathbb{N}$. Let $\mathscr{C} = \cup_{n=1}^{\infty} \mathscr{C}_n$. For $\mathbf{n} = (n_1, \ldots, n_k) \in \mathscr{C}$, let $\mathbf{n}^{j+} = (n_1, \ldots, n_{j-1}, n_j + 1, n_{j+1}, \ldots, n_k)$ for $j = 1, 2, \ldots, k$ and $\mathbf{n}^{(k+1)+} = (n_1, \ldots, n_k, 1)$. A function $p : \mathscr{C} \to [0,1]$ is called an (infinite) EPPF of (Π_n) if

(a) $p(1) = 1$;
(b) (addition rule) for all $\mathbf{n} \in \mathscr{C}$, $p(\mathbf{n}) = \sum_{j=1}^{k+1} p(\mathbf{n}^{j+})$; and
(c) $p|_{\mathscr{C}_n}$ is the EPPF of Π_n for all n.

The characterization of the exchangeable random partition through exchangeable random variables still holds. The infinite exchangeable random partition $\Pi_\infty = (\Pi_n)$ is said to be an exchangeable random partition of $\mathbb{N}$ generated by $(X_n)_{n \ge 1}$ if $\Pi_n = \Pi(X_1, \ldots, X_n)$ for all n. For a given Π_∞, let $(N_{n,i}^{\downarrow}, i \ge 1)$ be the decreasing arrangement of block sizes of Π_n, where $N_{n,i}^{\downarrow} = 0$ if Π_n has fewer than i blocks.

Theorem 6.6. (Kingman's representation) *Let $\Pi_\infty = (\Pi_n)$ be an exchangeable random partition of $\mathbb{N}$ and $(N_{n,i}^{\downarrow}, i \ge 1)$ be the decreasing arrangement of block sizes of Π_n for $n \ge 1$. Then, $\mathbf{n}^{\mathbf{k}+}$*

(a) $\dfrac{N_{n,i}^{\downarrow}}{n} \to P_i^{\downarrow}$ *a.s. for all* $i \ge 1$.

(b) $\Pi_\infty | (P_i^{\downarrow}, i \ge 1) \stackrel{d}{=} \Pi(X_1, X_2, \ldots)$, *where $X_1, X_2, \ldots$ follow iid F and F has ranked atoms $(P_i^{\downarrow}, i \ge 1)$.*

Remark 6.1. A consequence of Theorem 6.6 is that Π_∞ is an exchangeable random partition of $\mathbb{N}$ if and only if there exists a random probability distribution F such that $\Pi_\infty = \Pi(X_1, X_2, \ldots)$ where $X_1, X_2, \ldots | F \stackrel{iid}{\sim} F$.

Remark 6.2. The random distribution F is used as a nonparametric prior in the nonparametric Bayesian context. Thus, for each exchangeable random partition of $\mathbb{N}$, there is a corresponding random probability measure F.

6.2.1.2 Species Sampling Sequences

In this subsubsection, we take a closer look at the sequence of exchangeable random variables which induces an exchangeable random partition and consequently a random probability measure. We form a sequence of random variables $(X_1, X_2, \ldots)$ in the following way. Suppose we land on an unknown planet with infinitely many species. As we explore the planet, we encounter species and give a name X_i to the ith species encountered. We record whether the ith and jth species encountered are the same, $X_i = X_j$. For convenience, we pick a point in a complete separable metric space $\mathscr{X}$ randomly and use it as a name whenever we need a new name.

For a mathematical treatment, we introduce some notation. Let M_j be the index of the first appearance of the jth species, i.e., let $M_1 = 1$ and $M_j = \inf\{n : n > M_{j-1}, X_n \notin \{X_1, \ldots, X_{n-1}\}\}$ for $j \geq 2$, where $\inf \emptyset = \infty$. Let $\tilde{X}_j = X_{M_j}$ be the jth distinct species to appear which is defined on the event $M_j < \infty$. Let $n_j = n_{jn}$ be the number of times the jth species $\tilde{X}_j$ appears in $(X_1, \ldots, X_n)$

$$n_{jn} = \sum_{m=1}^{n} I(X_m = \tilde{X}_j), \quad j = 1, 2, \ldots$$
$$\mathbf{n} = \mathbf{n}_n = (n_{1n}, n_{2n}, \ldots) \text{ or } (n_{1n}, n_{2n}, \ldots, n_{k_n,n}),$$

where $k = k_n = \max\{j : n_{jn} > 0\}$ be the number of different species to appear in $(X_1, \ldots, X_n)$.

Let ν be a diffuse (or atomless) probability measure on $\mathscr{X}$. We now give the definition of the SSS. An exchangeable sequence $(X_1, X_2, \ldots)$ is called a species sampling sequence if $X_1 \sim \nu$ and

$$X_{n+1} | X_1, \ldots, X_n \sim \sum_{j=1}^{k_n} p_j(\mathbf{n}_n) \delta_{\tilde{X}_j} + p_{k_n+1}(\mathbf{n}_n) \nu,$$

where δ_x is the degenerate probability measure at x. Here ν is called the base probability measure and the sequence of functions $(p_1, p_2, \ldots)$ is called a sequence of predictive probability functions (PPF). Each p_j is a positive real-valued function defined on $\mathscr{C}$ and satisfies $\sum_{j=1}^{k_n+1} p_j(\mathbf{n}) = 1$ for all $\mathbf{n} \in \mathscr{C}$. Note that

$$p_j(\mathbf{n}) = \mathbb{P}(X_{n+1} = \tilde{X}_j | X_1, \ldots, X_n), \; j = 1, \ldots, k,$$
$$p_{k_n+1}(\mathbf{n}) = \mathbb{P}(X_{n+1} \notin \{X_1, \ldots, X_n\} | X_1, \ldots, X_n).$$

The SSS defines an exchangeable random partition by $(\Pi_n) = (\Pi(X_1, \ldots, X_n))$ and its EPPF p and PPF (p_j) have the following relation:

$$p_j(\mathbf{n}) = \frac{p(\mathbf{n}^{j+})}{p(\mathbf{n})}, \text{ for all } \mathbf{n} \in \mathscr{C},\ 1 \le j \le k+1.$$

It is clear from the definition that the distribution of the SSS is characterized by the PPF (or EPPF) and base measure.

The following theorem (Pitman, 1996) characterizes the directing measure of a SSS.

Theorem 6.7. *Let $(X_n)_{n\ge1}$ be a SSS with PPF (p_j) and the base probability measure ν. Then,*

(i) $\dfrac{n_{jn}}{n} \to \tilde{P}_j$ *a.s. for some positive random variable $\tilde{P}_j$ for all $j \ge 0$.*

(ii) $||F_n - F|| \to 0$ *a.s., as $n \to \infty$, where $||\cdot||$ is the total variation norm, and*

$$F_n = \sum_{j=1}^{k_n} p_j(\mathbf{n}_n)\delta_{\tilde{X}_j} + p_{k_n+1}(\mathbf{n}_n)\nu,$$
$$F = \sum_j \tilde{P}_j\delta_{\tilde{X}_j} + (1 - \sum_j \tilde{P}_j)\nu. \tag{6.2.4}$$

(iii) $(\tilde{X}_j)$ *are independent of* $(\tilde{P}_j)$ *and follow iid ν.*

(iv) *Conditionally on F, (X_i) are iid F.*

Remark 6.3. A sequence of random variables (X_n) is a SSS if and only if $X_1, X_2, \ldots | F$ is random sample from F where

$$F = \sum_{i=1}^{\infty} P_i\delta_{\tilde{X}_i} + R\nu \tag{6.2.5}$$

for some sequence of positive random variables (P_i) and R such that $1 - R = \sum_{i=1}^{\infty} P_i \le 1$, $(\tilde{X}_i)$ is a random sample from ν, and (P_i) and $(\tilde{X}_i)$ are independent.

Remark 6.4. The random probability distribution (6.2.4) is the directing measure of the species sampling sequence (X_i) and it is called the species sampling model of (X_i). We will also sometimes call it the species sampling process (or prior). Since it is characterized by the EPPF p (or PPF (p_j)) and a diffuse probability measure ν, we will denote it by $SSM(p, \nu)$ or $SSM((p_j), \nu)$. Similarly, the distribution of the SSS is denoted by $SSS(p, \nu)$ or $SSS((p_j), \nu)$.

Example 6.1. Dirichlet process. The celebrated Dirichlet process (Ferguson, 1973) is a special case of the SSM. Suppose $(X_1, X_2, \ldots)$ is a sample from F which follows $DP(\theta\nu)$, the Dirichlet process with parameter $\theta\nu$ with $\theta > 0$. Then, marginally $(X_1, X_2, \ldots)$ is a Pólya urn sequence with distribution $X_1 \sim \nu$ and

$$X_{n+1} | X_1, \ldots, X_n \sim \sum_{j=1}^{k_n} \frac{n_{jn}}{n+\theta}\delta_{\tilde{X}_j} + \frac{\theta}{n+\theta}\nu.$$

Thus, $(X_1, X_2, \ldots)$ is a SSS with the base probability measure ν and PPF

$$p_j(\mathbf{n}_n) = \frac{n_{jn}}{n+\theta} I(1 \le j \le k_n) + \frac{\theta}{n+\theta} I(j = k_n + 1). \tag{6.2.6}$$

Sethuraman (1994) showed that the Dirichlet process can be represented in form (6.2.5). Let $W_1, W_2, \ldots$ be an iid sequence from $Beta(1,\theta)$. From (W_i), discrete probability masses (P_i) are constructed by the stick breaking process. In particular,

$$\begin{aligned} P_1 &= W_1, \\ P_j &= W_j \prod_{i=1}^{j-1} (1 - W_i), \quad j = 2, 3, \ldots. \end{aligned} \tag{6.2.7}$$

Suppose $\tilde{X}_1, \tilde{X}_2, \ldots$ is an iid sequence from ν, and (P_i) and $(\tilde{X}_i)$ are independent. Then,

$$F = \sum_{j=1}^{\infty} P_j \delta_{\tilde{X}_j} \sim DP(\theta\nu).$$

Example 6.2. Pitman-Yor process. Pitman and Yor (1997) introduced an interesting class of discrete random measures which includes the Dirichlet process. Let a and b be real numbers with either $0 \le a < 1$ and $b > -a$ or $a < 0$ and $b = -ma$ for some $m = 1, 2, \ldots$ and let ν be a diffuse probability measure. For $j = 1, 2, \ldots,$ let

$$W_j \sim Beta(1 - a, b + ja).$$

Construct (P_i) from (W_j) by the stick breaking process as in (6.2.7). Let $(\tilde{X}_j)$ be an iid sequence from ν independent of (P_j). We call the random probability measure $F = \sum_{j=1}^{\infty} P_j \delta_{\tilde{X}_j}$ the Pitman-Yor process and denote it by $PY(a, b, \nu)$. The distribution of (P_j) is called Griffiths-Engen-McClosky distribution with parameter (a, b) or $GEM(a, b)$ and the ranked frequency $(P_j^{\downarrow})$ of (P_j) is called Poisson-Dirichlet distribution with parameter (a, b) or $PD(a, b)$, where $P_j^{\downarrow}$ is the jth largest value among P_j's. Note $PY(0, \theta, \nu)$ is $DP(\theta\nu)$. Sometimes Poisson-Dirichlet distribution is referred with only one parameter $\theta > 0$, in this case $PD(\theta) = PD(0, \theta)$.

The following theorem of Pitman (1996) characterizes the posterior distribution of the SSM.

Theorem 6.8. *Suppose* $X_1, X_2, \ldots, X_n | F \overset{iid}{\sim} F$ *and*

$$F = \sum_{i=1}^{\infty} \tilde{P}_j \delta_{\tilde{X}_j} + (1 - \sum_{j=1}^{\infty} \tilde{P}_j)\nu.$$

Then, the conditional distribution of F *given* $X_1, X_2, \ldots, X_n$ *are determined as follows.*

(a) $\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_{k_n}$ *are measurable functions of* $(X_1, X_2, \ldots, X_n)$.

(b) $\tilde{X}_j \overset{iid}{\sim} \nu$, *for* $j > k_n$.

(c) $(\tilde{P}_j)$ *and* $(\tilde{X}_j)$ *are independent and for all nonnegative measurable functions* f,

$$E(f(\tilde{P})|X_1,X_2,\ldots,X_n,\mathbf{N}=\mathbf{n}) = \frac{E(f(\tilde{P})\Pi(\mathbf{n},\tilde{P}))}{p(\mathbf{n})},$$

where $\Pi(\mathbf{n},\tilde{P}) = \left(\prod_{i=1}^{k-1}[\tilde{P}_i^{n_i-1}(1-\sum_{j=1}^{i}\tilde{P}_j)]\right)\tilde{P}_k^{n_k-1}$ *and* $p(\mathbf{n}) = E\Pi(\mathbf{n},\tilde{P})$.

Corollary 6.1. *Suppose* $F \sim PY(\alpha,\theta,\nu)$ *and* $X_1,X_2,\ldots,X_n|F \overset{iid}{\sim} F$, *where* $\tilde{X}_1,\ldots,\tilde{X}_k$ *are distinct values of* $X_1,X_2,\ldots,X_n$ *and* $n_i = |\{j : X_j = \tilde{X}_i\}|$. *Then,*

$$F \mid X_1,\ldots,X_n \overset{d}{=} \sum_{i=1}^{k}\tilde{P}_j\delta_{\tilde{X}_j} + \tilde{R}_kF_k,$$

where $(\tilde{P}_1,\tilde{P}_2,\ldots,\tilde{P}_k,\tilde{R}_k) \sim Dirichlet(n_1-\alpha,n_2-\alpha,\ldots,n_k-\alpha,\theta+k\alpha)$ *and* $F_k \sim PY(\alpha,\theta+k\alpha,\nu)$.

6.2.2 Construction Methods for EPPFs

A SSM is characterized by an EPPF (or PPF) and a diffuse base probability measure ν. Thus, to obtain different SSM, we need to have methods to construct an EPPF. In this subsection, we review three commonly used methods.

6.2.2.1 Generalization of Chinese Restaurant Process

Consider a restaurant which has infinite number of circular tables $1,2,\ldots$ and each table can accommodate an infinite number of customers. Entering the restaurant, customers are seated according to the following seating plan.

- Customer 1 sits at table 1.
- Suppose n customers entered the restaurant and they are seated at tables $1,2,\ldots,k$ with $n_j \geq 1$ customers at table j for $1 \leq j \leq k$ with $\sum_{j=1}^{k} n_j = n$. Upon entering the restaurant, customer $n+1$ is seated at table j with probability $p_j(\mathbf{n})$ for $1 \leq j \leq k$, or alone at table $k+1$ with probability $p_{k+1}(\mathbf{n})$, where $\sum_{j=1}^{k+1} p_j(\mathbf{n}) = 1$ and $p_j(\mathbf{n}) \geq 0$, where $\mathbf{n} = (n_1,\ldots,n_k)$.

Let $A_i = \{1 \leq j \leq n :$ customer j is seated at table $i\}$. Then $\Pi_n = \{A_1,A_2,\ldots,A_k\}$ defines a partition of $[n]$. Note that not every p_j results in an exchangeable random partition. For a more detailed discussion on this issue, see Pitman (1995).

Example 6.3. For $\theta > 0$, let

$$p_j(n_1,\ldots,n_k) = \begin{cases} \frac{n_i}{n+\theta}, & j = 1,2,\ldots,k, \\ \frac{\theta}{n+\theta}, & j = k+1. \end{cases}$$

In this case, the PPF results in an symmetric EPPF which is

$$p(n_1, n_2, \ldots, n_k) = \frac{\theta^k \prod_{i=1}^k (n_i - 1)!}{(\theta)_{n\uparrow}}, \tag{6.2.8}$$

where $(\theta)_{n\uparrow} = \theta(\theta+1)\cdots(\theta+n-1)$.

Example 6.4. For either $\alpha = -k < 0, \theta = mk$ for some $m = 1, 2, \ldots$ or $0 \le \alpha \le 1$ and $\theta > -\alpha$, let

$$p_j(n_1, \ldots, n_k) = \begin{cases} \frac{n_i - \alpha}{n+\theta}, & j = 1, 2, \ldots, k, \\ \frac{\theta + k\alpha}{n+\theta}, & j = k+1. \end{cases}$$

The above PPF also results in an EPPF and it is

$$p(n_1, n_2, \ldots, n_k) = \frac{(\theta)_{k\uparrow\alpha} \prod_{i=1}^k (1-\alpha)_{n_i - 1\uparrow}}{(\theta)_{n\uparrow}},$$

where $(\theta)_{k\uparrow\alpha} = \theta(\theta+\alpha)\cdots(\theta+(k-1)\alpha)$.

6.2.2.2 Poisson-Kingman Partitions

In Section 6.2.1, we have seen that an exchangeable random partition can be obtained by a random sample (X_n) from a random distribution with atoms (P_i). In this subsubsection, we review a method to construct (P_i) using the jump times of a non-homogeneous Poisson process on $(0, \infty)$ and its EPPF.

Suppose $(N(t), t \ge 0)$ is the Poisson process on $(0, \infty)$ with intensity measure $\Lambda(dx) = \rho(x)dx$ such that

$$\int_0^1 x\Lambda(dx) < \infty \text{ and } \Lambda[1, \infty) < \infty.$$

Let $J_1^{\downarrow} \ge J_2^{\downarrow} \ge \ldots$ be the ordered jump times of $N(t)$ and $T = \sum_{i=1}^{\infty} J_i^{\downarrow}$. For the following discussion, we assume that the random variable T has density $f(t)$ that is strictly positive and continuous on $(0, \infty)$. Let $P_i^{\downarrow} = J_i^{\downarrow}/T$ for $i \ge 1$. The distribution of $(P_i^{\downarrow})$ is called the Poisson-Kingman distribution with Levy density ρ and denoted by $PK(\rho)$. The conditional distribution of $(P_i^{\downarrow})$ given $T = t$ is denoted by $PK(\rho|t)$. We denote the distribution of $PK(\rho|t)$ with mixing distribution $\gamma(dt)$ by

$$PK(\rho, \gamma) = \int_0^{\infty} PK(\rho|t)\gamma(dt).$$

Define a random permutation $(\tau(i), i \in \mathbb{N})$ of $\mathbb{N}$ as follows: $\tau(1) = i$ with probability $P_i^{\downarrow}, i \ge 1$. After $\tau(1), \ldots, \tau(j-1)$ are determined, $\tau(j) = i$ with probability

$$\frac{P_i^{\downarrow}}{1-P_{\tau(1)}-P_{\tau(2)}-\ldots-P_{\tau(j-1)}}, \text{ for } i \in \mathbb{N}-\{\tau(1),\ldots,\tau(j-1)\}.$$

Let $\tilde{P}_i = P^{\downarrow}_{\tau(i)}$ for $i \geq 1$. The sequence $(\tilde{P}_i)$ is called the size-biased permutation of $(P_i^{\downarrow})$.

The EPPF of an exchangeable random partition $(\Pi(X_1,X_2,\ldots,X_n), n \geq 1)$ is obtained in the following theorem (Pitman, 2003), where $X_1,X_2,\ldots$ are a random sample from a random distribution F with atoms $(P_i^{\downarrow})$.

Theorem 6.9. *The EPPF of the partition induced by a driving measure with $PK(\rho|t)$ prior is given by*

$$p(n_1,n_2,\ldots,n_k|t) = t^{k-1}\int_0^1 p^{n+k-2} I(n_1,n_2,\ldots,n_k;tp)\tilde{f}(p|t)dp,$$

where $n=\sum_{i=1}^k n_i$, $I(n_1,n_2,\ldots,n_k;v)$ is defined as follows: for $k=1$ and $n_1=n$, $I(n;v)=1$ and

$$I(n_1,\ldots,n_k;v) = \frac{1}{\rho(v)}\int_{S_k}\prod_{i=1}^k \rho(v-u_i)u_i^{n_i}du_1\ldots du_{k-1},$$

$S_k = \{(u_1,\ldots,u_k) : u_i \geq 0, u_1+\ldots+u_k = 1\}$, and $\tilde{f}(p|t)$ is the conditional density of $\tilde{P}_1$ given $T=t$, which is given by

$$\tilde{f}(p|t) = pt\rho(pt)\frac{f((1-p)t)}{f(t)},\ 0<p<1.$$

The EPPF of $PK(\rho)$ partition is given by

$$p(n_1,n_2,\ldots,n_k) = \int_0^\infty \ldots \int_0^\infty \frac{f(v)dv\prod_{i=1}^k \rho(x_i)x_i^{n_i}dx_i}{(v+\sum_{i=1}^k x_i)^n}. \tag{6.2.9}$$

Example 6.5. If $\rho(x) = \theta x^{-1}e^{-bx}$, $\theta, b > 0$, then $T \sim Gamma(\theta,b)$ with mean θ/b. One can compute the EPPF using (6.2.9) and it is the same as (6.2.8). Thus, $PK(\rho) = PD(\theta)$.

6.2.2.3 Gibbs Partitions

Motivated by the form of the EPPF of the Dirichlet process, Gnedin and Pitman (2006) study the EPPF of Gibbs form. The exchangeable random partition Π on $\mathbb{N}$ is said to be of Gibbs form if there exists sequences random variables $(W_j, j \geq 1)$ and $(V_{n,k}, n, k \geq 1)$ such that the EPPF of Π is given by

$$p(n_1,\ldots,n_k) = V_{n,k}\prod_{j=1}^{k} W_{n_j}.$$

There are redundancies in the representation of the EPPF of Gibbs form. In fact, for any $\gamma > 0$, if we change W_j to $\gamma^j W_j$ and $V_{n,k}$ to $\gamma^{-n}V_{n,k}$, or W_j to γW_j and $V_{n,k}$ to $\gamma^{-k}V_{n,k}$, the EPPF does not change. The following proposition (Gnedin and Pitman, 2006) characterizes the EPPF of Gibbs form.

Proposition 6.1. *The sequences of random variables* $(W_j, j \geq 1)$ *and* $(V_{n,k}, n, k \geq 1)$ *define a partition of Gibbs form if and only if there exist* $b \geq 0$ *and* $a \leq b$ *such that*

(a) $W_j = (b-a)_{j-1\uparrow b}$, $j = 1,2,\ldots$; *and*
(b) $V_{n,k} = (bn-ak)V_{n+1,k} + V_{n+1,k+1}$, $1 \leq k \leq n$.

This proposition motivates the following definition. For $\alpha < 1$, an exchangeable random partition Π is said to follow Gibbs of type α if

$$W_j = \begin{cases} (1-\alpha)_{j-1\uparrow}, & -\infty < \alpha < 1, \\ 1, & \alpha = -\infty, \end{cases}$$
$$V_{n,k} = \gamma_{n,k}V_{n+1,k} + V_{n+1,k+1},$$

with $V_{1,1} = 1$ and

$$\gamma_{n,k} = \begin{cases} n-\alpha k, & -\infty < \alpha < 1, \\ k, & \alpha = -\infty. \end{cases}$$

6.2.3 Statistical Applications

6.2.3.1 Mixture Modeling with SSM

The realization of a SSM is with positive probability a discrete probability measure. A simple way to remedy this is to convolute the SSM with a family of distributions with continuous density $h(x|\theta)$ for $\theta \in \Theta$. Suppose G is a realization of $SSM(p,\nu)$ where p is an EPPF and ν is a diffuse probability measure. Define a probability measure F with continuous density by

$$f(x) = \int h(x|\theta)G(d\theta). \tag{6.2.10}$$

There are many choices for h. Popular choices of h are the densities of the normal distribution $N(\mu,\sigma^2)$ with $\mu \in \mathbb{R}$ and $\sigma > 0$, and the uniform distribution $U(\mu-\sigma,\mu+\sigma)$ with $\mu \in \mathbb{R}$ and $\sigma > 0$. In both cases, $\theta = (\mu,\sigma)$.

The SSM mixture model has the following structure. For an EPPF p and a diffuse probability measure ν,

$$G \sim SSM(p,\nu),$$
$$X_1, X_2, \ldots, X_n | F \overset{iid}{\sim} F, \tag{6.2.11}$$

where F has density of form (6.2.10). Model (6.2.11) is equivalent to

$$\begin{aligned} G &\sim SSM(p,\nu), \\ \theta_1,\theta_2,\ldots,\theta_n|G &\overset{iid}{\sim} G, \\ X_i|\theta_i &\overset{ind}{\sim} h(x|\theta_i),\ i=1,2,\ldots,n. \end{aligned} \tag{6.2.12}$$

By integrating out G from (6.2.12), the sequence $\theta_1,\theta_2,\ldots,\theta_n$ becomes the species sampling sequence, i.e.,

$$\theta_1,\theta_2,\ldots,\theta_n \sim SSS(p,\nu).$$

In a Markov chain Monte Carlo (MCMC) simulation of the posterior, one uses the following structure:

$$\begin{aligned} \theta_1,\theta_2,\ldots,\theta_n &\sim SSS(p,\nu), \\ X_i|\theta_i &\sim h(x|\theta_i),\ i=1,2,\ldots,n. \end{aligned}$$

Since the conditional distribution of θ_i given θ_{-i}, the vector obtained by removing θ_i from $(\theta_1,\theta_2,\ldots,\theta_n)$, is well known for the species sampling sequence, the conditional posterior distribution of θ_i given $\theta_{-i},X_1,\ldots,X_n$ is easily obtained and is

$$\pi(d\theta_i|\theta_{-i},X_1,\ldots,X_n) \propto h(x_i|\theta_i)\Big\{\sum_{j=1}^{k_{-i}} p_j(\mathbf{n}_{-i})\delta_{\theta^*_{-i,j}}(d\theta_i) + p_{k+1}(\mathbf{n}_{-i})\nu(d\theta_i)\Big\}, \tag{6.2.13}$$

where $\mathbf{n}_i$ is the composition of $n-1$ generated by θ_{-i}, k_{-i} is the number of blocks in θ_i, and $\theta^*_{-i,j}$ is the jth distinct value in θ_{-i}. The Gibbs sampling can be done by sampling θ_i from the distribution (6.2.13). For more elaborate computation algorithms, see Navarrete, Quintana, and Müller (2008).

6.2.3.2 Large Sample Properties

Jang, Lee, and Lee (2010) consider a simple nonparametric model

$$X_1,\ldots,X_n|F \sim F,\quad F \sim SSM(p,\nu), \tag{6.2.14}$$

and obtain necessary and sufficient conditions for posterior consistency under quite general assumptions. Let

$$F_0 = \sum_j q_j\delta_{z_j} + \lambda\mu \tag{6.2.15}$$

be the true distribution from which $X_1,X_2,\ldots,X_n$ are sampled, where $z_j \in \mathscr{X}$, $q_j \geq 0$, $\lambda = 1-\sum_j q_j \leq 1$ and μ is a diffuse probability measure. Let $\mathbf{n} = (n_1,n_2,\ldots,n_k)$ be the composition of n generated by the block sizes of $\Pi(X_1,X_2,\ldots,X_n)$, let $\tilde{X}_j$, $1 \leq j \leq k$, be the distinct values of $X_1,X_2,\ldots,X_n$, and $\mathscr{Z} = \{z_1,z_2,\ldots\}$.

Suppose the predictive probability function of $SSM(p,\nu)$ satisfies the smoothness condition:

$$S_n = S_n(\mathbf{n}) = \max_{1\le i\le k} \sum_{j=1}^{k} \left| p_j(\mathbf{n}) - p_j(\mathbf{n}^{i+}) \right| \to 0, \quad F_0 - a.s., \text{ as } n \to \infty, \tag{6.2.16}$$

and the support of the discrete part of F_0 satisfies the separability condition: there exists $\varepsilon > 0$ such that for all $i \neq j$

$$d(z_i, z_j) > \varepsilon, \tag{6.2.17}$$

where d is the metric of $\mathscr{X}$. Under these assumptions, Jang, Lee, and Lee (2010) obtained the following theorem.

Theorem 6.10. *Suppose $X_1, X_2, \ldots$ is an iid sequence from F_0 of form (6.2.15) with separability condition (6.2.17). Under the model (6.2.14) with prior $SSM(p,\nu)$ which satisfies the smoothness condition (6.2.16), the posterior given $X_1, \ldots, X_n$ is weakly consistent at P_0 if and only if the predictive probability function satisfies*

$$\lim_{n\to\infty} \sum_{j=1}^{k} |p_j(\mathbf{n}) - n_j/n| I(\tilde{X}_j \in \mathscr{Z}) = 0, \; F_0 - a.s.,$$

and one of the following holds

(i) $p_{k+1}(\mathbf{n}) \to 0$ *as* $n \to \infty$, $F_0 - a.s.$
(ii) F_0 *is a mixture of a discrete probability measure and the diffuse measure* ν.

The message of the theorem is that unless the predictive distribution behaves similar to the empirical distribution, posterior consistency is not granted. This restricts the class of $SSM(p,\nu)$ priors with posterior consistency to essentially the Dirichlet process. This point becomes clearer if one considers $PY(a,b,\nu)$. If the prior class is the Pitman-Yor process, the only priors which give rise to consistent posteriors under all true probability measures are Dirichlet processes. This fact has also been proved independently by James (2008). Also, Jang, Lee, and Lee (2010) show that some popular sub-classes of SSM's include no prior with consistent posteriors. If one considers the mixture models with SSM, however, the posterior consistency holds for much wider class of SSM. See Jang, Lee, and Lee (2010) and Lijoi, Prünster, and Walker (2005).

6.2.4 Discussion

In this section, we have reviewed the basic theory of SSMs and their statistical applications. An SSM is defined through the predictive distribution. Since the success of the Dirichlet process relies greatly on the simplicity of its predictive probability function, the SSM would seem to be a promising alternative for non-parametric

Bayesian inference. However, real success remains to be seen. From the outset, the SSM overcomes a shortcoming of Dirichlet process mixture models. Suppose in the mixture model (6.2.12), instead of $SSM(p, \nu)$ the Dirichlet proces $DP(\alpha\nu_\theta)$ is used, where $\alpha > 0$ and ν_θ belongs to a parametric family of distribution with parameter θ. The parameter α is often called the prior sample size, because it controls the amount of information the prior has. In many applications, one wishes to set $\alpha \approx 0$ to represent little prior knowledge. But, this has the unexpected effect that the mixture model becomes essentially the parametric model α_θ. This can be seen from Sethuraman's representation (6.2.7). It is due to the fact that α controls both the amount of the prior information and the distribution of the cluster size. It can be overcome by separating parameters for these two effects, for example, by adopting Pitman-Yor process. Thus, the flexibility of the general SSM does enhance Bayesian nonparametric modeling.

We have added a new tool to the Bayesian toolbox which is quite versatile and flexible. However, we need to hone this tool more to deploy in action. Some comments are in order. First, although in theory the SSM frees data analysts from the rigid restriction on the form of the PPF, practical use of wide class of PPFs seems to be limited currently, and computational methodology for general PPFs needs to be developed. Second, from the asymptotic point of view, it is not clear yet to what extent SSM's can be recommended, although there are some indications as discussed in Section 6.2.3. Thirdly, more freedom may confuse the data analysts even more. Practical guidelines for the choice of PPFs in actual data analysis will be beneficial to data analysts. Finally, examples of real data analysis with SSM's would convince more data analysts to use SSM.

6.3 Hierarchical Models, Nested Models, and Completely Random Measures

Michael I. Jordan

Statistics has both optimistic and pessimistic faces, with the Bayesian perspective often associated with the former and the frequentist perspective with the latter, but with foundational thinkers such as Jim Berger reminding us that statistics is fundamentally a Janus-like creature with two faces. In creating one field out of two perspectives, one of the unifying ideas emphasized by Berger and others is the Bayesian hierarchy, a modeling framework that simultaneously allows complex models to be created and tames their behavior.

Another general tool for creating complex models while controlling their complexity is by nesting simplified models inside of more complex models, an appeal to the principle of "divide-and-conquer." An example is the classical finite mixture model, where each data point is modeled as arising from a single mixture component. Note that this appeal to divide-and-conquer is quite different from the re-

cursive principle underlying hierarchical modeling—the latter strategy provides a way to share statistical strength among components while the former strategy tends to isolate components. Of course, many complex models involve a blend of these strategies.

If the need to exploit hierarchical and nested structures is compelling in parametric models, it is still more compelling in Bayesian nonparametrics, where the growth in numbers of degrees of freedom creates significant challenges in controlling model complexity. The basic idea of Bayesian nonparametrics is to replace classical finite-dimensional prior distributions with general stochastic processes, thereby allowing an open-ended number of degrees of freedom in a model. The framework expresses an essential optimism—only an optimist could hope to fit a model involving an infinite number of degrees of freedom based on finite data. But it also expresses the pessimism that simplified parametric models may be inadequate to capture many real-world phenomena, particularly in the setting of large data sets in which increasingly subtle aspects of those phenomena may be revealed. From either perspective, care needs to be taken to exploit and manage the large number of degrees of freedom available within a nonparametric model.

In this section we discuss hierarchical and nested modeling concepts within the framework of Bayesian nonparametrics. To keep the discussion focused, we restrict ourselves to a special class of stochastic processes known as "completely random measures." These random measures have the simplifying property that they assign independent random mass to disjoint regions of a probability space. This property turns out to imply that these measures are discrete (up to a deterministic component and a Brownian motion component that are of limited value for Bayesian modeling). While the discreteness is limiting for some applications, it also has some significant virtues. In particular it provides a natural tool for focusing on structural aspects of models, where the effects of hierarchy and nesting have relatively simple interpretations.

6.3.1 Completely Random Measures

Letting Ω denote a measurable space endowed with a sigma algebra $\mathscr{A}$, a *random measure* G is a stochastic process whose index set is $\mathscr{A}$. That is, $G(A)$ is a random variable for each set A in the sigma algebra. A *completely random measure* G is defined by the additional requirement that whenever A_1 and A_2 are disjoint sets in $\mathscr{A}$, the corresponding random variables $G(A_1)$ and $G(A_2)$ are independent. This idea generalizes the notion of "independent increments processes" that is familiar in the special case in which Ω is the real line.

Kingman (1967) presented a way to construct completely random measures based on the nonhomogeneous Poisson process. This construction has significant consequences for Bayesian modeling and computation; in particular, it allows connections to be made to the exponential family and to conjugacy. The construction is as follows (see Figure 6.1 for a graphical depiction). Consider the product space

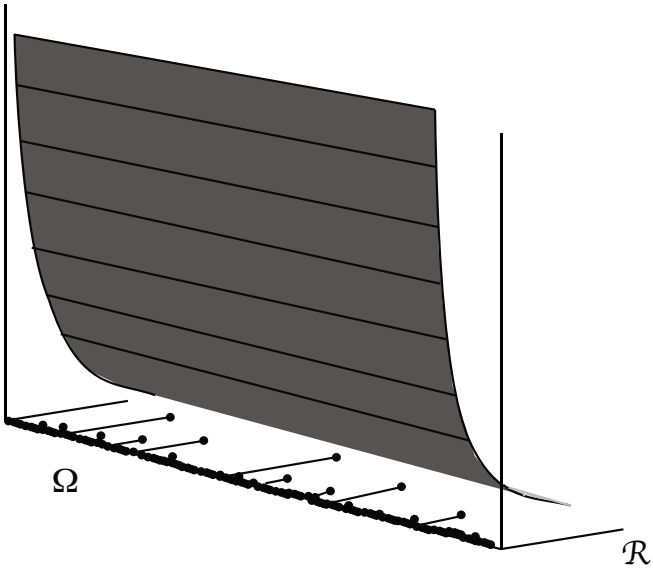


FIGURE 6.1. The construction of a completely random measure on Ω from a nonhomogeneous Poisson process on $\Omega \otimes \Re$.

$\Omega \otimes \Re$, and place a sigma-finite product measure η on this space. Treating η as the rate measure for a nonhomogeneous Poisson process, draw a sample $\{(\omega_i, p_i)\}$ from this Poisson process. From this sample, form a measure on Ω in the following way:

$$G = \sum_{i=1}^{\infty} p_i \delta_{\omega_i}. \tag{6.3.1}$$

We refer to $\{\omega_i\}$ as the *atoms* of the measure G and $\{p_i\}$ as the *weights*.

Clearly the random measure defined in (6.3.1) is completely random because the Poisson process assigns independent mass to disjoint sets. The interesting fact is that all completely random processes can be obtained this way (up to a deterministic component and a Brownian motion).

As an example of a completely random measure, define the rate measure η as a product of an arbitrary sigma-finite measure B_0 on Ω and an "improper" beta distribution on $(0,1)$:

$$\eta(d\omega, dp) = cp^{-1}(1-p)^{c-1} dp\, B_0(d\omega),$$

where $c > 0$. Note that the expression $cp^{-1}(1-p)^{c-1}$ integrates to infinity; this has the consequence that a countably infinite number of points are obtained from the Poisson process. The resulting completely random measure is known as the *beta*

process.[1] We denote a draw from the beta process as follows:

$$B \sim \mathrm{BP}(c, B_0),$$

where $c > 0$ is referred to as a *concentration parameter* and where B_0 is the *base measure*. Note that for the beta process the weights $\{p_i\}$ lie in the interval $(0,1)$. Their sum is finite (a consequence of Campbell's theorem), with the magnitude of the sum controlled by the concentration parameter c and by $B_0(\Omega)$. The locations of the atoms are determined by B_0.

As a second example, let the rate measure be a product of a base measure G_0 and an improper gamma distribution:

$$\eta(d\omega, dp) = cp^{-1}e^{-cp}dp\,G_0(d\omega). \tag{6.3.2}$$

Again the density on p integrates to infinity, yielding a countably infinite number of atoms. The resulting completely random measure is known as the *gamma process*. We write:

$$G \sim \mathrm{GP}(c, G_0)$$

to denote a draw from the gamma proess. Note that the weights $\{p_i\}$ lie in $(0,\infty)$ and their sum is again finite.

It is also of interest to consider random measures that are obtained from completely random measures by normalization. For example, returning to the rate measure defining the gamma process in (6.3.2), let $\{(\omega_i, p_i)\}$ denote the points obtained from the corresponding Poisson process. Form a random probability measure as follows:

$$G = \sum_{i=1}^{\infty} \pi_i \delta_{\omega_i}, \tag{6.3.3}$$

where $\pi_i = p_i / \sum_{j=1}^{\infty} p_j$. This is the famous *Dirichlet process (DP)* (Ferguson, 1973). We denote a draw from the DP as $G \sim \mathrm{DP}(\alpha_0, H_0)$, where $\alpha_0 = G_0(\Omega)$ and $H_0 = G_0/\alpha_0$. Note that the DP is *not* a completely random measure—for disjoint sets A_1 and A_2, the random variables $G(A_1)$ and $G(A_2)$ are negatively correlated due to the normalization.

6.3.2 Marginal Probabilities

At this point it is useful to recall De Finetti's theorem, which states that infinitely exchangeable sequences of random variables are obtained by drawing a random element G and then drawing the elements of the sequence independently conditional on G. Given the ubiquity of this conditional independence motif in Bayesian mod-

[1] For further details on this derivation of the beta process, see Thibaux and Jordan (2007). For an alternative derivation that does not make use of the framework of completely random measures, see Hjort (1990).

eling, it is of interest to ask what kinds of exchangeable sequences are obtained if G is one of the random measures discussed in the previous section.

In the case of the DP, the answer is classical—one obtains the *Pólya urn model* (Blackwell and MacQueen, 1973). This model can be described in terms of the closely related *Chinese restaurant process (CRP)* (Aldous, 1985). Consider a restaurant with an infinite number of tables, listed in some (arbitrary) order. Customers enter the restaurant sequentially. The first customer sits at the first table. Subsequent customers choose a table with probability proportional to the number of customers already sitting at that table. With probability proportional to a parameter α_0 they start a new table. This defines the CRP as a distribution on partitions of the customers. To define the Pólya urn we augment the model to place a parameter ϕ_k at the kth table, where the $\{\phi_k\}$ are drawn independently from some distribution G_0. All customers sitting at the kth table are assigned the parameter ϕ_k. Letting θ_i denote the parameter assigned to the ith customer, this defines an exchangeable sequence $(\theta_1, \theta_2, \ldots)$.

This connection between the DP and the CRP and Pólya urn model can be understood by noting that the representation of the DP in (6.3.3) is essentially a mixture model with a countably infinite number of components. We can view the CRP as a draw from this mixture model. In particular, let us associate an integer-valued variable W_i to the ith customer as follows:

$$\begin{aligned} G &\sim \mathrm{DP}(\alpha_0, G_0), \\ p(W_i = k \mid G) &= \pi_k, \quad i = 1, \ldots, n. \end{aligned}$$

In the language of the CRP, the event $\{W_i = k\}$ means that the ith customer sits at the kth table. In essence, the DP defines an infinite collection of random probabilities that, when integrated out according to the De Finetti construction, yield the CRP. The specific rule underlying the CRP—that customers sit at a table proportional to the number of customers already at that table—reflects classical Dirichlet-multinomial conjugacy.

Similar connections can be obtained for priors obtained from completely random measures. Consider in particular the beta process. Here the weights $\{p_i\}$ lie in $(0, 1)$, and thus it is natural to view the beta process as yielding an infinite collection of coin-tossing probabilities. Moreover, given the definition of a completely random measure, we are motivated to toss these coins *independently*. This defines the following hierarchy:

$$\begin{aligned} B &\sim \mathrm{BP}(c, B_0), \\ Z \mid B &\sim \mathrm{BeP}(B), \end{aligned}$$

where $Z = \sum_{k=1}^{\infty} z_k \delta_{\omega_k}$ is a completely random measure known as the *Bernoulli process*. The atoms $\{\omega_k\}$ are the same atoms as in B and the weights $\{z_k\}$ are binary values that are equal to one with probability p_k and equal to zero otherwise.

Returning to the De Finetti conditional independence motif, we can draw repeatedly from the Bernoulli process given an underlying draw from the beta process:

$$B \sim \mathrm{BP}(c, B_0),$$
$$Z_i \mid B \sim \mathrm{BeP}(B), \quad i = 1, \ldots, n. \tag{6.3.4}$$

This defines a binary-valued matrix with n rows and an infinite number of columns. Campbell's theorem can again be invoked to show that we obtain a sparse matrix in which the number of ones in each row is finite with probability one. Viewing the columns as "latent traits" or "features," this matrix can be viewed as a sparse featural representation of objects, or alternatively as a model in which each object is assigned to a subset of classes. Note that this differs from the Dirichlet process, where each object is assigned to a single class.

It is also possible to define this probability law directly via a sequential process that is known as the *Indian buffet process (IBP)* (Griffiths and Ghahramani, 2006). In the IBP, customers enter a restaurant sequentially and select dishes in the buffet line. Dishes that have been chosen previously by other customers are selected with probability proportional to the number of times they have been selected by the previous customers. Each customer also selects a random number of new dishes according to a Poisson distribution (with decreasing rate). As shown by Thibaux and Jordan (2007), this probability law can be obtained by marginalizing out the beta process in the hierarchy in (6.3.4). Their argument is the analog of the derivation of the CRP from the DP. In particular, as alluded to above, the CRP can be derived as a simple consequence of the fact that a posterior DP is itself a DP (Ferguson, 1973). A similar conjugacy relationship holds for the beta process—a posterior BP is itself a BP (Kim, 1999). The posterior BP contains atoms in its base measure, and these are necessarily present in any subsequent draw from the BP. Indeed, these act like independent coins relative to the rest of the random measure. Posterior updating of the probabilities associated with these coins is classical beta-Bernoulli updating, which is the rule underlying the IBP.

6.3.3 Hierarchical Models

Bayesian nonparametric models often incorporate classical finite-dimensional parameters—e.g., location parameters, scale parameters, regression parameters and correlation parameters—and it is common to build hierarchies on these parameters. In this section, however, we wish to consider a more thoroughgoing form of Bayesian hierarchy in which the infinite-dimensional parameters of nonparametric models are also linked via hierarchies. Specifically, we discuss conditional independence hierarchies in which a set of completely random measures, $\{G_1, G_2, \ldots, G_M\}$, are conditionally independent given a base measure G_0, and where G_0 is itself a completely random measure.

To see the value of this kind of construction, let us consider the case of the Dirichlet process, and consider M groups of data, $\{x_{1i}\}$, $\{x_{2i}\}$, and $\{x_{Mi}\}$, where each group is to be modeled as a DP mixture. If we have reason to believe that these groups are related, then we may wish to couple the underlying random probability

measures via the following *hierarchical Dirichlet process (HDP)* construction (Teh et al., 2006):

$$\begin{aligned} G_0 &\sim \mathrm{DP}(\gamma, H) \\ G_m \mid G_0 &\sim \mathrm{DP}(\alpha_0, G_0), \quad m = 1, \ldots, M \\ \theta_{mi} \mid G_m &\sim G_m, \quad m = 1, \ldots, M,\ i = 1, \ldots, N_m \\ x_{mi} \mid \theta_{mi} &\sim F_{\theta_{mi}}, \quad m = 1, \ldots, M,\ i = 1, \ldots, N_m, \end{aligned} \tag{6.3.5}$$

where $\{F_\theta\}$ is a parametric family of distributions. The nature of the coupling that is induced by this hierarchy is easily understood by considering a Chinese restaurant representation. Each random measure G_m corresponds to a Chinese restaurant where the atoms forming that random measure correspond to the tables in the restaurant. Some of these tables tend to have large occupancy—these correspond to the atoms with particularly large weights. All of the customers sitting around a single table can be viewed as belonging to a cluster; this is reflected in the fact that the corresponding parameters θ_{mi} are equal to each other. Now if we have reason to believe that the M groups are related, we might expect that a cluster discovered in group m will be useful in modeling the data in the other groups. To achieve this we need to share atoms not only within groups but also between groups. This is achieved by the specification in (6.3.5): the fact that G_0 is drawn from a DP means that it is atomic, and each G_m re-draws from among these atoms. Thus, atoms are shared among the $\{G_m\}$.

Note that it is also possible to couple the random measures $\{G_m\}$ via a classic parametric hierarchy, but this would not generally achieve the goal of sharing clusters among the groups. For example, suppose that G_0 were a parametric distribution depending on a location parameter μ. Bayesian inference for μ would share statistical strength among the groups by centering their base measure at a common location, but, due to the absolutely continuous nature of G_0, the atoms in G_m would be distinct from those in $G_{m'}$ for $m \neq m'$. That is, none of the θ_{mi} would be equal to $\theta_{m'i}$; there would be no sharing of clusters among groups.

The HDP has been used in a wide variety of applications, including social network analysis (Airoldi et al., 2008), genetic analysis (Xing, Jordan, and Sharan, 2007), computational vision (Sudderth, 2006; Kivinen, Sudderth, and Jordan, 2007), natural language parsing (Liang, Jordan, and Klein, 2010), information retrieval (Cowans, 2004), and music segmentation (Ren, Dunson, and Carin, 2008).

The sharing of atoms achieved via the hierarchical nature of the HDP is also useful for the beta process and other members of the family of completely random measures. Recall that the beta process can be viewed as providing a featural description of a collection of objects, where the weights $\{p_i\}$ are probabilities of the objects possessing or not possessing a given feature. In the setting of multiple collections of objects, it may be useful to transfer the features discovered in one collection to other collections. As an example of this *hierarchical beta process* construction, Thibaux and Jordan (2007) presented an application to document modeling, where the groups correspond to document corpora. Each document is represented as a binary vector indexed by the vocabulary items, and this probability vector is modeled as a draw

from a Bernoulli process. The sparseness of word occurrences in documents means that it is useful to transfer probabilities between corpora.

Similarly, it is useful to consider hierarchies based on the gamma process, where the data might consist (for example) of counts of items in some open-ended vocabulary and the gamma process encodes the Poisson rates underlying these counts.

6.3.4 Nested Models

Hierarchical models provide a way to share atoms among multiple random measures, a useful idea when these measures are viewed as related. It can also be worthwhile, however, to consider the opposite end of the spectrum, where atoms are separated into different, non-interacting groups. From a modeling point of view, this allows complex models to be built of simpler components. There is also a computational argument. When atoms are shared, the inferential problem of computing the posterior distribution becomes a highly coupled problem in which each data point has an inferential impact on each atom. Such highly coupled problems can be difficult to solve numerically; in particular, in the MCMC setting the correlations introduced by the coupling can increase the mixing time of MCMC algorithms. The coupling also has an effect on the difficulty of implementing inference algorithms; in particular, it makes it difficult to use divide-and-conquer strategies.

Nesting is a general strategy for building complex models out of simpler components. To illustrate, let us consider a nested version of the Chinese restaurant process (Blei, Griffiths, and Jordan, 2010). In the *nested CRP*, the restaurant metaphor is extended to a set of restaurants organized according to a branching structure, where individuals partake of a sequence of dinners as they proceed down a path in the tree. All individuals enter a fixed restaurant at the root node of the tree. They select a table according to the usual CRP rule (i.e., they sit at a table with probability proportional to the number of customers who have previously selected the table). The table also has a card on it giving the address of a restaurant where the customers will eat the following night. This construction recurses, yielding an infinitely-branching tree where each customer follows a particular path down the tree. After n customers have entered the tree, there will be up to n paths selected in the tree, with some paths selected by multiple customers. The depth of the tree can be fixed and finite, or it can be infinite.

The nested CRP defines a sequence of distributions on partitions, one for each level of the tree. To turn this into a random measure, we introduce atoms drawn from a base measure G_0 at the tables in the restaurants (one atom per table). One possible option is to consider a tree of fixed depth and to place atoms only at the tables in the restaurants at the leaves of the tree, but it can also be useful to place atoms throughout the tree. In either case, the construction separates atoms into distinct groups. In particular, having selected a branch at the root node, only the atoms in the clade below that branch are available.

This nested construction can also be expressed using random measures directly. In particular, consider the following specification of a two-level *nested Dirichlet process (nDP)* (Rodríguez, Dunson, and Gelfand, 2008):

$$G \sim \sum_{k=1}^{\infty} \pi_k^* \delta_{G_k^*},$$
$$G_k^* = \sum_{j=1}^{\infty} \pi_{kj} \delta_{\theta_{kj}}, \tag{6.3.6}$$

where the weights $\{\pi_{kj}\}$ and $\{\pi_k^*\}$ are obtained as in (6.3.3). We see that G is a draw from a DP that selects among an infinite set of components $\{G_k^*\}$, where each component is itself a DP. Note that the atoms associated with the lower-level DPs are distinct (assuming a continuous base measure). From the point of view of the nCRP formalism the specification in (6.3.6) corresponds to a two-level tree in which atoms are associated only with the tables at the leaves of the tree. The top-level restaurant implements the choice among the G_k and each G_k corresponds to a restaurant at the second level of the tree. More generally, the nDP can be extended to an arbitrary number of levels, and a K-level nCRP is obtained by integrating out the Dirichlet processes in a K-level nDP.

Rodríguez, Dunson, and Gelfand (2008) discussed an application of the two-level nDP to a problem in health care where the goal is to model an outcome variable for the hospitals in the fifty US states. The intra-state distribution is generally multimodal and thus it is natural to use DP mixtures for each state. Moreover, there are inter-state similarities as well, and one approach to capturing these similarities is to cluster the states. This is done at the higher level of the nDP by allowing similar states to select the same low-level DP. Note the difference between this approach and an HDP-based approach, where all atoms would be shared among all of the groups; here, atoms are shared only when states fall in the same cluster.

There are also natural applications of infinite-level nested models. Blei, Griffiths, and Jordan (2010) presented an application of the nCRP to the problem of discovering topics in document collections. A topic is defined to be a probability distribution across the words in some vocabulary. The nCRP is augmented to place a topic at each table at every restaurant in the tree. The generation of the words in a document is modeled as follows. The first step is to select a path down the (infinite) tree according to the nCRP. Fixing that path, we repeatedly pick a level in the tree using the GEM ("stick-breaking") distribution and pick a word from the topic distribution at that level on the selected path.[2] Given that nodes at the higher levels in the tree tend to be shared across multiple documents (e.g., the root node is shared across all documents), there is statistical pressure to force topics at higher levels in the tree to concentrate on words that are useful across many documents.

[2] The GEM distribution is closely related to the Dirichlet process; the GEM probabilities can be obtained by randomly permuting the weights $\{\pi_k\}$ in the Dirichlet process according to a size-biased permutation.

Topics at lower levels can focus on more specialized words. Thus the model yields an *abstraction hierarchy*.

It is also of interest to consider nesting for random measures other than the DP. In particular, we can define a *nested beta process* in the following way:

$$B \sim \text{BeP}\Big(\sum_{k=1}^{\infty} p_k^* \delta_{B_k^*}\Big),$$
$$B_k^* = \sum_{j=1}^{\infty} p_{kj} \delta_{\theta_{kj}}.$$

This defines a random measure B that is a collection of atoms, each of which is a beta process. Instead of picking a single path down a tree as in the nDP, this definition makes it possible to pick multiple paths down a tree. This construction is quite natural in applications; indeed, in the setting of document modeling it is a natural generalization of *latent Dirichlet allocation (LDA)* (Blei, Ng, and Jordan, 2003). LDA is a topic model that allow documents to range over arbitrary collections of topics. In the nCRP model, topics are restricted to lie along a single path of the tree, differing only in level of abstraction but not in thematic content. A model based on the nested BP would allow a document to range over both thematic content and level of abstraction.

Similarly, it is of interest to consider a *nested gamma process* construction, which could be used to select multiple branches at each level of a tree, each of which is associated with a count or a rate.

6.3.5 Discussion

We have reviewed some recent developments in Bayesian nonparametrics. Our discussion has focused on completely random measures, a broad class of random measures that have simplifying representational and computational properties. Additional random measures can be obtained by normalization of such measures; in particular, the Dirichlet process can be obtained in this way.

The proliferation of parameters (i.e., atoms) in models based on completely random measures calls for organizational principles to control these models in statistically and computationally sound ways. We have focused on two such principles—hierarchy and nesting. While familiar in the parametric setting, these principles have only recently begun to be exploited fully and explicitly in the nonparametric setting. We anticipate many further developments in this vein. We also note that the theory of Bayesian nonparametrics is in an early stage of development, and in particular we have not yet seen theoretical results in which hierarchy and nesting play an explicit role. Given the important role these concepts play in parametric theory, we expect to see analogous theory emerging in nonparametrics. Finally, we note that hierarchy and nesting are applicable to a wide range of Bayesian nonparametric models that

lie outside of the class of completely random measures; indeed, the Pólya tree is one instance of nesting that is well known in Bayesian nonparametrics.

Acknowledgments: I would like to thank Percy Liang, Kurt Miller, Erik Sudderth, and Yee Whye Teh for helpful comments.

Chapter 7
Bayesian Influence and Frequentist Interface

Under the Bayesian paradigm to statistical inference the posterior probability distribution contains in principle all relevant information. All statistical inference can be deduced from the posterior distribution by reporting appropriate summaries. This coherent nature of Bayesian inference can give rise to problems when the implied posterior summaries are unduly sensitive to some detail choices of the model. This chapter discusses summaries and diagnostics that highlight such sensitivity and ways to choose a prior probability model to match some desired (frequentist) summaries of the implied posterior inference.

7.1 Bayesian Influence Methods

Hongtu Zhu, Joseph G. Ibrahim, Hyunsoon Cho, and Niansheng Tang

A formal Bayesian analysis for analyzing data $\mathbf{Y}$ involves the specification of a sampling distribution for $\mathbf{Y}$, denoted by $p(\mathbf{Y}|\theta)$, and a prior distribution for all parameters of interest θ, denoted by $p(\theta)$. The posterior distribution of θ given $\mathbf{Y}$ is given by

$$p(\theta|\mathbf{Y}) = \frac{p(\theta)p(\mathbf{Y}|\theta)}{\int p(\theta)p(\mathbf{Y}|\theta)d\theta},$$

which forms a basis for carrying out any formal Bayesian analysis. In practice, however, posterior quantities such as the Bayes factor can be sensitive to three kinds of discrepancies: influential observations relative to the fitted model, the prior distribution, and the sampling distribution. There is a large body of literature on developing various Bayesian influence methods for detecting these discrepancies.

Three primary classes of Bayesian influence methods, including Bayesian case influence measures, Bayesian local robustness, and Bayesian global robustness have been proposed to assess the influence of various discrepancies in a Bayesian analy-

M.-H. Chen et al. (eds.), *Frontiers of Statistical Decision Making and Bayesian Analysis*, DOI 10.1007/978-1-4419-6944-6_7,

sis (Kass, Tierney, and Kadane, 1989; McCulloch, 1989; Berger, 1990, 1994; Weiss, 1996; Gustafson, 2000; Sivaganesan, 2000; Oakley and O'Hagan, 2004; Cho, 2009; Cho et al, 2009; Zhu, Ibrahim, and Tang, 2009; Zhu, Ibrahim, and Cho, 2010). Bayesian case influence measures primarily assess the influence of individual observations (or generally, a set of observations) in a Bayesian analysis in order to identify outliers and high leverage points. Considerable research has been devoted to developing single case influence measures for various specific statistical models including generalized linear models, time series models, and survival models (Box and Tiao, 1973; Geisser, 1975, 1993; Johnson and Geisser, 1983; Johnson, 1985; Pettit, 1986; Chaloner and Brant, 1988; Kass, Tierney, and Kadane, 1989; McCulloch, 1989; Carlin and Polson, 1991a; Gelfand, Dey, and Chang, 1992; Weiss and Cook, 1992; Blyth, 1994; Peng and Dey, 1995; Weiss, 1996; Bradlow and Zaslavsky, 1997; Christensen, 1997). The influence of individual observations is often assessed either on the posterior distributions or the predictive distributions through case deletion. The two most popular Bayesian case influence measures are the Conditional Predictive Ordinate (CPO) (Gelfand, Dey, and Chang, 1992; Geisser, 1993) and the Kullback-Leibler (KL) divergence (Peng and Dey, 1995). For instance, Pettit (1986) uses the KL divergence in detecting influential observations in his review of Bayesian diagnostics, whereas Weiss (1996) and Weiss and Cho (1998) assess the influence of deleting a single case as well as establishing its relationship to the KL divergence and CPO. Zhu et al. (2010) introduce three types of Bayesian case influence measures based on case deletion, namely ϕ*-divergence*, *Cook's posterior mode distance*, and *Cook's posterior mean distance*, and then evaluate the effects of deleting a set of observations on these three measures in general parametric models. Another class of simple Bayesian case influence measures is to check whether the posterior distribution of a checking function, which is a function of $\mathbf{Y}$ and/or θ, is far from an appropriate measure of center (Zellner, 1975; Chaloner and Brant, 1988; Meng, 1994; Gelman, Meng, and Stern, 1996).

The global robustness (Berger, 1984, 1990, 1994) approach is based on computing a range of posterior quantities in a perturbation set based on perturbations to $\mathbf{Y}$, $p(\theta)$ and/or $p(\mathbf{Y}|\theta)$ (Basu, 1999; Moreno, 2000; Sivaganesan, 2000). Global influence methods, however, are generally computationally intensive even in relatively simple models with several parameters (Berger, Ríos Insua, and Ruggeri, 2000; Sivaganesan, 2000). For a given posterior quantity $h(\theta)$, one may conclude that a small range indicates robustness. A serious issue associated with examining the range is that the value of the range itself might be misleading because it strongly depends on the appropriate scale chosen for $h(\theta)$ and the size of the perturbation set. For instance, for any given set and posterior quantity $h(\theta)$, the range of $E[ch(\theta)|\mathbf{Y}]$ is c times the range of $E[h(\theta)|\mathbf{Y}]$, where c is any nonzero scalar. To address this issue, several scaled versions of the range including a relative sensitivity have been introduced (Sivaganesan, 1991; Gustafson, 1994; Boratinska, 1996; Ruggeri and Sivaganesan, 2000). By utilizing the concept of differential geometry, Zhu, Ibrahim, and Tang (2009) propose several global influence measures to quantify the degree of various perturbations to statistical models.

The local robustness approach primarily computes the derivatives of posterior quantities with respect to a small deviation from $\mathbf{Y}$, $p(\theta)$ and/or $p(\mathbf{Y}|\theta)$ (Berger, 1994; Gustafson and Wasserman, 1995; Shi and Wei, 1995; Gustafson, 1996a; Berger, Ríos Insua, and Ruggeri, 2000; Perez, Martín, and Rufo, 2006; Zhu, Ibrahim, and Tang, 2009). The local robustness approach has strong connections with Cook's (1986) seminal local influence approach for perturbing $p(\mathbf{Y}|\theta)$ for detecting influential observations and assessing model misspecifications (Tsai and Wu, 1992; Poon and Poon, 1999; Zhu and Lee, 2001; Zhu et al., 2003, 2007). McCulloch (1989) generalizes Cook's (1986) local influence to assess the effects of perturbing the prior in a Bayesian analysis. Further research has been developed for carrying out local robustness utilizing the Frechét derivative of the posterior with respect to the prior (Berger, 1994; Gustafson and Wasserman, 1995; Gustafson, 1996a; Berger, Ríos Insua, and Ruggeri, 2000). Clarke and Gustafson (1998) consider simultaneously perturbing $(\mathbf{Y}, p(\theta), p(\mathbf{Y}|\theta))$ in the context of independent and identically distributed data. Zhu, Ibrahim, and Tang (2009) propose several local influence measures to quantify the degree of simultaneously perturbing $(\mathbf{Y}, p(\theta), p(\mathbf{Y}|\theta))$ for a large class of statistical models, while accounting for missing data.

The rest of this section is organized as follows. In Section 7.1.1, we give a thorough review of Bayesian case influence measures. In Section 7.1.2, we provide an overview of Bayesian global and local robustness approaches. In Section 7.1.3, we illustrate the existing methodologies with a real data set.

7.1.1 Bayesian Case Influence Measures

Various Bayesian case influence measures have been proposed to identify outliers and influential points, but we primarily review four major types, including the posterior probability of outlying sets, posterior outlier statistics, predictive diagnostics, and posterior diagnostics. The first type of Bayesian case influence measures includes Box and Tiao's (1968) approach and its extensions. The key idea of Box and Tiao (1968) is to calculate the posterior probability of an event A_S that a set of observations, denoted by $\mathbf{Y}_S$, is an outlying set, where S denotes a subset of all observations. This approach requires at least three strong assumptions as follows. The first one is to specify the prior probability of A_S such that

$$P(A_S) \geq 0 \text{ and } \sum_S P(A_S) = 1.$$

The second assumption is to specify an explicit sampling distribution of $\mathbf{Y}$ given the fact that A_S is true, denoted by $p(\mathbf{Y}|A_S, \theta)$, for each A_S. The third assumption is that when A_S is true,

$$p(\mathbf{Y}_S|A_S,\theta) = \int p(\mathbf{Y}|A_S,\theta)d\mathbf{Y}_{[S]} \neq p(\mathbf{Y}_S|\theta) = \int p(\mathbf{Y}|\emptyset,\theta)d\mathbf{Y}_{[S]},$$
$$p(\mathbf{Y}_{[S]}|A_S,\theta) = \int p(\mathbf{Y}|A_S,\theta)d\mathbf{Y}_S = p(\mathbf{Y}_{[S]}|\theta) = \int p(\mathbf{Y}|\emptyset,\theta)d\mathbf{Y}_S,$$

where $\emptyset$ denotes the empty set, $\mathbf{Y}_S$ denotes a subsample of $\mathbf{Y}$ consisting of all the observations in S and $\mathbf{Y}_{[S]}$ denotes a subsample of $\mathbf{Y}$ with all observations in S (or $\mathbf{Y}_S$) deleted. Finally, for each set A_S, we calculate the posterior probability of A_S given $\mathbf{Y}$ as follows:

$$P(A_S|\mathbf{Y}) = \frac{p(\mathbf{Y}|A_S)P(A_S)}{\sum_S p(\mathbf{Y}|A_S)P(A_S)},$$

where $p(\mathbf{Y}|A_S) = \int p(\mathbf{Y}|A_S,\theta)p(\theta)d\theta$ is the marginal distribution of $\mathbf{Y}$ under A_S. If $P(A_S|\mathbf{Y})$ is relatively large for a specific set S, then it is fair to claim that all observations $\mathbf{Y}_S$ are outlying observations.

Example 7.1. Consider a linear model $y_i = \mathbf{x}_i'\beta + \varepsilon_i$ for $i = 1,\cdots,n$, where $\mathbf{x}_i$ and β are $p \times 1$ vectors. For a set S, it is assumed that $\varepsilon_i \sim N(0,\sigma^2)$ for $i \notin S$ and $\varepsilon_i \sim N(0,\omega^2\sigma^2)$ otherwise. Let α be the prior probability that any $(\mathbf{x}_i, y_i)$ is an outlier. Thus, $P(A_S) = \alpha^k(1-\alpha)^{n-k}$, where k is the size of S. We consider a Normal-Gamma prior for (β,σ^{-2}). Let $\mathbf{X} = (\mathbf{x}_1',\ldots,\mathbf{x}_n')'$ and $\mathbf{X}_S$ contain all $\mathbf{x}_i$ for $i \in S$ and $\mathbf{X}_{[S]}$ is a submatrix of $\mathbf{X}$ with all observations $\mathbf{x}_i$ with $i \in S$ deleted. After some calculations, it has been shown in Box and Tiao (1968) that $p(A_S|\mathbf{Y})$ is given by

$$C\left(\frac{\alpha}{1-\alpha}\right)^k \omega^{-k}|\mathbf{X}'\mathbf{X}|^{1/2}|\mathbf{X}'\mathbf{X} - (1-\omega^{-2})\mathbf{X}_S'\mathbf{X}_S|^{-1/2}\left(\frac{\mathrm{RSS}_\omega(S)}{\mathrm{RSS}}\right)^{-(n-p)/2},$$

where C is a constant, $\mathrm{RSS} = (\mathbf{Y}-\mathbf{X}\beta)'(\mathbf{Y}-\mathbf{X}\beta)$ and

$$\begin{aligned}\mathrm{RSS}_\omega(S) = {} & (\mathbf{Y}_{[S]} - \mathbf{X}_{[S]}\hat{\beta}_\omega(S))'(\mathbf{Y}_{[S]} - \mathbf{X}_{[S]}\hat{\beta}_\omega(S)) \\ & + \omega^{-2}(\mathbf{Y}_S - \mathbf{X}_S\hat{\beta}_\omega(S))'(\mathbf{Y}_S - \mathbf{X}_S\hat{\beta}_\omega(S)),\end{aligned}$$

in which $\hat{\beta}_\omega(S) = (\mathbf{X}_{[S]}'\mathbf{X}_{[S]} + \omega^{-2}\mathbf{X}_S'\mathbf{X}_S)^{-1}(\mathbf{X}_{[S]}'\mathbf{Y}_{[S]} + \omega^{-2}\mathbf{X}_S'\mathbf{Y}_S)$.

While Box and Tiao's approach is conceptually simple, it is difficult to implement in many scenarios. The major difficulty is to specify an explicit distribution $p(\mathbf{Y}|\mathbf{A}_\mathbf{S},\theta)$ for characterizing the outlying observations in A_S. Another difficulty is to compute the marginal distribution $p(\mathbf{Y}|A_S)$ for a large class of models for missing data, such as models with missing covariates and generalized linear mixed models with/without missing data. Yet another difficulty is that the computation of $p(A_S|\mathbf{Y})$ is numerically infeasible given the large number of all possible events A_S. Given these difficulties, this approach is limited to simple models, such as linear regression with conjugate prior distributions and some models for time series (Abraham and Box, 1979).

The second type of Bayesian case influence measures include Chaloner and Brant's (1988) posterior residual approach and its extensions. The original idea of Chaloner and Brant (1988) is to use the posterior distribution of the realized error

terms for residual analysis to define outliers and calculate posterior probabilities of observations being outliers in a linear model (Zellner, 1975; Zellner and Moulton, 1985). This method has been further extended to generalized linear models, survival models, latent variable models, state space models, and many others (Chaloner, 1991; Albert and Chib, 1993; Lee and Zhu, 2000). The key idea is summarized as follows. One first specifies an objective function $g(\mathbf{Y}_S, \theta)$, which may be a function of the data and unknown parameters, and then calculates the posterior probability of the event $E_S = \{g(\mathbf{Y}_S, \theta) > K_g\}$, called the posterior outlier statistic, which is given by

$$P(E_S|\mathbf{Y}) = \int \mathbf{1}(g(\mathbf{Y}_S, \theta) > K_g) p(\theta|\mathbf{Y}) d\theta,$$

where K_g is a prespecified constant determined by the model assumptions and $\mathbf{1}(\cdot)$ is an indicator function of an event. Moreover, one may allow for missing data in $g(\mathbf{Y}_S, \theta)$. If one considers all observations $\mathbf{Y}$ and sets $K_g = g(\mathbf{Y}^{rep}, \theta)$, then one obtains the *posterior predictive p−value*, which was formally introduced by Meng (1994) and Gelman, Meng, and Stern (1996). The posterior predictive $p-$value may be used to assess the plausibility of a proposed statistical model. If $P(E_S|\mathbf{Y})$ is relatively small for a given K_g, then it is fair to claim that all observations in $\mathbf{Y}_S$ are outlying observations.

Example 7.2. Consider a linear model $y_i = \mathbf{x}_i'\beta + \varepsilon_i$ for $i = 1, \cdots, n$, where $\mathbf{x}_i$ and β are $p \times 1$ vectors and $\varepsilon_i \sim N(0, \sigma^2)$ for all i. We consider a Normal-Gamma prior for (β, σ^{-2}), that is $\beta|\sigma^{-2} \sim N(\mathbf{m}_0, \sigma^2\Sigma_0^{-1})$ and $\sigma^{-2} \sim \Gamma(\alpha_0, \alpha_1)$. It has been shown in Chaloner and Brant (1988) that $\varepsilon_i|\mathbf{Y} \sim t(e_i, a_1 b_1 h_{i,i}^{-1}, 2a_1)$, where $e_i = y_i - \mathbf{x}_i'\mathbf{m}_1$, $\mathbf{m}_1 = (\Sigma_0 + \mathbf{X}'\mathbf{X})^{-1}(\Sigma_0\mathbf{m}_0 + \mathbf{X}'\mathbf{Y})$, $h_{i,i}$ is the (i,i)th element of $H = \mathbf{X}(\Sigma_0 + \mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, $a_1 = \alpha_0 + n/2$, and $b_1 = \alpha_1 + 0.5[(\mathbf{Y} - \mathbf{X}\mathbf{m}_1)'\mathbf{Y} + (\mathbf{m}_0 - \mathbf{m}_1)'\Sigma_0\mathbf{m}_0]$. Thus, it can be shown that the posterior probability that $|\varepsilon_i| > m\sigma$ is given by

$$P\{|\varepsilon_i| > m\sigma|\mathbf{Y}\} = \int_0^\infty [1 - \Phi(z_1) + \Phi(z_2)] p(\tau|\mathbf{Y}) d\tau,$$

where $\tau = \sigma^{-2}$, $\Phi(\cdot)$ is the distribution function of the standard normal distribution, $z_1 = (m - \tilde{e}_i\tau^{1/2})h_{i,i}^{-1/2}$, and $z_2 = (-m - \tilde{e}_i\tau^{1/2})h_{i,i}^{-1/2}$.

The second class of case influence measures has its merits and limitations. Its merit is that it is computationally straightforward even for very complicated statistical models, such as hierarchical models with missing response and/or covariate data. Specifically, the posterior outlier statistic can be directly estimated by using samples drawn from Markov chain Monte Carlo methods (MCMC) (Chen, Shao, and Ibrahim, 2000). However, because $P(E_S|\mathbf{Y})$ and its associated posterior predictive $p-$value were criticized for double use of the data, they do not have 'standard' probability interpretations, that is, the uniform scale is not appropriate (Bayarri and Berger, 2000; Robins, van der Vaart, and Ventura, 2000; de la Horra and Rodriguez-Bernal, 2001; Hjort, Dahl, and Steinbakk, 2006). Hjort, Dahl, and Steinbakk (2006) propose a calibration of the posterior predictive p (ppp) values, which results in calibrated ppp values, while the calibrated ppp values are uniform in [0, 1] under a correct model specification. Computing the calibrated ppp values, however, gener-

ally involves in a double-simulation scheme and can be computationally intensive. Further development of the calibrated ppp values for hierarchical models is given in Steinbakk and Storvik (2009). Instead of calibrating ppp values, several *measures of surprise*, such as the $p-$value and the relative predictive surprise, have been proposed for outlier detection and model assessment in some relatively simple models (Bayarri and Berger, 2000; Bayarri and Morales, 2003; Bayarri and Castellanos, 2007).

The third type of Bayesian case influence measures include all predictive diagnostics (Box, 1980; Geisser, 1993; Gelfand and Dey, 1994; Sinha and Dey, 1997). The key idea of predictive diagnostics is to assess the discordancy of a set of observation $\mathbf{Y}_S$ based on the predictive distribution of $\mathbf{Y}_S$ using

$$E[g(\mathbf{Y}_S, \mathbf{Y}_S^*)|\mathbf{Y}_{[S]}] = \int g(\mathbf{Y}_S, \mathbf{Y}_S^*) p(\mathbf{Y}_S^*|\mathbf{Y}_{[S]}) d\mathbf{Y}_S^*,$$

where $g(\mathbf{Y}_S, \mathbf{Y}_S^*)$ is a checking function and $\mathbf{Y}_S^*$ is a future value of $\mathbf{Y}_S$. A particular predictive diagnostic is

$$P(\mathbf{Y}_S^* \in R_S|\mathbf{Y}_{[S]}) = \int \mathbf{1}(\mathbf{Y}_S^* \in R_S) p(\mathbf{Y}_S^*|\mathbf{Y}_{[S]}) d\mathbf{Y}_S^*,$$

where $p(\mathbf{Y}_S^*|\mathbf{Y}_{[S]})$ is the predictive distribution of $\mathbf{Y}_S^*$ given $\mathbf{Y}_{[S]}$ given by

$$p(\mathbf{Y}_S^*|\mathbf{Y}_{[S]}) = \frac{p(\mathbf{Y}_S^*, \mathbf{Y}_{[S]})}{p(\mathbf{Y}_{[S]})} = \frac{1}{\int [p(\mathbf{Y}_S^*|\mathbf{Y}_{[S]}, \theta)]^{-1} p(\theta|\mathbf{Y}_S^*, \mathbf{Y}_{[S]}) d\theta}.$$

Moreover, the region R_S can be a region around either the predictive mean, mode, or median of $p(\mathbf{Y}_S|\mathbf{Y}_{[S]})$, or $\{T_S(\mathbf{Y}_S^*) \geq T_S(\mathbf{Y}_S)\}$, in which $T_S(\cdot)$ is a scalar function. By setting $g(\mathbf{Y}_S, \mathbf{Y}_S^*) = \mathbf{1}(\mathbf{Y}_S : ||\mathbf{Y}_S - \mathbf{Y}_S^*|| \leq \varepsilon)/V(\varepsilon, S)$, the conditional predictive ordinate statistic $p(\mathbf{Y}_S|\mathbf{Y}_{[S]})$ for the set S, denoted by CPO_S, is the limit of $E[g(\mathbf{Y}_S, \mathbf{Y}_S^*)|\mathbf{Y}_{[S]}]$ as $\varepsilon \to 0$, where $V(\varepsilon, S)$ is the volume of a sphere with radius ε in $[\dim(\mathbf{Y}_S) + 1]-$dimensional Euclidean space (Geisser, 1993; Gelfand and Dey, 1994; Sinha and Dey, 1997; Chen, Shao, and Ibrahim, 2000). If CPO_S is relatively small for a specific set S, then one could reject that $\mathbf{Y}_S$ is concordant with $\mathbf{Y}_{[S]}$. These predictive diagnostics have been used to create model selection criteria including pseudo-Bayes factors and posterior Bayes factors for model selection (Aitkin, 1991; Geisser, 1993; Berger and Pericchi, 1996a; Gelfand and Ghosh, 1998; Ibrahim, Chen, and Sinha, 2001; Chen, Dey, and Ibrahim, 2004).

Example 7.3. Consider an independent and identically distributed sample $Y_1, \cdots, Y_n$ from $p(y|\theta) = \theta \exp(-\theta x)$. Suppose that $p(\theta) \propto \theta^{N_0 - 1} \exp(-\theta N_0 \bar{y}_0)$. First, by setting $S = \{i\}$ and $g(Y_i, \mathbf{Y}_i^*) = \mathbf{1}(\mathbf{Y}_i^* \geq \mathbf{Y}_i)$, we have

$$P_i = E[\mathbf{1}(\mathbf{Y}_i^* \geq \mathbf{Y}_i)|\mathbf{Y}_{[i]}] = \left(\frac{N_0 \bar{y}_0 + n\bar{y}}{N_0 \bar{y}_0 + n\bar{y} + Y_i} \right)^{n+N_0}.$$

The predictive diagnostic P_i is conceptually simple, but it can be computationally difficult for complicated statistical models, such as multilevel models with missing data. Specifically, except for very simple models, approximating $P(\mathbf{Y}_S^* \in R_S|\mathbf{Y}_{[S]})$ involves drawing samples using MCMC methods and the approximation of $\int [p(\mathbf{Y}_S^*|\mathbf{Y}_{[S]},\theta)]^{-1} p(\theta|\mathbf{Y}_S^*,\mathbf{Y}_{[S]})d\theta$ in the denominator is typically numerically unstable (Chen, Shao, and Ibrahim, 2000). Moreover, to accurately approximate $P(\mathbf{Y}_S^* \in R_S|\mathbf{Y}_{[S]})$, we must draw many samples from $p(\theta|\mathbf{Y}_S^*,\mathbf{Y}_{[S]})$ for a large number of different $\mathbf{Y}_S^*$. For complex models, however, it is very challenging to obtain a suitable and accurate approximation of these model selection criteria based on the predictive distribution.

The fourth type of Bayesian case influence measures include all posterior diagnostics (Csiszár, 1967; Johnson and Geisser, 1985; Pettit and Smith, 1985; Pettit, 1986; Weiss and Cook, 1992; Weiss, 1996). In contrast to predictive diagnostics, posterior diagnostics compare the posterior distributions of θ given the complete data $\mathbf{Y}$ and the reduced data $\mathbf{Y}_{[S]}$. A well-known example is the $\phi-$influence of $\mathbf{Y}_{[S]}$, defined by

$$\mathrm{D}_\phi(S) = \int \phi(R_{[S]}(\theta)) p(\theta|Y) d\theta,$$

where $R_{[S]}(\theta) = p(\theta|Y_{[S]})/p(\theta|Y)$ and $\phi(\cdot)$ is a convex function with $\phi(1) = 0$ (Weiss and Cook, 1992; Weiss, 1996). It has been shown in Cho (2009) and Zhu, Ibrahim, and Tang (2009) that $\mathrm{D}_\phi(S)$ can be approximated by *Cook's posterior mode distance*, denoted by $\mathrm{CP}(S)$, and *Cook's posterior mean distance*, denoted by $\mathrm{CM}(S)$. A large value of $\mathrm{D}_\phi(S)$ indicates that S contains an influential set of observations. These posterior diagnostics can also be used to form model selection criteria, which have strong connections with the Deviance Information Criterion (Spiegelhalter et al., 2002; Cho, 2009; Zhu, Ibrahim, and Tang, 2009).

Similar to predictive diagnostics, posterior diagnostics are also conceptually simple, but can be computationally difficult for complicated statistical models, particularly, hierarchical models with missing data. Let $p_S(\theta)$ be $p(\mathbf{Y}_S|\mathbf{Y}_{[S]},\theta)$. Then, we have $p(\theta|Y_{[S]}) = [p_S(\theta)]^{-1} p(\mathbf{Y}|\theta)p(\theta)/\int [p_S(\theta)]^{-1} p(\mathbf{Y}|\theta)p(\theta)d\theta$ and the computational formula for $\mathrm{d}_\phi(S)$ can be obtained as

$$\mathrm{d}_\phi(S) = E_{\theta|Y}\left[\phi\left(\frac{[p_S(\theta)]^{-1}}{E_{\theta|Y}\{[p_S(\theta)]^{-1}\}}\right)\right],$$

where $E_{\theta|Y}$ denotes the expectation taken with respect to the posterior distribution $p(\theta|\mathbf{Y})$. It is challenging to accurately calculate $\mathrm{d}_\phi(S)$ for all possible sets S. It has been shown that $\mathrm{D}_\phi(S)$ can be approximated by a quadratic form in $\partial_{\boldsymbol{\theta}} \log p_S(E[\theta|\mathbf{Y}])$ even for S with a relatively large number of observations (Cho, 2009; Zhu, Ibrahim, and Tang, 2009). This result provides a one-step approximation to $\mathrm{D}_\phi(S)$, which is computationally simple.

Example 7.4. We consider the normal linear model as $y_i = \mathbf{x}_i'\beta + \varepsilon_i$, where β is $p \times 1$ parameter vector, $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_n)' \sim N_n(0, \tau^{-1}\mathrm{I})$, and $\tau = 1/\sigma^2$ is assumed known. Thus, we have $\mathbf{Y}|\beta,\tau \sim N_n(\mathbf{X}\beta, \tau^{-1}\mathrm{I})$, where $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$ and $\mathbf{Y} =$

$(y_1,\ldots,y_n)'$. We consider a conjugate prior for β as $N_p(\mu_0,\tau^{-1}\Sigma_0)$. The posterior distributions of β based on the full data, $p(\beta|\mathbf{Y})$, and with the ith case deleted, $p(\beta|\mathbf{Y}_{[i]})$, are given by

$$\beta|\mathbf{Y}\sim N_p(\tilde{\beta},\ \tau^{-1}(\mathbf{X}'\mathbf{X}+\Sigma_0^{-1})^{-1})\ \text{ and }\ \beta|\mathbf{Y}_{[i]}\sim N_p(\tilde{\beta}_{[i]},\ \tau^{-1}(\mathbf{X}'_{[i]}\mathbf{X}_{[i]}+\Sigma_0^{-1})^{-1}),$$

where $\tilde{\beta}=(\mathbf{X}'\mathbf{X}+\Sigma_0^{-1})^{-1}(\mathbf{X}'\mathbf{Y}+\Sigma_0^{-1}\mu_0)$, $\tilde{\beta}_{[i]}=(\mathbf{X}'_{[i]}\mathbf{X}_{[i]}+\Sigma_0^{-1})^{-1}(\mathbf{X}'_{[i]}\mathbf{Y}_{[i]}+\Sigma^{-1}\mu_0)$, $\mathbf{X}_{[i]}$ is $\mathbf{X}$ with $\mathbf{x}_i'$ deleted, and $\mathbf{Y}_{[i]}$ is $\mathbf{Y}$ with y_i deleted. Note that $\mathbf{X}'_{[i]}\mathbf{X}_{[i]}=\mathbf{X}'\mathbf{X}-\mathbf{x}_i\mathbf{x}_i'$ and $\mathbf{X}'_{[i]}\mathbf{Y}_{[i]}=\mathbf{X}'\mathbf{Y}-\mathbf{x}_i\mathbf{y}_i$.

For the K-L divergence, we can get an exact theoretical form for $\mathrm{d}_\phi(i)$ by applying the formula for the K-L divergence between two normal distributions (Cook and Weisberg, 1982) as

$$\begin{aligned}\mathrm{d}_\phi(i)=0.5[&\tau(\tilde{\beta}-\tilde{\beta}_{[i]})'(\mathbf{X}'\mathbf{X}+\Sigma_0^{-1})(\tilde{\beta}-\tilde{\beta}_{[i]})-\tau(\tilde{\beta}-\tilde{\beta}_{[i]})'(\mathbf{x}_i\mathbf{x}_i')(\tilde{\beta}-\tilde{\beta}_{[i]})\\&-\log|1-\mathbf{x}_i'(\mathbf{X}'\mathbf{X}+\Sigma_0^{-1})^{-1}\mathbf{x}_i|-\mathrm{tr}\{\mathbf{x}_i'(\mathbf{X}'\mathbf{X}+\Sigma_0^{-1})^{-1}\mathbf{x}_i\}].\end{aligned}$$

Let $\mathbf{Q}=\mathbf{X}(\mathbf{X}'\mathbf{X}+\Sigma_0^{-1})^{-1}\mathbf{X}'$ with ith diagonal element $q_{ii}=\mathbf{x}_i'(\mathbf{X}'\mathbf{X}+\Sigma_0^{-1})^{-1}\mathbf{x}_i$. Then we have

$$\tilde{\beta}-\tilde{\beta}_{[i]}=\frac{1}{1-q_{ii}}(\mathbf{X}'\mathbf{X}+\Sigma_0^{-1})^{-1}\mathbf{x}_i(y_i-\mathbf{x}_i'\tilde{\beta})$$

and therefore, we obtain

$$\mathrm{d}_\phi(i)=0.5\left\{\tau\cdot\frac{q_{ii}}{1-q_{ii}}(y_i-\mathbf{x}_i'\tilde{\beta})'(y_i-\mathbf{x}_i'\tilde{\beta})-\log(1-q_{ii})-q_{ii}\right\}.$$

7.1.2 Bayesian Global and Local Robustness

Bayesian global and local robustness methods assess the degree of sensitivity of the results obtained from a Bayesian analysis to the data, the prior, and the sampling distribution (Sivaganesan, 1991; Berger, 1994; Gustafson, 1994; Gustafson and Wasserman, 1995; Shi and Wei, 1995; Boratinska, 1996; Gustafson, 1996a; Basu, 1999; Berger, Ríos Insua, and Ruggeri, 2000; Moreno, 2000; Ruggeri and Sivaganesan, 2000; Sivaganesan, 2000; Perez, Martín, and Rufo, 2006; Zhu, Ibrahim, and Tang, 2009). Bayesian robustness methods are sequentially executed in two steps: defining a perturbation class, denoted by Γ, which includes a class of posterior densities representing the uncertainty in the prior and sampling distribution, and computing global or local influence measures based on Γ. For instance, for a given Γ and a target quantity $h(\theta)$, the most commonly used measure in Bayesian global robustness is the range of a posterior quantity given by

$$R_h=\overline{h}-\underline{h}=\max_{p(\boldsymbol{\theta}|\mathbf{Y})\in\mathscr{F}}E[h(\theta)|\mathbf{Y}]-\min_{p(\boldsymbol{\theta}|\mathbf{Y})\in\mathscr{F}}E[h(\theta)|\mathbf{Y}],\qquad(7.1.1)$$

where $E[h(\theta)|\mathbf{Y}]$ is a posterior expectation of $h(\theta)$. The value R_h measures the maximum variation caused by the uncertainty in $\mathscr{F}$. If R_h is small, then one may claim that the Bayesian analysis is robust.

7.1.2.1 Perturbation Class

We briefly review different perturbation classes to characterize specific perturbation schemes to the data, the prior or/and sampling distributions. First, we outline various sophisticated perturbation classes of priors in the literature. We refer the reader to Berger (1994) and Ruggeri and Sivaganesan (2000) for a detailed review. Some prior classes include parametric classes (Berger, 1990, 1994), classes with a given shape and smoothness (Bose, 1990, 1994), classes with specified generalized moments (Goutis, 1994; Betro, Meczarski, and Ruggeri, 1994), contamination classes (Huber, 1973; Sivaganesan and Berger, 1989, 1993; Gelfand and Dey, 1991), classes with a topological neighborhood (Fortini and Ruggeri, 1994a,b; Basu, 1995), the density ratio class (DeRobertis and Hartigan, 1981), mixture classes (Bose, 1994), and marginal and independence classes (Lavine, Wasserman, and Wolpert, 1991; Berger and Moreno, 1994). Among them, the parametric classes, the contamination classes, and the mixture classes are most popular, because they are conceptually easy and easily generalizable even for commonly used hierarchical models.

We consider the contamination classes as an example. The mixture contamination classes have the form

$$\Gamma_{CC} = \{p(\theta) : p(\theta) = (1-\varepsilon)p_0(\theta) + \varepsilon g(\theta), g \in \mathscr{G}\},$$

where $\varepsilon \in [0,1]$, $p_0(\theta)$ is a base density, and $\mathscr{G}$ is a class of densities. This class has been widely used in robust statistics (Huber, 1973). Other contamination classes include the linear and nonlinear perturbation classes (Gustafson, 1996a) and the general ε and geometric contamination classes (Perez, Martín, and Rufo, 2006). The contamination classes are easy to elicit and can be easily generalized to hierarchical models with multivariate parameters. There has been extensive research on finding a suitable $\mathscr{G}$ (Sivaganesan, 1988, 1989, 1991; Sivaganesan and Berger, 1989; Moreno and Gonzalez, 1990). For large $\mathscr{G}$, the class Γ can be too large, which is common for high dimensional θ, and thus robustness can be hard to achieve.

Another interesting class is the distribution band class defined as

$$\Gamma_{DBC} = \{p(\theta) : L(\theta) \leq \int_{-\infty}^{\theta} p(x)dx \leq U(\theta)\},$$

where $L(\theta)$ and $U(\theta)$ are, respectively, specified cumulative density functions with $L(\theta) \leq U(\theta)$. This class includes the well known Kolmogorov and Levy neighborhoods (Ruggeri, Ríos Insua, and Martín, 2005). Moreover, it is straightforward to compute the associated influence measures. However, it is hard to choose an appropriate $L(\theta)$ and $U(\theta)$. We refer the reader to Berger (1994) and Ruggeri, Ríos Insua, and Martín (2005) for further details.

We consider two types of sampling distribution classes, including single-case perturbation classes and global perturbation classes. The single-case perturbation classes are to independently perturb individual observations Y_i in order to identify influential observations (Cook, 1986; Zhu and Lee, 2001, 2003; Zhu et al., 2007). We consider an independent-type-incomplete-data (ITID) model defined by $p(\mathbf{Y};\theta) = \prod_{i=1}^{n} p(Y_i;\theta)$, which subsumes most commonly used models, such as generalized linear models with missing responses and/or covariates (Ibrahim, Chen, and Lipsitz, 1999, 2005). In this case, a single-case perturbation class can be defined by

$$\Gamma_{SC} = \{p(\mathbf{Y};\theta,\omega_S) = \prod_{i=1}^{n} p(Y_i;\theta,\omega_{S,i}) : \omega_S \in R^n\},$$

where $\omega_S = (\omega_{S,1},\cdots,\omega_{S,n})$ and $\omega_{S,i}$ denotes the perturbation to the ith observation.

The global perturbation classes are to introduce a global perturbation to the sampling distribution in order to assess the goodness of fit of the sampling distribution. A class of commonly used perturbation densities is defined by

$$\Gamma_{GP} = \{p(\mathbf{Y};\theta)\exp\{\sum_{j=1}^{m}\omega_{G,j}u_j(\mathbf{Y};\theta) - 0.5\sum_{j=1}^{m}\omega_{G,j}^2 u_j(\mathbf{Y};\theta)^2 - C(\theta,\omega_G)\} : \omega_G \in R^m\},$$

where $C(\theta,\omega_G)$ is the normalizing constant, $\omega = (\omega_j)$ is an $m\times 1$ vector, and $u_j(\mathbf{Y};\theta)$ is any scalar function having zero mean under $p(\mathbf{Y};\theta)$. In this case, the $m\times 1$ vector $\omega_G^0(\mathbf{Y},\theta) = (0,\cdots,0)'$ represents no perturbation. The number m can either be as small as one or increase with n (Gustafson, 2000; Troxel, Ma, and Heitjan, 2004; Copas and Eguchi, 2005; Zhu et al., 2007). It is also possible to introduce the additive $\varepsilon-$contamination class and the geometric contamination class for the sampling distribution. For instance, we may consider an additive $\varepsilon-$contamination class defined as

$$\Gamma_{G\varepsilon} = \{p(\mathbf{Y};\theta)^{1-\varepsilon}g(\mathbf{Y};\theta)^{\varepsilon} : g(\mathbf{Y};\theta) \in \mathscr{G}_{\mathbf{Y}}\},$$

where $\mathscr{G}_{\mathbf{Y}}$ is a class of densities of $\mathbf{Y}$ and $\varepsilon \in [0,1]$.

Besides individual perturbations to the prior and the sampling distribution, we can also consider a perturbation class to simultaneously perturb the data, prior and sampling distribution. Based on the joint perturbation class, we may measure the amount of perturbation, the extent to which each component of a perturbation model contributes to, and the degree of orthogonality for the components of the perturbation model. Such a quantification is very useful for rigorously assessing the relative influence of each component in the Bayesian analysis, which can reveal any discrepancy among the data, the prior, or the sampling model. To the best of our knowledge, Clarke and Gustafson (1998) is one of the very few papers on simultaneously perturbing the data, prior and sampling distribution in the context of independent and identically distributed data. Recently, Zhu, Ibrahim, and Tang (2009) propose the simultaneous perturbation class in general parametric models. Specifically,

Zhu, Ibrahim, and Tang (2009) develop a Bayesian perturbation manifold (BPM) for the perturbation class. Suppose that the perturbation class Γ can be written as $\{p(\mathbf{Y},\theta;\omega):\omega\in\Omega\}$, where Ω can be either a finite dimensional set or an infinite dimensional set. We assume that $C(t):p_Y(\omega(t))=p(\mathbf{Y},\theta;\omega(t))$ is a differentiable function mapping from $t\subset I\in R$ to the manifold Γ with $p_Y(\omega(0))=p(\mathbf{Y},\theta;\omega)$, where I is an open interval covering 0. Let $\dot{p}_Y(\omega(t))=dp_Y(\omega(t))/dt$. At each ω, there is a tangent space $T_\omega\Gamma$ of Γ spanned by $\dot{p}_Y(\omega(0))$. The inner product of any two tangent vectors $\mathbf{v}_1(\omega)$ and $\mathbf{v}_2(\omega)$ in $T_\omega\Gamma$ is defined as

$$g(\mathbf{v}_1,\mathbf{v}_2)(\omega)=\int\frac{\mathbf{v}_1(\omega)}{p_Y(\omega)}\frac{\mathbf{v}_2(\omega)}{p_Y(\omega)}p_Y(\omega)d\mathbf{Y}d\theta.$$

Let $\mathbf{u}(\omega)=\mathbf{u}(p_c(\omega))$ and $\mathbf{v}(\omega)=\mathbf{v}(p_c(\omega))$ be two smooth vector fields defined from Γ to $\mathscr{T}\Gamma$. We can define the covariant derivative for the Levi-Civita connection $\nabla_{\mathbf{v}}\mathbf{u}$ and a *geodesic* on the manifold between any two densities $p(\mathbf{Y},\theta;\omega_1)$ and $p(\mathbf{Y},\theta;\omega_2)$. The geodesic is a direct extension of the straight line $\omega(t)=\omega^0+t\mathbf{h}$ in finite dimensional Euclidean space (Amari, 1990; Kass and Vos, 1997). A *Bayesian perturbation manifold* $(\Gamma,g(\mathbf{u},\mathbf{v}),\nabla_{\mathbf{v}}\mathbf{u})$ is the manifold Γ with an inner product $g(\mathbf{u},\mathbf{v})$ and a covariant derivative for the Levi-Civita connection denoted by $\nabla_{\mathbf{v}}\mathbf{u}$. There are several advantages of building the Bayesian perturbation manifold framework for the perturbation class.

(i) It allows one to quantify the 'shortest' distance between any two densities in the class. Then, we can calculate the maxima of the distances between any two densities in the class to determine the size of the perturbation class. In addition, we may measure the amount of perturbation and the extent to which each component of a perturbation model contributes.

(ii) The Bayesian perturbation manifold is relatively simple even for simultaneous perturbations to the prior, the data, and the sampling distribution (Zhu, Ibrahim, and Tang, 2009). Based on the inner product $g(\mathbf{v}_1,\mathbf{v}_2)$ of the Bayesian perturbation manifold for the joint perturbation class, we may measure the degree of orthogonality for the components of the perturbation model. Such a quantification is very useful for rigorously assessing the relative influence of the data, prior and sampling distribution in a Bayesian analysis.

The last issue is how to choose the different perturbation classes. Discussions on the choice of reasonable perturbation classes are given in Berger (1990), Lavine (1991), Pericchi and Walley (1991), Sivaganesan (1991), Walley (1991), Wasserman (1992), and Moreno and Pericchi (1993). The consensus from these papers can be summarized as follows:

(i) The class should be relatively easy to elicit and interpret.

(ii) The class should be relatively easy to compute. Generally, the computational burden is strongly associated with the dimension of the class and the number of elicited features of the prior and sampling distribution.

(iii) The size of the class should be appropriately chosen. For a small class, one may easily conclude robustness compared with a larger class, whereas for a large class, one might conclude that robustness is lacking. Thus, one needs to quantify the size of a specific class by measuring the maximum of the distances between any two densities in the class. Recently, by using concepts in differential geometry, Zhu, Ibrahim, and Tang (2009) suggest computing the geodesic distance between any two densities in the class in order to measure the size of the class.

(iv) The class should be relatively easy and generalizable to a wide range of statistical models. Many prior classes may be hard to use for models with multiple parameters.

7.1.2.2 Global Influence Measures

Various global influence measures have been developed for carrying out global robustness inference (Berger, 1994; Ruggeri, Ríos Insua, and Martín, 2005). The most commonly used measure in Bayesian global robustness is the range of a posterior quantity in (7.1.1). Its value quantifies the degree of variation caused by the uncertainty in the perturbation class. The range itself, however, can be misleading since it depends strongly on two quantities: (i) the size of the perturbation class and (ii) the scale of $h(\theta)$. To address this drawback, several scaled versions of the range have been proposed for the prior perturbation class (Sivaganesan, 1991; Gustafson, 1994; Ruggeri and Wasserman, 1995; Ruggeri and Sivaganesan, 2000). Let $p_0(\theta|\mathbf{Y})$ be the unperturbed (baseline) posterior density. For a given perturbation class Γ, Ruggeri and Sivaganesan (2000) consider a *relative sensitivity* defined by

$$\mathrm{RS}(h,p) = \frac{\{E[h(\theta)|\mathbf{Y},p] - E[h(\theta)|\mathbf{Y},p_0]\}^2}{\mathrm{Var}[h(\theta)|\mathbf{Y},p]},$$

where $E[\cdot|\mathbf{Y},p]$ and $\mathrm{Var}[\cdot|\mathbf{Y},p]$, respectively, denote the conditional expectation and variance with respect to $p(\theta|\mathbf{Y}) \in \Gamma$. The relative sensitivity $\mathrm{RS}(h,p)$ is a linear functional and scale invariant, because $\mathrm{RS}(h,p) = \mathrm{RS}(kh,p)$ for any $k \neq 0$. The relative sensitivity has its limitations. It also depends on the size of Γ because $\mathrm{RS}(h,p)$ does not account for the distance between $p(\theta|\mathbf{Y})$ and the baseline posterior density. In addition, the relative sensitivity is defined only for the linear functional, but not for nonlinear functionals.

Example 7.5: (Ruggeri and Sivaganesan, 2000). We consider a normal prior with the mean in a range. Suppose that Y is distributed as $N(\theta,\sigma^2)$ with $\Gamma = \{N(\mu,\tau^2) : -6 < \mu < 6\}$ and $p_0(y) = N(0,\tau^2)$ in which $\sigma^2 = \tau^2 = 32$. Also let $h(\theta) = \theta$ and $y = 0$. Then, $E[h(\theta)|y,p_0] = 0$, and for $p(y) \in \Gamma$, $E[h(\theta)|Y,p] = \mu/2$ and $\mathrm{Var}[\theta|Y,p] = 16$. Thus, the range $E[h(\theta)|Y,p]$ is (-3,3), which would probably be regarded as large, indicating a lack of robustness. But $\sup \mathrm{RS}(h,p) = 1/4$, which means that, for any $p(y) \in \Gamma$, the value of $E[h(\theta)|Y,p_0]$ is at most one-half of a posterior standard deviation away from $E[h(\theta)|Y,p]$. Hence, the value of $\sup \mathrm{RS}(h,p)$ is likely to be thought of as small, indicating robustness.

Based on the Bayesian perturbation manifold, Zhu, Ibrahim, and Tang (2009) introduce several global influence measures to assess the degree of variation caused by the uncertainty in the perturbation class. Let $p_Y(\omega^0)$ and $p_Y(\omega)$ be, respectively, the baseline and perturbed posterior densities, where ω^0 and ω belong to Ω. Let $C(t) = p_Y(\omega(t)) : [-\delta, \delta] \to \Gamma$ be a smooth curve on Γ joining $p_Y(\omega^0)$ and $p_Y(\omega)$ such that $C(0) = p_Y(\omega^0)$ and $C(1) = p_Y(\omega)$, where $\delta > 1$. We consider a function of the baseline and perturbed posterior distributions as a smooth function of interest $f(\omega) = f(p_Y(\omega)) : \Gamma \to R$ for sensitivity analysis. Some well-known examples include the ϕ-divergence function, Cook's posterior mean distance, or the logarithm of the Bayes factor. Thus, $f(\omega(t)) : [-\delta, \delta] \to R$ is a real function of t. One can assess the global influence of perturbing the posterior distribution based on the objective function $f(\omega(t))$. Specifically, Zhu, Ibrahim, and Tang (2009) introduce the *intrinsic global influence measure* for comparing $p_Y(\omega^0)$ and $p_Y(\omega)$ given by

$$\mathrm{IGI}_f(\omega^0, \omega) = \frac{\{f(\omega) - f(\omega^0)\}^2}{d(\omega^0, \omega)^2},$$

where $d(\omega^0, \omega)$ is the minimal distance between $p_Y(\omega^0)$ and $p_Y(\omega)$. $\mathrm{IGI}_f(\omega^0, \omega)$ can be interpreted as the ratio of the change of the objective function over the minimal distance between $p_Y(\omega^0)$ and $p_Y(\omega)$ on Γ.

The intrinsic global influence measure has its merits and limitations. $\mathrm{IGI}_f(\omega^0, \omega)$ as a sensitivity measure explicitly accounts for the size of a specific perturbation relative to the baseline density, whereas $\mathrm{RS}(h, p)$ does not. The intrinsic global influence measure allows us to directly compare the degree of a specific perturbation relative to ω^0. Moreover, $\mathrm{IGI}_f(\omega^0, \omega)$ is invariant to any diffeomorphism transformation of ω and can be defined for both linear and nonlinear functionals. For large classes, however, it can be computationally challenging to compute $d(\omega^0, \omega)$. The calibration of the intrinsic global influence measure is also a topic of current research.

7.1.2.3 Local Influence Measures

A variety of local influence measures have been developed to assess the effect of minor perturbations to the data, prior and sampling distribution (Cook, 1986; McCulloch, 1989; Birmiwal and Dey, 1993; Ruggeri and Wasserman, 1993; Sivaganesan, 1993; Basu, 1996; Dey, Ghosh, and Lou, 1996; Gustafson, 1996a,b, 2000; Zhu et al., 2007; Zhu, Ibrahim, and Tang, 2009). For instance, for the prior perturbation class, almost all local influence measures involve differentiating a posterior quantity with respect to a changing prior and then evaluating it at a *baseline* prior. Diaconis and Freedman (1986) used Fréchet derivatives with respect to the prior distribution and proposed the use of the norm of the derivative as a sensitivity measure. Gateaux derivatives are considered in Srinivasan and Truszaczynska (1990), Ruggeri and Wasserman (1995), Dey, Ghosh, and Lou (1996), Sivaganesan (2000), Perez, Martín, and Rufo (2006), and among others. These measures differ from each

other in the ways that the prior distributions are perturbed, and that either the posterior distribution or a scalar posterior summary is used.

As an illustration, we consider the Fréchet derivative of $E[h(\theta)|Y,p]$ with respect to the prior $p(\theta)$. For $p(\theta)$, we perturb it with a signed measure $\delta(\theta)$ with zero mass and then obtain a posterior distribution $p_\delta(\theta|\mathbf{Y})$ based on the prior $p(\theta)+\delta(\theta)$. Suppose that there is the Fréchet derivative of $E[h(\theta)|Y,p]$ in the direction $\delta(\theta)$, denoted by $\dot{E}[h(\theta)|Y,p](\delta)$, satisfying

$$|E[h(\theta)|Y,p_\delta]-E[h(\theta)|Y,p]-\dot{E}[h(\theta)|Y,p](\delta)|=o(||\delta||),$$

where $||\delta||$ is a specific norm of δ. A local influence measure is based on the norm of the linear operator $\dot{E}[h(\theta)|Y,p](\delta)$, given by

$$||\dot{E}[h(\theta)|Y,p](\delta)||=\sup_\delta \frac{|\dot{E}[h(\theta)|Y,p](\delta)|}{||\delta||}.$$

Furthermore, we may consider $d(p(\theta|\mathbf{Y}),p_\delta(\theta|\mathbf{Y}))$ as a target of interest and then we can similarly use a derivative norm given by

$$\sup_\delta \frac{d(p(\theta|\mathbf{Y}),p_\delta(\theta|\mathbf{Y}))}{d(p(\theta),p(\theta)+\delta(\theta))}.$$

Computationally, it is relatively straightforward to approximate these local influence measures using Markov chain Monte Carlo methods (Chen, Shao, and Ibrahim, 2000).

Although the computation of these local influence measures is relatively straightforward, they may not have desirable properties as the sample size increases to ∞ (Gustafson and Wasserman, 1995; Gustafson, 1996a, 2000). Particularly, it has been shown that the derivative norm diverges to ∞ as the sample size increases to ∞, leading to the wrong conclusion that prior influence increases with sample size (Gustafson and Wasserman, 1995) Furthermore, consider posterior means under linear prior perturbations and the L_m norm for $||\delta||$. It is only when $m=2$ that the local influence measure asymptotically depends jointly on the sample size and the prior (Gustafson, 1996a).

Example 7.6: (Gustafson and Wasserman, 1995). Let $X_1,\ldots,X_n|\theta\sim N(\theta,1)$ and $\theta\sim N(0,1)$. Then, we have

$$||\dot{E}[\theta|X,p](\delta)||=\sqrt{n+1}\exp\left\{\left(\frac{n}{n+1}\right)\frac{\bar{x}_n^2}{2}\right\}\approx\sqrt{n}\exp\left\{\frac{\bar{x}_n^2}{2}\right\},$$

which diverges at rate $\sqrt{n}$ for almost all sample paths, where $X=\{X_1,\ldots,X_n\}$, $\bar{x}_n=n^{-1}\sum_{i=1}^n x_i$, and p denotes the prior density of θ.

Based on the Bayesian perturbation manifold, Zhu, Ibrahim, and Tang (2009) systematically examine local influence measures to assess the effects of perturbing the posterior distribution. By using a Taylor's series expansion, they obtain

$$f(\omega(t)) = f(\omega(0)) + \dot{f}(\omega(0))t + 0.5\ddot{f}(\omega(0))t^2 + o(t^2),$$

where $\dot{f}(\omega(0))$ and $\ddot{f}(\omega(0))$ denote the first- and second order derivatives of $f(\omega(t))$ with respect to t evaluated at $t = 0$. There are two cases: $\dot{f}(\omega(0)) \neq 0$ for some smooth curves $\omega(t)$ and $\dot{f}(\omega(0)) = 0$ for all smooth curves $\omega(t)$. When $\dot{f}(\omega(0)) \neq 0$ holds for some smooth curves $\omega(t)$, the *first-order local influence measure* is given by

$$\mathrm{FI}_f[\mathbf{v}](\omega(0)) = \frac{\{df[\mathbf{v}](\omega(0))\}^2}{g(\mathbf{v},\mathbf{v})(\omega(0))},$$

which can be regarded as the first-order measure in an infinite-dimensional manifold and it is invariant with respect to any reparametrizations of the curve $\omega(t)$. Furthermore, Zhu, Ibrahim, and Tang (2009) use $\ddot{f}(\omega(0))$ to assess the second-order local influence of ω to a statistical model (Wu and Luo, 1993; Zhu et al., 2007; Zhu, Ibrahim, and Tang, 2009). Specifically, they introduced a *second-order influence measure* (SI) in the direction $\mathbf{v} \in T_{\omega(0)}\Gamma$ given by

$$\mathrm{SI}_f[\mathbf{v}](\omega(0)) = \frac{\mathrm{Hess}(f)(\mathbf{v},\mathbf{v})(\omega(0))}{g(\mathbf{v},\mathbf{v})(\omega(0))},$$

where $\mathrm{Hess}(f)(\mathbf{v},\mathbf{v})(\omega(0)) = \ddot{f}(\mathrm{Exp}_{\omega(0)}(t\mathbf{v}))|_{t=0}$ is a covariant (or Riemmanian) Hessian (Lang, 1995). $\mathrm{SI}_f[\mathbf{v}](\omega(0))$ is invariant with respect to scalar transformations. $\mathrm{FI}_f[\mathbf{v}](\omega(0))$ and $\mathrm{SI}_f[\mathbf{v}](\omega(0))$ are defined for finite-dimensional and infinite-dimensional spaces and have nice geometric interpretations. However, the calibration of these local influence measures is a current topic of research.

7.1.3 An Illustrative Example

We first illustrate the methodology with a logistic regression example. We considered data on 200 men taken from the Los Angeles Heart Study conducted under the supervision of John M. Chapman (Dixon and Massey, 1983). The response variable is occurrence or nonoccurrence of a coronary incident in the previous ten years. Of the 200 cases, 26 had coronary incidents and the dataset contains six other covariates: Age (x_1) (mean= 42.56, sd=11.65), Systolic blood pressure (x_2) (mean=121.64, sd=16.70), Diastolic blood pressure (x_3) (mean=81.59, sd=9.99), Cholesterol (x_4) (mean=285.11, sd=65.04), Height (x_5) (mean=65.58, sd=2.5) and Weight (x_6) (mean=165.19, sd=24.94). The logistic regression analysis of these data has been done by Christensen (1997) from a frequentist viewpoint. Here, we illustrate the proposed Bayesian diagnostic and model selection statistics under both uniform improper and normal priors for β.

The full model is given by

$$\log\{p_i/(1-p_i)\} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6}, \quad (7.1.2)$$

where p_i is the probability of the occurrence of coronary incident for the ith case for $i = 1, \cdots, 200$, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$ in which $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{i6})'$. For the normal prior for β, we took $\beta \sim N(\mathbf{0}, \kappa(\mathbf{X}'\mathbf{X})^{-1})$, and considered several values of κ including $\kappa = 1$, 3, 10, and 100. The posterior samples were obtained using Adaptive Rejection Metropolis Sampling (ARMS) (Gilks, Best, and Tan, 1995) Gibbs and 40,000 MCMC posterior samples were used in the analysis after burn-in. For numerical stability in the MCMC sampling, we standardized all of the covariates. The posterior means (standard deviations) for $\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6)'$ were, respectively, given by: -2.375 (0.289), 0.559 (0.285), 0.113 (0.356), -0.069 (0.400), 0.433 (0.245), -0.196 (0.274), and 0.528 (0.258).

TABLE 7.1. Case influence diagnostics for Chapman data.

Uniform				Normal (κ=10)				Normal (κ=100)			
Case	KL(i)	Cal.	CM(i)	Case	KL(i)	Cal.	CM(i)	Case	KL(i)	Cal.	CM(i)
86	0.202	0.788	0.404	41	0.067	0.677	0.134	41	0.150	0.754	0.299
151	0.191	0.782	0.382	5	0.055	0.661	0.110	151	0.139	0.747	0.279
192	0.179	0.774	0.358	19	0.050	0.654	0.099	192	0.122	0.732	0.243
41	0.177	0.773	0.355	151	0.048	0.651	0.096	126	0.121	0.732	0.242
126	0.166	0.766	0.331	126	0.043	0.644	0.087	86	0.121	0.732	0.242
48	0.150	0.755	0.300	48	0.041	0.641	0.083	48	0.111	0.724	0.223
129	0.143	0.749	0.286	192	0.039	0.637	0.078	5	0.105	0.718	0.210
5	0.123	0.734	0.246	113	0.037	0.633	0.073	129	0.100	0.713	0.200
21	0.108	0.720	0.216	129	0.034	0.628	0.068	19	0.088	0.701	0.177
159	0.106	0.718	0.212	111	0.032	0.624	0.064	21	0.071	0.682	0.143

Cal. denotes calibration of KL(i), computed by the methods in McCulloch (1989) and Cho et al. (2009).

To examine the performance of the proposed diagnostic measures, we computed the K-L divergence ($\phi(u) = -\log(u)$), denoted by KL(i), and CM(i). The changes in the posterior estimates across the cases were computed as well. Table 7.1 shows the top ten most influential cases based on the uniform and normal priors. We observe from Table 7.1 that case 86 (KL=0.202, CM=0.404) is identified as the most influential case followed by cases 151, 192, and 41 for the uniform prior. Under the normal prior with κ=10 and κ=100, case 41 is identified as the most influential, whereas for the normal prior with κ=1 and κ=3, essentially no influential cases are identified. This is due to the fact that the prior is very informative and thus dominates the likelihood. When κ gets large, the normal prior becomes more noninformative and thus yields similar results to the uniform prior. The changes in the posterior estimates also describe the influence of the identified cases very well (results not shown for brevity). After an investigation as to the reasons why these identified cases were more influential than others, we found that one or two of the covariate values are extreme, or that a coronary incident occurred for the cases having covariate values corresponding to those of lower risk of a coronary incident. For example, case 86 has low values of the covariates, corresponding to a low risk of a coronary incident (age (x_1)=34, cholesterol(x_4)=214, and weight (x_6)=139), but a coronary incident had oc-

curred for this case. Case 41 has an exceptionally high cholesterol value (x_4=520), and case 151 has low weight (x_6=128), which is the smallest weight among those who had coronary incidents.

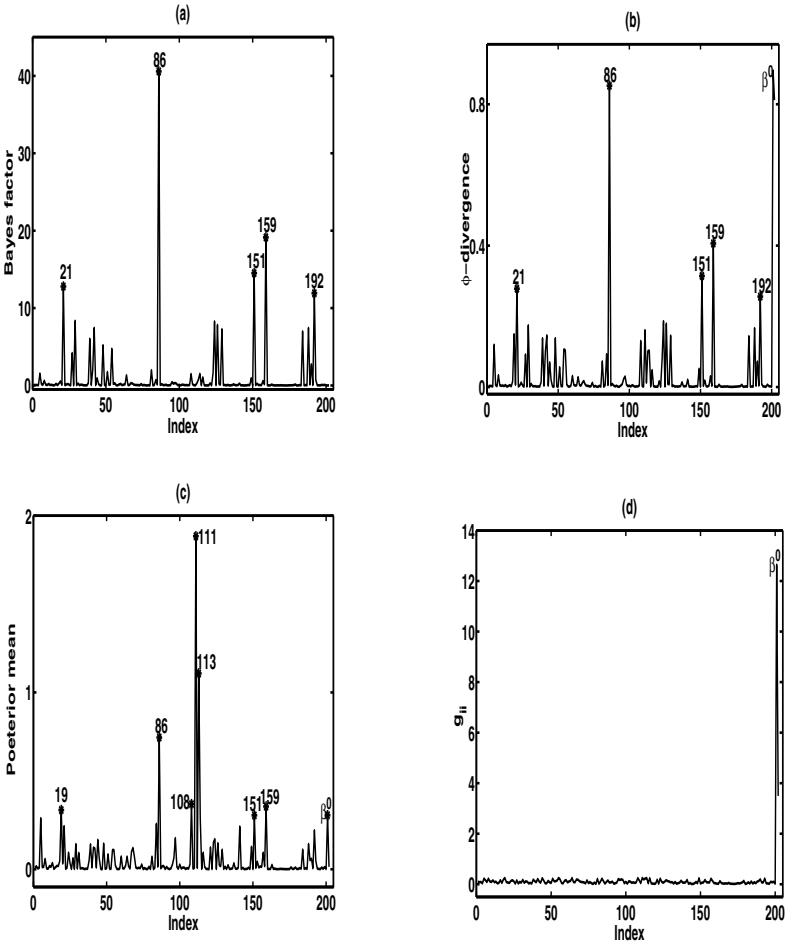


FIGURE 7.1. Bayesian local-influence measures from Chapman data. Index plots of local influence measures (a) $\mathbf{v}^B_{\max}$; (b)$\mathrm{SI}_{\mathrm{D}_{\phi},\mathbf{e}_j}$; (c) $\mathrm{SI}_{CM_h,\mathbf{e}_j}$; and (d) g_{ii} for simultaneous perturbation.

We use model (7.1.2) to fit the above data under the prior $\beta \sim N\left(\beta^0, (\mathbf{X}'\mathbf{X})^{-1}\right)$, where β^0 is taken to be the maximum likelihood estimate of β. We simultaneously perturbed the model (7.1.2) and the prior distribution of β, whose perturbed log-posterior is given by

$$\begin{aligned}
&\log p(\beta|\mathbf{X},\mathbf{Y},\omega)\\
&= \sum_{i=1}^{n}[y_i(\mathbf{x}_i+\omega_i\mathbf{1}_p)'\beta - \log(1.0+\exp(\mathbf{x}_i+\omega_i\mathbf{1}_p))] + \frac{p}{2}\log(\omega_\kappa)\\
&\quad - \frac{\omega_\kappa}{2}(\beta-\beta^0-\omega_\beta\mathbf{1}_p)'\mathbf{X}'\mathbf{X}(\beta-\beta^0-\omega_\beta\mathbf{1}_p) + \text{Constant}, \qquad (7.1.3)
\end{aligned}$$

where $\omega = (\omega_1,\ldots,\omega_n,\omega_\beta,\omega_\kappa)'$, $n = 200$, $p = 7$, $\mathbf{Y} = (y_1,\ldots,y_n)'$ and $\mathbf{1}_p = (1,\ldots,1)'$. In this case, $\omega^0 = (0,\ldots,0,0,1)'$ represents no perturbation. After some calculations, we have

$$\mathbf{G}(\omega^0) = \text{diag}(g_{11},\ldots,g_{nn},\mathbf{1}_p'\mathbf{X}'\mathbf{X}\mathbf{1}_p,p/2),$$

where $g_{ii} = E\{\exp(\mathbf{x}_i'\beta)(\mathbf{1}_p'\beta)^2/(1.0+\exp(\mathbf{x}_i'\beta))^2\}$ for $i = 1,\ldots,n$ and the expectation is taken with respect with the prior distribution of β. Then, we chose a new perturbation scheme $\tilde{\omega} = \omega^0 + \mathbf{G}(\omega^0)^{1/2}(\omega-\omega^0)$ and calculated the associated local influence measures $\mathbf{v}^B_{\max} = \text{argmax}\{\text{FI}_B[\mathbf{v}](\tilde{\omega}(0))\}$, $\text{SI}_{\text{D}_{\phi,\mathbf{e}_j}}$ and $\text{SI}_{CM_h,\mathbf{e}_j}$ via 40,000 MCMC posterior samples after burn-in, in which $\phi(.)$ was chosen to be the Kullback divergence and $h(\beta) = \beta$. Examination of Figures 7.1 (a) and 7.1 (b) indicates that cases 21, 86, 151, 159, and 192 were detected to be influential by our local influence measures; Figure 7.1 (c) shows that cases 19, 86, 108, 111, 113, 151, and 159 were detected to be influential by our local influence measures, and the prior for β has a large effect on the analysis; Figure 7.1 (d) shows that the metric tensor $g_{ii}(\omega^0)$ for perturbation (7.1.3) changes very little for all individuals, but there is a large change for the metric tensor $g_{ii}(\omega^0)$ corresponding to a perturbation in the prior mean of β.

To examine the robustness of the sampling model, we consider the following global perturbation to $p(\mathbf{X},\mathbf{Y}|\beta)$:

$$p(\mathbf{X},\mathbf{Y}|\beta,\omega) = \exp\{\sum_{i=1}^{n}[y_i(\mathbf{x}_i'\beta+\omega) - \log(1.0+\exp(\mathbf{x}_i'\beta+\omega))]\}.$$

In this case, $\omega = \omega^0 = 0$ represents no perturbation. After some calculations, we have $\mathbf{G}(\omega^0) = E\{\sum_{i=1}^{n}\exp(\mathbf{x}_i'\beta)/(1.0+\exp(\mathbf{x}_i'\beta))^2\}$. Three local influence measures, including the Bayes factor, ϕ-divergence and posterior mean with $h(\beta) = \beta$ for the above Chapman dataset are calculated with the prior $\beta \sim \left(\beta^0,(\mathbf{X}'\mathbf{X})^{-1}\right)$, and are denoted by t_B, t_ϕ and t_P, respectively. Let $\tilde{\beta}$ denote the posterior mean of β obtained via the Gibbs sampler. For 100 bootstrap datasets which are generated from the fitted model (7.1.2) with $\beta = \tilde{\beta}$, the three corresponding local influence measures, denoted by t^k_{bB}, $t^k_{b\phi}$ and t^k_{bP} $(k = 1,\ldots,100)$, are calculated using the prior of $\beta \sim N\left(\tilde{\beta},(\mathbf{X}'\mathbf{X})^{-1}\right)$ for each generated bootstrap dataset. The three corresponding p-values are given by 0.70, 0.53, and 0.10, respectively. These results show that these influences measures are not "significant" and therefore the sampling model may be correctly specified.

Acknowledgments: This work was supported in part by NSF SES-06-43663 and BCS-08-26844, NIH grants UL1-RR025747-01, MH086633 and AG033387 to Dr. Zhu, NIH grants GM 70335 and CA 74015 to Dr. Ibrahim, and NSFC 10961026 and NCET-07-0737 to Dr. Tang. The work was done while Hyunsoon Cho was at the University of North Carolina at Chapel Hill.

7.2 The Choice of Nonsubjective Priors on Hyperparameters for Hierarchical Bayes Models

Gauri S. Datta and J.N.K. Rao

We start with a quote from Berger, Strawderman, and Tang (2005), "Hierarchical modeling is wonderful and here to stay, ...". Hierarchical models, which can also be represented as random effects models, are quite common in statistics. These models are used for simultaneous estimation of many similar parameters representing some common characteristic for several subpopulations. Among many applications of such models, small area estimation is at the forefront. Random effects models have played a central role in obtaining reliable indirect estimators of small area means when the traditional area-specific direct estimators are not suitable due to small area-specific samples. Usefulness of this approach relies heavily on the availability of suitable external auxiliary information. For a detailed account of the model-based approach to small area estimation we refer the reader to Rao (2003); see Datta (2009) for a recent review.

Both Bayesian and frequentist approaches to model fitting are popular in model-based estimation of small area means. In the Bayesian approach small area means are assigned a prior distribution, while in the frequentist approach they are expressed as functions of random effects and regression on some auxiliary variables. The random effects attempt to explain possible lack of fit in a regression between the small area mean of the response and the explanatory variables. While a Bayesian modeling is inherently hierarchical, the steps in frequentist modeling are essentially the same with a hierarchical structure. In the frequentist approach, a sampling model for the data (or direct estimators) is specified in the first-stage, conditional on the random effects, and in the second-stage a model for the random effects is specified. Here all the parameters including the variance components (of the random effects and/or the sampling error) and the regression parameter(s) are collectively called model parameters. In the HB approach, the prior distribution of the first-stage parameters involves other parameters, which are called hyperparameters. Usually, the hyperparameters and model parameters have a large overlap.

In the frequentist approach, a Bayes predictor of a small area mean is first obtained and the unknown model parameters in the predictor is then replaced by suitable estimators, leading to an empirical Bayes (EB) predictor of the small area mean. On the other hand, in the HB approach some suitable prior distribution (also known

as hyperprior) is assigned to the hyperparameters. The posterior distribution of the hyperparameters is used to integrate them out from the Bayes predictor of a small area mean. The latter approach, which is fully Bayesian, is called the HB approach.

An advantage of the Bayesian approach is that the information on a parameter based on data (or likelihood) and a prior distribution, possibly subjective, can be combined in a probabilistic manner. Unfortunately, in many circumstances no suitable subjective prior is available. In such situations, to implement the Bayesian machinery one usually considers some nonsubjective priors. Such priors are usually diffuse or noninformative, and not unique. Furthermore, not all those choices necessarily lead to "good" Bayesian solutions. There is a large literature devoted to the development of suitable nonsubjective priors which produce reasonable Bayesian solutions. Two major developments are based on the reference prior idea of Berger, Bernardo and others, and the probability matching approach by Welch and Peers (1963) and others. However, such developments are mostly aimed at standard non-hierarchical models.

Both the reference prior and the probability matching prior approaches partition the parameter vector into a vector of interest parameters and a vector of nuisance parameters. A reference prior is obtained by maximizing certain expected Kullback-Leibler distance of the prior and the resulting posterior. Derivation of such a prior is usually very involved. One may find details of this approach in Berger and Bernardo (1992a), Ghosh and Mukerjee (1992), and Berger, Bernardo, and Sun (2009). The probability matching approach, on the other hand, imposes the requirement on a nonsubjective prior to generate a Bayesian solution that also possesses good frequentist property. For a thorough account of this approach, we refer the reader to Ghosh (1994), Ghosh and Mukerjee (1998), Datta and Mukerjee (2004), and Datta and Sweeting (2005).

The above mentioned concepts for the development of nonsubjective priors have been mostly used to propose suitable priors for specified parameters of interest in non-hierarchical models. They have also been used in such models to derive appropriate priors when the interest is not in estimation but in the prediction of a future observation. Kuboki (1998) developed reference priors and Datta et al. (2000) and Datta and Mukerjee (2003) developed probability matching priors for prediction of a future observation in non-hierarchical models. However, the development of suitable nonsubjective hyperpriors for the hyperparameters in HB models has been very sparse. It is worthwhile to point out that in HB models usually at the second-stage of the hierarchy, proper priors, typically some suitable conjugate priors, are specified for the parameters appearing in the first-stage, which is usually the likelihood part, of the model. In small area estimation, the first-stage parameters are related to small area means. In an HB model in the absence of suitable hyperpriors, hyperparameters are usually assigned some diffuse or noninformative priors. In this context, we refer the reader to Berger and Strawderman (1996) and Berger, Strawderman, and Tang (2005) on the importance on the selection of appropriate hyperpriors. In the context of normal HB model problems, they noticed that these default hyperpriors can lead to an inadequate overall prior (for the first-stage parameters) which often leads to inadmissible Bayes estimates.

Usually the parameters of interest in an HB model are the first-stage parameters. Conditional on some hyperparameters, the first-stage parameters are assigned a proper distribution. However, it is necessary to derive suitable nonsubjective priors on the hyperparameters of the model. Derivation of such priors is usually quite complicated (see Section 7.2.3). Indeed, to the best of our knowledge there is no published article on reference hyperpriors. However, based on the probability matching approach Datta et al. (2000) and Chang, Kim, and Mukerjee (2003) derived nonsubjective hyperpriors in simple hierarchical models. They derived these priors by treating the unobservable random effects as objects of inferential interest. The random effects are actually related to the first-stage parameters or the small area means in HB models.

In small area estimation usually two types of data are available. In most frequently occurring applications, only area-level summary data are available both for the response variable and the auxiliary variables. Occasionally, there are also applications where such data are available at the unit-level within each small area. Two popular hierarchical models in small area estimation are the model by Fay and Herriot (1979) for area-level data, and the nested error linear regression model of Battese, Harter, and Fuller (1988) for unit-level data. The latter model is really an extension of a one-way ANOVA model for random effects to include explanatory variables.

In small area estimation the goal is to obtain reliable point estimates of small area means and accurate estimates of variability of the point estimates. In the frequentist approach to small area estimation, point estimates are provided by EB predictors, and nearly unbiased estimates of the mean-squared error (MSE) of the estimates are provided for the variability of the estimates. Accurate asymptotic estimation of the MSE of the EB predictor has been extensively discussed in the small area estimation literature (see for example, Chapters 7 and 9 of Rao, 2003).

In an HB approach to small area estimation, posterior means and posterior variances provide the necessary inference. To derive these Bayesian estimates, default noninformative priors on hyperparameters have been routinely used in small area estimation (see for example, Datta, Fay, and Ghosh, 1991; Datta and Ghosh, 1991; Nandram and Sedransk, 1993). Even though the resulting posterior means and the EB predictors are in reasonable agreement, at least for the normal linear model, the corresponding frequentist and Bayesian measures of uncertainty are not necessarily close. Datta, Rao, and Smith (2005) showed that for the Fay-Herriot area-level model it is possible to choose a prior for the hyperparameters so that the resulting posterior variance of a small area mean agrees with the estimated MSE of the corresponding EB predictor in certain asymptotic sense. The underlying principle in this solution is similar to that in the probability matching approach. In both cases, suitable frequentist performance of a Bayesian procedure is evaluated and the priors are selected in order to ensure good frequentist performance of the resulting Bayes procedure. While in probability matching context, usually the frequentist validity of a posterior quantile is sought, Datta, Rao, and Smith (2005) sought frequentist validity of a posterior variance as a suitably accurate estimator of the MSE of the empirical best linear unbiased prediction (EBLUP). In both cases these priors are typically

obtained by solving certain partial differential equations. We present details of this method in Section 7.2.2. In Section 7.2.1, we present probability matching priors in small area estimation based on the Fay-Herriot model. This is done by ensuring an approximate frequentist validity of the posterior quantile of a small area mean. In Section 7.2.3 we present a discussion outlining the difficulty in developing suitable reference hyperpriors in an HB model.

7.2.1 Probability Matching in Small Area Estimation

In this section we first briefly review the probability matching equation by Chang, Kim, and Mukerjee (2003) for prediction of the first-stage parameters in an HB model. These authors considered random variables $Y_1, \ldots, Y_m$ and $\theta_1, \ldots, \theta_m$, with their joint distribution indexed by an unknown parameter $\delta = (\delta_1, \ldots, \delta_q)'$. They specified

(a) conditional on $\theta_1, \ldots, \theta_m$ and δ, the observable random variables $Y_1, \ldots, Y_m$ are independently distributed with a density of Y_i given by $f(y_i|\theta_i, \delta)$; and
(b) conditional on δ, the unobservable first-stage parameters $\theta_1, \ldots, \theta_m$ are independent and identically distributed, each with a common density $g(\cdot|\delta)$. It is assumed that θ_i's are scalars.

Note that given δ, marginally $Y_1, \ldots, Y_m$ are iid with the joint density

$$\prod_{i=1}^{m} \psi(y_i|\delta) = \prod_{i=1}^{m} \int_{-\infty}^{\infty} f(y_i|\theta_i, \delta) g(\theta_i|\delta) d\theta_i. \tag{7.2.1}$$

Let $\lambda(\delta)$ denote the average log-likelihood $m^{-1} \sum_{i=1}^{m} \log \psi(Y_i|\delta)$ for δ based on the marginal distribution. We use $\hat{\delta}$ to denote the MLE of δ based on this likelihood. It is assumed that $I(\delta)$, the per unit observation information matrix for δ based on the distribution in (7.2.1), is positive definite. Elements of $I(\delta)^{-1}$ are denoted by $I^{st}(\delta)$. We use the notation $D_j \equiv \partial/\partial\delta_j$. Suppose $c_{jr} = -\{D_j D_r \lambda(\delta)\}_{\delta=\hat{\delta}}$, $a_{jrs} = \{D_j D_r D_s \lambda(\delta)\}_{\delta=\hat{\delta}}$. Let C^{-1} with elements c^{jr} denote the inverse of the observed information matrix C. Let a prior $\pi(\delta)$ on the hyperparameter δ be positive and three times continuously differentiable. Let $h(\theta_i|Y_i, \delta)$ be the conditional density of θ_i given Y_i and δ, and for $0 < \alpha < 1$, let $q(\delta, \alpha, y_i)$ be the $(1-\alpha)-$quantile of the density $h(\theta_i|y_i, \delta)$. Let $\tilde{\pi}(\theta_i|\mathbf{Y})$ denote the posterior predictive density of θ_i given the data $\mathbf{Y} = (Y_1', \ldots, Y_m')'$, under the hyperprior $\pi(\delta)$.

We further define: $\pi_j(\delta) = D_j \pi(\delta)$, $h_j(\theta_i|Y_i, \delta) = D_j h(\theta_i|Y_i, \delta)$ and $h_{jr}(\theta_i|Y_i, \delta) = D_j D_r h(\theta_i|Y_i, \delta)$. Using summation convention, Chang, Kim, and Mukerjee (2003) showed

$$\tilde{\pi}(\theta_i|\mathbf{Y})$$
$$= h(\theta_i|Y_i,\hat{\delta}) + \frac{1}{2m}\left[c^{st}\left\{c^{jr}a_{jrs} + \frac{2\pi_s(\hat{\delta})}{\pi(\hat{\delta})}\right\}h_t(\theta_i|Y_i,\hat{\delta}) + c^{jr}h_{jr}(\theta_i|Y_i,\hat{\delta}) + o(1)\right].$$

Then Chang, Kim, and Mukerjee (2003) showed that a prior $\pi(\delta)$ is probability matching, in the sense of ensuring frequentist validity of the posterior quantiles of θ_i, with a margin of error up to $o(m^{-1})$, if and only if it solves the partial differential equation

$$D_s\{I^{st}V_t(\delta,\alpha)\pi(\delta)\} = 0, \tag{7.2.2}$$

where, using $E_\delta(\cdot)$ to denote the expectation under the marginal density (7.2.1), $V_t(\delta,\alpha) = E_\delta\left\{\int_{q(\delta,\alpha,Y_i)}^{\infty} h_t(\theta_i|Y_i,\delta)d\theta_i\right\}$.

We now focus on developing probability matching hyperprior for the basic Fay-Herriot small area model. This model is given by

$$(i)\ Y_i|\theta_i,\delta \overset{ind}{\sim} N(\theta_i,C_i),\ i=1,\ldots,m;\ \ (ii)\ \theta_i|\delta \overset{ind}{\sim} N(\mathbf{x}_i'\beta,\tau^2),\ i=1,\ldots,m. \tag{7.2.3}$$

Here Y_i is a direct estimator of the small area mean θ_i, and C_i is the sampling variance of Y_i assumed to be known. Further, $\mathbf{x}_i$ is a known $p\times 1$ vector of auxiliary variables associated with area i and $\delta = (\beta',\tau^2)'$ with $q = p+1$. In practice, the sampling variances are unknown, but methods have been proposed in the literature to produce stable estimates of the C_i's which are then treated as if known (see Rao, 2003, Chapter 7 for examples). You and Chapman (2006) studied a fully HB model by using a prior on the C_i, assuming that independent estimators of the sampling variances C_i are available.

Due to the presence of the auxiliary variables $\mathbf{x}_i$'s and possibly unequal C_i's, the Y_i's are not identically distributed marginally, but under suitable conditions on the $\mathbf{x}_i$'s and C_i's (such as $\max_{1\le i\le m}C_i/\min_{1\le i\le m}C_i$ bounded), it can be established that the matching equation (7.2.2) still holds.

Remark 7.1. Suppose τ^2 is known. In this case, $\delta=\beta$. It can be checked that the elements I^{st} and V_t in (7.2.2) are free from δ. Then the flat prior $\pi(\delta)\equiv 1$ satisfies the differential equation (7.2.2). Indeed in this case it can be shown that the matching property is exact under the flat prior. For similar exact matching results for estimation and prediction in non-hierarchical models, we refer the reader to Welch and Peers (1963), Datta, Ghosh, and Mukerjee (2000), Severini, Mukerjee, and Ghosh (2002), and Berger and Sun (2008).

Now suppose that both β and τ^2 are unknown. In this case, $\delta=(\beta',\tau^2)'$. It can be checked that $I(\delta)$ is given by

$$I(\beta,\tau^2) = \mathrm{diag}\left(m^{-1}\sum_{i=1}^{m}(\tau^2+C_i)^{-1}\mathbf{x}_i\mathbf{x}_i', (2m)^{-1}\sum_{i=1}^{m}(\tau^2+C_i)^{-2}\right),$$

depending only on τ^2. In view of the orthogonality of β and τ^2, and Remark 7.1, it is compelling to use $\pi(\delta)$ of the form $\pi^0(\tau^2)$, depending only on τ^2. It can also

be checked that V_t's depend only on τ^2 and $V_q = u(z_\alpha)C_i/\{\tau^2(\tau^2+C_i)\}$, where $u(z_\alpha)$ is a known function of z_α, the $\alpha-$quantile of standard normal distribution. The matching prior is given by $\pi(\beta,\tau^2) = \pi^0(\beta)\pi^0(\tau^2)$ with $\pi^0(\beta) = 1$ and

$$\pi^0(\tau^2) \propto \frac{\tau^2(\tau^2+C_i)}{C_i} \sum_{i=1}^{m} (\tau^2+C_i)^{-2}. \tag{7.2.4}$$

While the prior in (7.2.4) is not proper, it can be shown that the resulting posterior will be proper for $m \geq p+2$ where p is the full column rank of the matrix $\mathbf{X} = (\mathbf{x}_1,\ldots,\mathbf{x}_m)'$. An unpleasant feature of this prior is that it depends on the index i of the small area. However, this is consistent with the idea of a probability matching prior which is always obtained for a specific parameter or parametric function. If C_i's are all equal to C_0, thereby leading to the balanced Fay-Herriot model, then the prior in (7.2.4) simplifies to $\pi^0(\tau^2) \propto \tau^2/(\tau^2+C_0)$. Datta, Ghosh, and Mukerjee (2000), derived this prior earlier through a direct argument. Note that this prior is different from the uniform prior for τ^2 on $(0,\infty)$, recommended by Morris (1983) and Berger (1985).

The area specific prior (7.2.4) is appropriate for making inference on the small area mean θ_i. However, a quantile matching prior for a weighted mean $\sum_{i=1}^m w_i\theta_i$ of all areas with specified weights $w_1,\ldots,w_m, \sum_{i=1}^m w_i = 1$, can be derived. In a somewhat informal way, it can be shown that a matching prior in this case will have the form

$$\pi^0(\tau^2) \propto \frac{\tau^2 \sum_{i=1}^m w_i^2 C_i(\tau^2+C_i)^{-1}}{\sum_{i=1}^m w_i^2 C_i^2(\tau^2+C_i)^{-2}} \times \sum_{i=1}^{m} (\tau^2+C_i)^{-2}.$$

If $w_i \propto C_i^{-1}$ then $\pi^0(\tau^2)$ is proportional to $\sum_{i=1}^m \tau^2/\{C_i(\tau^2+C_i)\}$.

7.2.2 Frequentist Evaluation of Posterior Variance

Second-order accurate approximation to the MSE of an EBLUP or EB predictor of a small area mean and second-order (or nearly) unbiased estimation of the MSE have played a dominant role in frequentist approach to small area estimation; see for example Prasad and Rao (1990) for an early development and Rao (2003) for subsequent developments. In the second-order approximation, terms of order smaller than $O(m^{-1})$ are neglected. For the Fay-Herriot model in (7.2.3), for known δ, the best (or Bayes) predictor of θ_i under squared error loss is given by

$$\tilde{\theta}_i^B(\delta) = E(\theta_i|Y_i) = Y_i - \frac{C_i}{\tau^2+C_i}(Y_i - \mathbf{x}_i'\beta) = \{1-B_i(\tau^2)\}Y_i + B_i(\tau^2)\mathbf{x}_i'\beta,$$

where $B_i(\tau^2) = C_i/(\tau^2+C_i)$. Let $\hat{\delta}$ denote an estimator of δ. Then an EB predictor or an EBLUP of θ_i is $\tilde{\theta}_i^B(\hat{\delta})(\equiv \hat{\theta}_i^{EB})$. While different estimators of δ have been considered in the literature, we will focus here on the maximum likelihood or the residual maximum likelihood estimator. The MSE of the resulting EBLUP predictor

is asymptotically smaller than the MSE of other EBLUP predictors resulting from different choices of estimators of δ (Datta, Rao, and Smith, 2005; Datta, 2009).

Datta and Lahiri (2000) showed that

$$\text{MSE}(\hat{\theta}_i^{EB}) = g_{1i}(\tau^2) + g_{2i}(\tau^2) + g_{3i}(\tau^2) + o(m^{-1}), \tag{7.2.5}$$

where $g_{1i}(\tau^2) = C_i(1 - B_i(\tau^2))$, $g_{2i}(\tau^2) = \{B_i(\tau^2)\}^2 \mathbf{x}_i' \Big[\sum_{i=1}^m (\tau^2 + C_i)^{-1} \mathbf{x}_i \mathbf{x}_i'\Big]^{-1} \mathbf{x}_i$ and

$$g_{3i}(\tau^2) = 2C_i^2(\tau^2 + C_i)^{-3} \Big\{ \sum_{i=1}^m (\tau^2 + C_i)^{-2} \Big\}^{-1}.$$

In the decomposition of the $\text{MSE}(\hat{\theta}_i^{EB})$ in (7.2.5), the first term is of the order $O(1)$, while the last two terms are of the order $O(m^{-1})$. Datta, Rao, and Smith (2005) derived a hyperprior $\pi(\delta)$ by requiring that the posterior variance $V(\theta_i|\mathbf{Y})$ is a second-order unbiased estimator of $\text{MSE}(\hat{\theta}_i^{EB})$ in the sense that

$$E_\delta\{V(\theta_i|\mathbf{Y})\} = \text{MSE}(\hat{\theta}_i^{EB}) + o(m^{-1}). \tag{7.2.6}$$

The posterior variance of θ_i for such a prior will have a dual justification. Such a dual justification of a measure of variability of small area estimators is appealing to survey practitioners.

As in Section 7.2.1, we first assume that τ^2 known and use a uniform prior on β. For known τ^2, the frequentist predictor is the BLUP and is identical with the hierarchical Bayes (HB) predictor of θ_i. Moreover the MSE is known, and exactly the same as the posterior variance, and it is given by $g_{1i}(\tau^2) + g_{2i}(\tau^2)$.

We now turn to the case of unknown τ^2 and, as in the last section, consider prior $\pi(\delta)$ of the form $\pi^0(\tau^2)$. The posterior variance of θ_i may be written as

$$V(\theta_i|\mathbf{Y}) = E[g_{1i}(\tau^2)|\mathbf{Y}] + V[B_i(\tau^2)(Y_i - \mathbf{x}_i'\beta)|\mathbf{Y}]. \tag{7.2.7}$$

To derive an expansion of the posterior variance, given by (7.2.7), we employ the following lemma, using the notation of Section 7.2.1.

Lemma 7.1. *For a smooth function $b(\delta)$ and prior $\exp(\rho(\delta))$,*

$$\begin{aligned} E[b(\delta)|\mathbf{y}] &= b(\hat{\delta}) + \frac{1}{m}\{\hat{b}_j\hat{\rho}_k c^{jk} + \frac{1}{2}a_{jrs}\hat{b}_k c^{jr}c^{ks} + \frac{1}{2}\hat{b}_{jk}c^{jk}\} + o(m^{-1}) \\ &= b(\hat{\delta}) + \frac{1}{m}\{\hat{b}_j\hat{\rho}_k \hat{I}^{jk} + \frac{1}{2}a_{jrs}\hat{b}_k \hat{I}^{jr}\hat{I}^{ks} + \frac{1}{2}\hat{b}_{jk}\hat{I}^{jk}\} + o(m^{-1}), \\ V[b(\delta)|\mathbf{y}] &= \frac{1}{m}\hat{b}_j\hat{b}_k c^{jk} + o(m^{-1}) = \frac{1}{m}\hat{b}_j\hat{b}_k \hat{I}^{jk} + o(m^{-1}). \end{aligned}$$

In the above, we use the notation $\hat{f}$ to denote $f(\hat{\delta})$. Note that up to the order of $o(m^{-1})$ the posterior variance of $b(\delta)$ is free from the prior.

Using the second part of Lemma 7.1, after some simplifications it can be shown that

$$E_\delta\Big[V\{B_i(\tau^2)(Y_i-\mathbf{x}_i'\beta)|\mathbf{Y}\}\Big]=g_{2i}(\tau^2)+g_{3i}(\tau^2)+o(m^{-1}). \tag{7.2.8}$$

Now by the first part of Lemma 7.1,

$$\begin{aligned}E_\delta\Big[E\{g_{1i}(\tau^2)|\mathbf{Y}\}\Big]&=E_\delta\{g_{1i}(\hat\tau^2)\}+\frac{1}{m}E_\delta\{g_{1i,j}(\hat\tau^2)\hat\rho_k^0\hat I^{jk}\}\\&\quad+\frac{1}{2m}E_\delta\{a_{jrs}g_{1i,k}(\hat\tau^2)I^{jr}I^{ks}\}+\frac{1}{2m}E_\delta\{g_{1i,jk}(\hat\tau^2)\hat I^{jk}\}\\&\quad+o(m^{-1}).\end{aligned} \tag{7.2.9}$$

Let $\Sigma(\tau^2)=\mathrm{diag}(\tau^2+C_1,\ldots,\tau^2+C_m)$. After considerable simplifications we get:

$$\begin{aligned}E_\delta\{g_{1i}(\hat\tau^2)\}&=g_{1i}(\tau^2)-\frac{B_i^2(\tau^2)\mathrm{tr}\Big[(\mathbf{X}'\Sigma^{-1}(\tau^2)\mathbf{X})^{-1}(\mathbf{X}'\Sigma^{-2}(\tau^2)\mathbf{X})\Big]}{\sum_{i=1}^m(\tau^2+C_i)^{-2}}\\&\quad-g_{3i}(\tau^2)+o(m^{-1}),\end{aligned} \tag{7.2.10}$$

$$\text{second term on the rhs of (7.2.9)}=\frac{2B_i^2(\tau^2)\rho^{0\prime}(\tau^2)}{\sum_{i=1}^m(\tau^2+C_i)^{-2}}+o(m^{-1}), \tag{7.2.11}$$

$$\begin{aligned}\text{third term on the rhs of (7.2.9)}&=\frac{B_i^2(\tau^2)\mathrm{tr}\Big[(\mathbf{X}'\Sigma^{-1}(\tau^2)\mathbf{X})^{-1}(\mathbf{X}'\Sigma^{-2}(\tau^2)\mathbf{X})\Big]}{\sum_{i=1}^m(\tau^2+C_i)^{-2}}\\&\quad+\frac{4B_i^2(\tau^2)\sum_{i=1}^m(\tau^2+C_i)^{-3}}{\{\sum_{i=1}^m(\tau^2+C_i)^{-2}\}^2}+o(m^{-1}),\end{aligned} \tag{7.2.12}$$

$$\text{and fourth term on the rhs of (7.2.9)}=-g_{3i}(\tau^2)+o(m^{-1}). \tag{7.2.13}$$

Equation (7.2.10) follows from equation (7) and Remark 1 of Datta and Lahiri (2000). Also, in (7.2.11), we write $\rho^0(\tau^2)=\log\pi^0(\tau^2)$. From (7.2.9) to (7.2.13), we get

$$\begin{aligned}E_\delta\Big[E\{g_{1i}(\tau^2)|\mathbf{Y}\}\Big]&=g_{1i}(\tau^2)-2g_{3i}(\tau^2)+2B_i^2(\tau^2)\rho^{0\prime}(\tau^2)\Big\{\sum_{i=1}^m(\tau^2+C_i)^{-2}\Big\}^{-1}\\&\quad+\frac{4B_i^2(\tau^2)\sum_{i=1}^m(\tau^2+C_i)^{-3}}{\{\sum_{i=1}^m(\tau^2+C_i)^{-2}\}^2}+o(m^{-1}).\end{aligned} \tag{7.2.14}$$

Now from (7.2.5), (7.2.7), (7.2.8), and (7.2.14), in order to achieve (7.2.6) we need to solve the following differential equation:

$$-2g_{3i}(\tau^2)+2B_i^2(\tau^2)\rho^{0\prime}(\tau^2)\Big\{\sum_{i=1}^m(\tau^2+C_i)^{-2}\Big\}^{-1}+\frac{4B_i^2(\tau^2)\sum_{i=1}^m(\tau^2+C_i)^{-3}}{\{\sum_{i=1}^m(\tau^2+C_i)^{-2}\}^2}=0.$$

A solution of the differential equation is given by

$$\pi_0(\tau^2) \propto (\tau^2 + C_i)^2 \{\sum_{i=1}^{m} (\tau^2 + C_i)^{-2}\}. \tag{7.2.15}$$

The MSE matching prior (7.2.15) was obtained by Datta, Rao, and Smith (2005). Note that this prior is usually different from the uniform prior of Morris. Also, it is dependent on the individual small area we are interested in. If $C_i = C_0$ for all areas i, then (7.2.15) reduces to the uniform prior. Ganesh and Lahiri (2008) made a slight generalization of the above result. Note that the MSE matching prior (7.2.15) is different from the probability matching prior (7.2.4).

Suppose as in Section 7.2.1 for some suitable weights w_i's one is interested in matching prior for the weighted mean $\sum_{i=1}^{m} w_i \theta_i$. Then it can be shown that this prior is given by $\pi(\tau^2) \propto \{\sum_{i=1}^{m} (\tau^2 + C_i)^{-2}\}\{\sum_{i=1}^{m} w_i^2 C_i^2 (\tau^2 + C_i)^{-2}\}^{-1}$. In the special case when $w_i \propto C_i^{-1}$, the above solution simplifies to the uniform prior on $(0, \infty)$.

We now turn to the MSE matching prior for the nested error regression model. This model is useful for unit-level data and is given by

$$Y_{ij} = \mathbf{x}_{ij}'\beta + v_i + e_{ij}, j = 1, \ldots, n_i, i = 1, \ldots, m,$$

where v_i's and e_{ij}'s are all independent, with $v_i \sim N(0, \tau^2)$ and $e_{ij} \sim N(0, \sigma^2)$ and unknown variance parameters τ^2 and σ^2. Auxiliary variables $\mathbf{x}_{ij}$'s are such that the design matrix is of full column rank. In small area estimation, the goal is the prediction of $\theta_i = \bar{\mathbf{X}}_i'\beta + v_i$, where $\bar{\mathbf{X}}_i$ is the finite population mean vector of the auxiliary variables for the ith small area, and is assumed to be known. Based on our discussion of the Fay-Herriot model, it is clear that the auxiliary variables do not play any role in determining "matching hyperpriors." In fact, a uniform prior for β is standard, and our goal is to determine hyperprior for the variance components τ^2 and σ^2. For this purpose, one can assume $\beta = 0$. This will simplify the problem to the prediction of v_i, the random effect based on the nested error model given above without the $\mathbf{x}_{ij}'\beta$ term.

First, we consider the balanced setup where all n_i's are equal to n. After considerable simplifications, the MSE matching hyperprior for τ^2 and σ^2 is given by

$$\pi(\tau^2, \sigma^2) \propto \sigma^{-2}(\sigma^2 + n\tau^2)^{2/(n-1)}.$$

It can be checked that the resulting posterior will be proper. For the balanced nested error model, Chang, Kim, and Mukerjee (2003) obtained the quantile matching prior (for predicting v_i) given by $\pi(\tau^2, \sigma^2) \propto \tau^2 \sigma^{-2}(\sigma^2 + n\tau^2)^{(3-n)/(n-1)}$. For the unbalanced nested error model, the MSE matching equation is very complicated and apparently there is no explicit solution to the equation.

7.2.3 Discussion

In this section we have discussed some methods for determining nonsubjective priors for the hyperparameters in hierarchical Bayes (HB) models. We have taken the

frequentist validation approach to determine such priors. In particular, for HB models in small area estimation, we have used both the quantile matching and the MSE matching criteria of a resulting HB procedure to determine suitable nonsubjective priors. The hyperpriors that we obtained in our examples are less attractive in the sense that, unlike in non-hierarchical models, these hyperpriors usually depend on the individual small area under consideration. While this is not a problem for parameters in non-hierarchical models, the dependence of the hyperprior on an individual small area implies that the same hyperprior will not have good frequentist properties for all the small areas. However, the situation is not really any different from non-hierarchical models. In non-hierarchical models as well, for different parameters of interest we usually get different matching priors. While in non-hierarchical models we may not have more than one interest parameter, in contrast we are usually interested in all the first-stage parameters in HB models.

While we have presented alternative methods to determine nonsubjective hyperpriors, the most popular approach in determining nonsubjective priors, at least for non-hierarchical models, is the reference prior approach by Berger, Bernardo and their collaborators. According to this approach, for a non-hierarchical model $p(\mathbf{y}|\delta)$ when δ is the parameter of interest, determination of a reference prior $p(\delta)$ for the parameter vector δ considers the Kullback-Leibler divergence between the posterior $p(\delta|\mathbf{y})$ and the prior $p(\delta)$; in particular, the functional

$$J\{p(\delta), \mathbf{Y}\} = E^{(\mathbf{Y},\delta)}\left[E^{\delta|\mathbf{Y}}\left\{\log \frac{p(\delta|\mathbf{Y})}{p(\delta)}\right\}\right], \tag{7.2.16}$$

where the expectation $E^{\delta|\mathbf{Y}}(\cdot)$ is with respect to the posterior distribution $p(\delta|\mathbf{Y})$, and the expectation $E^{(\mathbf{Y},\delta)}(\cdot)$ is with respect to the joint distribution of $\mathbf{Y}$ and δ. Through suitable technical modifications of the functional in (7.2.16), Bernardo (1979) showed that Jeffreys' prior is obtained as the reference prior for δ. When only a subvector of δ is of interest, and the other components of δ are nuisance parameters, appropriate modifications have been proposed by Berger and Bernardo (1989, 1992a).

To apply the reference prior approach to determine a nonsubjective hyperprior for the hierarchical model presented in the beginning of Section 7.2.1, we need to consider the first-stage parameters $\theta_1, \ldots, \theta_m$ as the vector of interest parameter. A reasonable functional which is analogous to the functional in (7.2.16) is given by

$$J\{p(\delta), \mathbf{Y}\} = E^{(\mathbf{Y},\delta)}\left[E^{\theta|\mathbf{Y}}\left\{\log \frac{p(\theta|\mathbf{Y})}{p(\theta)}\right\}\right], \tag{7.2.17}$$

where $p(\theta)$ (with slight abuse of notation) is a marginal prior density for θ given by $\int g(\theta|\delta)p(\delta)d\delta$. The hyperprior $p(\delta)$ appears in the functional indirectly through $p(\theta)$. It is technically quite challenging to maximize the functional in (7.2.17) to obtain $p(\delta)$. As a simple illustration of the method, consider a special case of the Fay-Herriot model in (7.2.3) with $\mathbf{x}_i'\beta = \mu$, a scalar, which is treated as unknown but τ^2 is considered known. Also, if we restrict choosing prior for μ from the class of all normal distributions $N(\nu, r^{-1})$, then it can be shown that the functional in (7.2.17)

simplifies to

$$\frac{1}{2}\Big[\sum_{i=1}^{m}\log\{(\tau^2+C_i)/C_i\}+\log\{1+r^{-1}\sum_{i=1}^{m}(\tau^2+C_i)^{-1}\}\Big],$$

which is maximized with respect to r by taking it to be zero. The resulting hyperprior for μ becomes a uniform prior on the real line. This result is in agreement with the other solutions that we obtained earlier through other criteria. However, the problem in its full generality is far from being solved.

Acknowledgments: G.S. Datta's research on this project was partially supported by the U.S. National Science Foundation and the U.S. National Security Agency. J.N.K. Rao's research was supported by the Natural Sciences and Engineering Research Council of Canada.

7.3 Exact Matching Inference for a Multivariate Normal Model

Luyan Dai and Dongchu Sun

A p-dimensional multivariate normal population, $\mathbf{x}=(x_1,\ldots,x_n)'\sim N_p(\mu,\Sigma)$, has the probability density function,

$$f(\mathbf{x}\mid\mu,\Sigma)=(2\pi)^{-p/2}|\Sigma|^{-1/2}\exp\Big\{-\frac{1}{2}(\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu)\Big\},\qquad(7.3.18)$$

where $\mu=(\mu_1,\ldots,\mu_n)'$ are the normal means and Σ is the covariance matrix.

In Bayesian analysis, the most commonly used priors are flat Gaussian priors for normal means μ and inverse Wishart priors for the covariance matrix Σ. These priors are used for ease of computation and convenience due to the conjugate properties. However, Stein (1956) showed that such priors may lead to inferior inferences and be lack of flexibility. Later, Brown (2001) developed the generalized inverse Wishart priors to allow more flexibility and overcome the major deficiency of that distribution whilst retaining much of the ease of the inverted Wishart distribution in Bayesian analysis for normal sampling.

Another extensively discussed priors require a Cholesky decomposition of precision matrix, Σ^{-1},

$$\Sigma^{-1}=\Psi'\Psi,$$

where Ψ is a $p\times p$ lower triangular matrix with positive diagonal elements ψ_{ii} and off-diagonal ones ψ_{ij} for $j=1,\ldots i-1$ and $i=1,\ldots,p$. There is a class of such priors for (μ,Ψ) that include many famous and popular ones as special cases. Let $\mathbf{a}=(a_1,\ldots,a_p)'$. Those priors have the form,

$$\pi_{\mathbf{a}}(\mu,\Psi)d\mu d\Psi=\prod_{i=1}^{p}\frac{1}{\psi_{ii}^{a_i}}d\mu d\Psi,\qquad(7.3.19)$$

which suggests a constant prior for μ be independent of the prior component for Ψ. When $a_i = p - i$, the prior in (7.3.19) corresponds to the Jeffreys (1961) (rule) prior π_J, one of the earliest objective priors discussed and applied in the Bayesian literature. A prior closely related to π_J is the independence−Jeffreys prior π_{IJ} when $a_i = p - i + 1$ in $\pi_{\mathbf{a}}$. It is derived using a constant prior for normal mean μ and the Jeffreys (rule) prior separately for Ψ conditioning on μ. These two Jeffreys priors are generally successful for bivariate normal distributions. Nevertheless, they perform badly when the dimension p goes higher.

Geisser and Cornfield (1963) proposed a prior yielding the exact frequentist matching inference for all means and variances. It also belongs to the class of priors defined in (7.3.19) and corresponds to the one with $a_i = 2 - i$. In bivariate normal models, it agrees with π_J, while it is equivalent to π_{IJ} in univariate models. However, this prior fails to prevail in estimation of correlation parameters, predictions, or other inferences involving a multivariate normal distribution. Since more desirable priors were in need, Sun and Berger (2007) derived the reference prior of (μ, Ψ) for the ordered group $\{\mu_1, \ldots, \mu_p, \psi_{11}, (\psi_{21}, \psi_{22}), \ldots, (\psi_{p1}, \ldots, \psi_{pp})\}$, which is given by

$$\pi_{R1}(\mu, \Psi) \propto \prod_{i=1}^{p} \frac{1}{\psi_{ii}}.$$

It corresponds to $\pi_{\mathbf{a}}$ with $a_i = i$ for $i = 1, \ldots, p$.

When $a_i = i$, it is the right Haar measure prior π_H introduced by Stein (1956). It could be used to calculate certain distributions arising in invariant multivariate situations. According to the discussion by Eaton (1989), the best invariant decision rule is often the formal Bayes rule for right Haar measure. Recently, Berger and Sun (2008) considered the objective inference for parameters in bivariate normal distributions with special focus on development of objective confidences or credible sets. A variety of surprising results were found including the availability of objective priors that yield exact frequentist inferences for many functions of the bivariate normal parameters, such as the correlation coefficient. The right Haar measure prior is the one that yields exact frequentist inference most frequently. Sun and Berger (2007) generalized part of the results into multivariate normal distributions particularly on reference priors for many parameters and functions of parameters. Some of the results about exact matching for the right Haar measure prior are also given in Sun and Berger (2007). We should point out that general results on probability matching priors can be found in Datta and Mukerjee (2004).

In this section, we focus on exploring exact matching inferences of the class of the objective priors $\pi_{\mathbf{a}}$ for more multivariate normal parameters and their related functions. We use the tool of **constructive posterior distributions** (see, e.g., Berger and Sun, 2008) to prove the results. Section 7.3.1 illustrates various key concepts. Several critical results of the posterior distributions relating to the objective priors $\pi_{\mathbf{a}}$ are summarized. Whereafter, Section 7.3.2 gives the main results on the objective inference of the parameters and some functions of the parameters.

7.3.1 The Background

7.3.1.1 Frequentist Coverage Probability and Exact Matching

Suppose $\pi(\theta|\mathbf{x})$ is the posterior density of parameter θ provided by data information $\mathbf{x}$. Let η be a function of θ. Define the one-sided interval $(\eta_L, q_{1-\alpha}(\mathbf{X})]$ of parameter η, where η_L is the lower bound of η and $q_{1-\alpha}(\mathbf{X})$ is the $(1-\alpha)$ posterior quantile (i.e., $P\{\eta \le q_{1-\alpha}(\mathbf{x}) \mid \mathbf{x}\} = 1-\alpha$). An appealing objective prior is to yield the frequentist coverage probability of corresponding confidence interval, i.e.,

$$P\big\{\eta \le q_{1-\alpha}(\mathbf{X}) \mid \theta\big\}$$

as close as possible to the nominal $1-\alpha$. If it equals $1-\alpha$ exactly, we conclude the prior is the exact matching prior for η.

The following lemma will be repeatedly utilized in the later context. The proof is straightforward and omitted here.

Lemma 7.2. *Let $Y_{1-\alpha}$ denote the $1-\alpha$ quantile of random variable Y.*

(a) If $g(\cdot)$ is a monotonically increasing function, $[g(Y)]_{1-\alpha} = g(Y_{1-\alpha})$ for any $\alpha \in (0,1)$.

(b) If W is a positive random variable, $(WY)_{1-\alpha} \ge 0$ if and only if $Y_{1-\alpha} \ge 0$.

7.3.1.2 Constructive Posterior Distributions

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a random sample from multivariate normal distribution $N_p(\mu, \Sigma)$ with the density function as in (7.3.18). It is known that the sufficient statistics are sample means $\bar{\mathbf{X}}_n$ and sample variance $\mathbf{S}$ such that,

$$\bar{\mathbf{X}}_n = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i, \quad \mathbf{S} = \sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}}_n)(\mathbf{X}_i - \bar{\mathbf{X}}_n)'.$$

Recall that $\Sigma^{-1} = \Psi'\Psi$. The likelihood function of (μ, Ψ) can be uniquely determined and expressed by the sufficient statistics such that,

$$\begin{aligned} &L(\mu, \Psi \mid \bar{\mathbf{X}}_n, \mathbf{S}) \\ &= (2\pi)^{-np/2}|\Psi|^n \exp\Big\{-\frac{n}{2}(\bar{\mathbf{X}}_n - \mu)'\Psi'\Psi(\bar{\mathbf{X}}_n - \mu) - \frac{1}{2}tr(\mathbf{S}\Psi'\Psi)\Big\}. \end{aligned} \tag{7.3.20}$$

Under the prior $\pi_{\mathbf{a}}$, the conditional posterior distributions of μ are p-variate multivariate normal with mean $\bar{\mathbf{X}}_n$ and variance $\frac{1}{n}\Psi^{-1}(\Psi^{-1})'$. The marginal posterior distribution of Ψ, obtained by integrating μ, is determined by sample variance $\mathbf{S}$ such that,

$$[\Psi \mid \mathbf{S}] \propto \prod_{i=1}^{p}\psi_i^{n-a_i-1}\exp\Big\{-\frac{1}{2}tr(\Psi\mathbf{S}\Psi')\Big\}. \tag{7.3.21}$$

We would like to express the marginal posteriors (7.3.21) in computational closed form, allowing direct Monte Carlo simulation. Berger and Sun (2008) called this representation as the **Constructive Posterior Distributions**. For instance, a random variable Z from the Chi-square distribution χ^2_n with n degrees of freedom can be directly expressed as the way how to sample it, i.e., $Z = \chi^2_n$. For simplicity, $\chi_n \equiv \sqrt{\chi^2_n}$. We follow the notation used by Sun and Berger (2007) that $*$ distinguishes the constructive posterior distribution from later expression involving data randomness.

From (7.3.21), the constructive posterior distributions of $\psi = (\psi_{11}, \psi_{21}, \psi_{22}, \ldots, \psi_{p1}, \ldots, \psi_{pp})$ can be formulated based on sample covariance $\mathbf{S}$. First, we consider the decomposition of $\mathbf{S}$ by lower triangular matrix $\mathbf{V}$ such that $\mathbf{S} = \mathbf{V}\mathbf{V}'$, where the diagonal elements are $v_{ii} > 0$ and off-diagonal elements are v_{ij} for $j = 1, \ldots, i-1$, $i = 1, \ldots, p$. Second, we write

$$\mathbf{V}_1 = v_{11}, \mathbf{V}_2 = \begin{pmatrix} v_{11} & 0 \\ v_{21} & v_{22} \end{pmatrix}, \cdots, \mathbf{V}_p = \begin{pmatrix} \mathbf{V}_{p-1} & 0 \\ \mathbf{v}'_{p,p-1} & v_{pp} \end{pmatrix},$$

where $\mathbf{v}_{i,i-1}$ denotes the $(i-1) \times 1$ vector of the last row of $\mathbf{V}_i$ excluding v_{ii}, $i = 2, \ldots, p$. Note that $\mathbf{V}_p = \mathbf{V}$. Then, the constructive posterior distributions of $(\psi_{11}, \psi_{21}, \psi_{22}, \ldots, \psi_{p1}, \ldots, \psi_{pp})$ given data information $\mathbf{V}$ can be expressed as

$$\begin{aligned} \psi^*_{ii} &= \frac{\chi^*_{n-a_i}}{v_{ii}}, \quad i = 1, \cdots, p, \\ \psi^*_{i,i-1} &= \mathbf{V}^{-1'}_{i-1}\mathbf{z}^*_{i,i-1} - \frac{\chi^*_{n-a_i}}{v_{ii}}\mathbf{V}^{-1'}_{i-1}\mathbf{v}_{i,i-1}, \quad i = 2, \cdots, p, \end{aligned} \tag{7.3.22}$$

where $\mathbf{z}^*_{i,i-1}$ denotes independent draws from a multivariate normal distribution $N_{i-1}(\mathbf{0}, \mathbf{I}_{i-1})$, and $\chi^*_{n-a_i}$ denotes the positive square root of the independent draws from a Chi-square distribution with degree of freedom $n - a_i$. The detailed proof of the results can be found in Section 3.2.1 in Sun and Berger (2007). They showed that the right Haar measure prior π_H can yield the exact frequentist matching inference for diagonal elements ψ_{ii}, the conditional variance $d_i = 1/\psi^2_{ii}$, and the determinant of all the upper and left corner blocks in covariance matrix, $|\Sigma_i| = \prod^i_{j=1} d_i$ in the multivariate normal model (7.3.20). As a special case, bivariate normal models have been paid more attention and discussed thoroughly by Berger and Sun (2008) in terms of the exact matching inference provided by objective priors (e.g., right Haar measure prior). One of the most important findings is that the right Haar measure prior yields the exact matching inferences for off diagonal element ψ_{21} in matrix Ψ. Regardless, it is still unknown if the properties retain for off-diagonal elements ψ_{ij} and functions of $\psi = (\psi_{11}, \psi_{21}, \psi_{22}, \ldots, \psi_{p1}, \ldots, \psi_{pp})'$ for $p > 2$.

7.3.1.3 Two Lemmas

Theoretically, we can always adopt the constructive posterior distributions in (7.3.22) to explore the properties of exact matching inference for off-diagonal elements

ψ_{ij}. However, the practical proof is troublesome due to the matrix inverse $\mathbf{V}_{i-1}^{-1}$ in (7.3.22). The next lemma reveals an explicit form of the constructive posterior distributions of Ψ in a matrix form.

Lemma 7.3. *The constructive posterior distributions of* Ψ *given* $\mathbf{V}$ *described in (7.3.22) is equivalent to*

$$\mathbf{E}^* = \Psi^* \mathbf{V} \quad or \quad \Psi^* = \mathbf{E}^* \mathbf{V}^{-1}, \tag{7.3.23}$$

where

$$\mathbf{E}^* = \begin{pmatrix} \chi^*_{n-a_1} & 0 & 0 & 0 & 0 \\ Z^*_{21} & \chi^*_{n-a_2} & 0 & 0 & 0 \\ Z^*_{31} & Z^*_{32} & \chi^*_{n-a_3} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ Z^*_{p1} & Z^*_{p2} & Z^*_{p3} & \cdots & \chi^*_{n-a_p} \end{pmatrix}_{p\times p},$$

assuming matrix elements Z^*_{ij}*'s and* $\chi^*_{n-a_i}$*'s are independent for* $j = 1,\ldots,i-1$, $i = 1,\ldots,p$. Z^*_{ij}*'s denote independent draws from a standard normal distribution* $N(0,1)$ *and* $\chi^*_{n-a_i}$*'s denote positive square root of a Chi-square distribution with degree of freedom* $n - a_i$.

Proof. The results can be verified from (7.3.22) by some tedious derivations.

Note that the constructive posterior distribution of Σ is simply $\Sigma^* = \Psi^{*-1}(\Psi^{*-1})'$. It is also can be shown that the indicated data distribution of $\mathbf{v} = (v_{11}, v_{21}, v_{22}, \ldots, v_{p1}, \ldots, v_{pp})'$ is

$$\begin{aligned} \psi_{ii} v_{ij} + \sum_{k=j}^{i-1} \psi_{ik} v_{kj} = Z_{ij} \sim N(0,1), \quad j = 1, \cdots, i-1 \\ (v_{ii}\psi_{ii})^2 \sim \chi^2_{n-i}, \quad i = 1, \ldots p \end{aligned} \tag{7.3.24}$$

given $n \geq p$. The detail of the results can be found in Section 3 of Sun and Berger (2007). Similarly, the above indicated distributions have a matrix form.

Lemma 7.4. *The indicated distributions in (7.3.24) can be written into a matrix form with*

$$\mathbf{E} = \Psi \mathbf{V} \quad or \quad \mathbf{V} = \Psi^{-1} \mathbf{E}, \tag{7.3.25}$$

where

$$\mathbf{E} = \begin{pmatrix} \chi_{n-1} & 0 & 0 & 0 & 0 \\ Z_{21} & \chi_{n-2} & 0 & 0 & 0 \\ Z_{31} & Z_{32} & \chi_{n-3} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ Z_{p1} & Z_{p2} & Z_{p3} & \cdots & \chi_{n-p} \end{pmatrix},$$

assuming matrix elements Z_{ij}*'s and* χ_{n-a_i}*'s are independent for* $j = 1,\ldots,i-1$, $i = 1,\ldots,p$. Z_{ij}*'s denote independent draws from a standard normal distribution*

$N(0,1)$ and χ_{n-a_i}'s denote positive square root of a Chi-square distribution with degree of freedom $n-i$.

Proof. The proof follows (7.3.24) and can be verified easily.

In fact, (7.3.23) and (7.3.25) suggest that the multivariate normal model (7.3.20) marginalizing population mean μ belong to the functional models discussed by Dawid and Stone (1982). Some related discussion can be seen in the next section.

7.3.2 Main Results

7.3.2.1 The Exact Matching Inference of Ψ

We now show the properties of exact frequentist matching inferences for any element ψ_{ij} in Ψ under $\pi_{\mathbf{a}}$ in the following theorem. The proof applies the matrix form of constructive posterior distributions (7.3.23) and the marginal distribution of $\mathbf{V}$ given Ψ (7.3.25).

Theorem 7.1. *For fixed i, $i=1,\ldots,p$, denote the $(1-\alpha)$ posterior quantile of ψ_{ij} under the prior $\pi_{\mathbf{a}}$ by $(\psi^*_{ij})_{1-\alpha}$, $j=1,\ldots,i$. For any $\alpha\in(0,1)$,*

$$P\{\psi_{ij}\le(\psi^*_{ij})_{1-\alpha}\mid\psi_{jj},\psi_{j+1,j},\ldots,\psi_{pj}\}=1-\alpha,$$

if and only if $a_i=i$.

Proof. Let $\mathbf{l}_i$ denote a $p\times 1$ vector with entry 1 in the ith location and 0 elsewhere for $i=1,\cdots,p$. Clearly, $\psi_{ij}=\mathbf{l}_i'\Psi\mathbf{l}_j$. It follows from (7.3.25) that $\Psi=\mathbf{E}\mathbf{V}^{-1}$ and thus $\psi_{ij}=\mathbf{l}_i'\mathbf{E}\mathbf{V}^{-1}\mathbf{l}_j$. By Lemma 7.3, the $(1-\alpha)$ posterior quantile of ψ_{ij} is

$$(\psi^*_{ij})_{1-\alpha}=(\mathbf{l}_i'\Psi^*\mathbf{l}_j)_{1-\alpha}=(\mathbf{l}_i'\mathbf{E}^*\mathbf{V}^{-1}\mathbf{l}_j)_{1-\alpha},$$

Let $\psi=(\psi_{11},\psi_{21},\psi_{22},\ldots,\psi_{p1},\ldots,\psi_{pp})$. Consequently,

$$\begin{aligned}
&P\Big\{\psi_{ij}\le(\psi^*_{ij})_{1-\alpha}\mid\psi\Big\}\\
&=P\Big\{\mathbf{l}_i'\mathbf{E}\mathbf{V}^{-1}\mathbf{l}_j\le(\mathbf{l}_i'\mathbf{E}^*\mathbf{V}^{-1}\mathbf{l}_j)_{1-\alpha}\mid\psi\Big\}\\
&=P\Big\{\mathbf{l}_i'\mathbf{E}\mathbf{E}^{-1}\Psi\mathbf{l}_j\le(\mathbf{l}_i'\mathbf{E}^*\mathbf{E}^{-1}\Psi\mathbf{l}_j)_{1-\alpha}\mid\psi_{jj},\psi_{j+1,j},\ldots,\psi_{pj}\Big\}.
\end{aligned}\tag{7.3.26}$$

The last equation is true since $\Psi\mathbf{l}_j=(\psi_{jj},\psi_{j+1,j},\ldots,\psi_{pj})'$.

If (7.3.26) equals $(1-\alpha)$ for any $\psi_{ij}\in\mathbb{R}$, consider the coverage probability of ψ_{ii}, i.e., $j=i$. It is known from (7.3.22) that $\psi^*_{ii}=\frac{\chi^*_{n-a_i}}{v_{ii}}$ and $v_{ii}=\frac{\chi_{n-i}}{\psi_{ii}}$. Then,

$$P\{\psi_{ii}\le(\psi^*_{ii})_{1-\alpha}\mid\psi\}=P\Big\{\psi_{ii}\le\Big(\frac{\chi^*_{n-a_i}}{\chi_{n-i}}\psi_{ii}\Big)_{1-\alpha}\mid\psi_{ii}\Big\}=P\{\chi_{n-i}\le(\chi^*_{n-a_i})_{1-\alpha}\}$$

equals $(1-\alpha)$ if $a_i = i$. Then necessity is verified.

In (7.3.26), it is clear that

$$\mathbf{l}_i'\mathbf{E} = (Z_{i1}, \ldots, Z_{i,i-1}, \chi_{n-i})' \text{ and } \mathbf{l}_i'\mathbf{E}^* = (Z_{i1}, \ldots, Z_{i,i-1}, \chi^*_{n-a_i})'.$$

If $a_i = i$, $\mathbf{l}'\mathbf{E}$ and $\mathbf{l}_i'\mathbf{E}^*$ are the same in terms of distributions. Provided by mutually independent elements in $\mathbf{E}^*$ and $\mathbf{E}$, it is obvious that a sufficient condition of coverage probability in (7.3.26) to be $(1-\alpha)$ is that $a_i = i$ conditioning on $(\psi_{jj}, \psi_{j+1,j}, \ldots, \psi_{pj})$ when fixing i.

Since the right Haar measure π_H implies $a_i = i$ for all i's and $\mathbf{E}^*$ and $\mathbf{E}$ are the same in terms of distributions, π_H is the exact matching prior for every element in Ψ by Theorem 7.1. Furthermore, π_H yields a fiducial model (Dawid and Stone, 1982) formulated by the functional model. Recall that

$$\mathbf{V} = \Psi^{-1}\mathbf{E} \text{ and } \Psi = \mathbf{E}\mathbf{V}^{-1},$$

where data information $\mathbf{V}$ is uniquely determined by Ψ and $\mathbf{E}$ has a known distribution over the space $\mathscr{E}$, whatever the value of Ψ. Thus Ψ is the pivotal statistic. By the results of Section 4.1 in Dawid and Stone (1982), fix set $\mathbf{A} \subset \mathscr{E}$ with $P(\mathbf{A}) = 1-\alpha$ and define

$$\mathbf{A}\mathbf{v}^{-1} = \{\mathbf{e}\mathbf{v}^{-1} : \mathbf{e} \in \mathbf{A}\} = \{\psi \in \Psi_v : \psi\mathbf{v} \in \mathbf{A}\}.$$

We have

$$P_\psi(\psi \in \mathbf{A}\mathbf{V}^{-1}) = P_\psi(\psi\mathbf{v} \in \mathbf{A}) = P(\mathbf{A}) = 1-\alpha.$$

Also

$$\prod_v(\mathbf{A}\mathbf{v}^{-1}) = P(\{\mathbf{e}^* : \mathbf{e}^*\mathbf{v}^{-1} \in \mathbf{A}\mathbf{v}^{-1}\}) = P(\mathbf{A}) = 1-\alpha,$$

where $\prod_v$ denotes the posterior distribution of Ψ. For any ψ_{ij}, we just need to carefully specify the set $\mathbf{A}$ to satisfy $P(\mathbf{A}) = 1-\alpha$. The results hold consequently. It requires that $a_i = i$ for all the i's simultaneously, although the same conclusion can be obtained by using the theory of pivotal simple fiducial models. However, Theorem 7.1 provides a much relaxed condition because it only requires $a_i = i$ in $\pi_{\mathbf{a}}$ for any fixed i to yield the exact matching inference for ψ_{ij} for arbitrary j, $j = 1, \ldots, i-1$ in the multivariate normal model (7.3.20).

7.3.2.2 The Exact Matching Inference for Functions of Ψ

Many critical parameters in mutlivariate normal models can be uniquely determined by ψ.

Example 7.7. Consider the decomposition of precision matrix as $\Sigma^{-1} = \mathbf{T}'\tilde{\Psi}^2\mathbf{T}$, where $\tilde{\Psi} = \text{diag}(\psi_{11}, \cdots, \psi_{pp})$ and

$$t_{ij} = \begin{cases} 0, & \text{if } i < j, \\ 1, & \text{if } i = j, \\ \frac{\psi_{ij}}{\psi_{ii}}, & \text{if } i > j. \end{cases} \tag{7.3.27}$$

Pourahmadi (1999) pointed out t_{ij}'s are the negatives of the coefficients of the best linear predictor of x_i based on $(x_1, \cdots, x_{i-1})$, and ψ_{ii}^2's are the precisions of the predictive distribution, since

$$x_1 \sim N(\mu_1, \psi_{11}^{-2}), \quad x_i \sim N\Big(\mu_i - \sum_{j=1}^{i-1} t_{ij}(x_j - \mu_j), \psi_{ii}^{-2}\Big), \quad i \geq 2.$$

Theorem 7.2. *Let* $\mathbf{t}_j = (t_{21}, t_{31}, t_{32}, \ldots, t_{j1}, \ldots, t_{j,j-1})'$ *and* $\tilde{\psi}_j = (\psi_{11}, \ldots, \psi_{jj})'$, $j = 2, \ldots, p$. *For fixed* i, $i = 2, \ldots, p$, *define* $(t_{ij}^*)_{1-\alpha}$ *be the* $(1-\alpha)$ *posterior quantile of* t_{ij} *under the prior* $\pi_{\mathbf{a}}$ *in (7.3.19) for arbitrary* $j \in \{1, \ldots, i-1\}$. *Then, for any* $\alpha \in (0,1)$, *the frequentist coverage probability of the credit interval* $(-\infty, (t_{ij}^*)_{1-\alpha}]$ *is*

$$P\{t_{ij} \leq (t_{ij})_{1-\alpha}^* \mid \mathbf{t}_j, \tilde{\psi}_j\} = 1 - \alpha$$

if and only if $a_i = i$.

Proof. From (7.3.22), it is easy to verify the constructive posteriors of $\mathbf{t}_{i,i-1} = (t_{i1,}, \ldots, t_{i,i-1})$ to be

$$\mathbf{t}_{i,i-1}^* = \mathbf{V}_{i-1}^{-1}\mathbf{z}_{i,i-1}^* \frac{v_{ii}}{\chi_{n-a_i}^*} - \mathbf{V}_{i-1}^{-1}\mathbf{v}_{i,i-1}.$$

The indicated data distributions of $\mathbf{t}_{i,i-1}$ are

$$\mathbf{t}_{i,i-1} = \mathbf{V}_{i-1}^{-1}\mathbf{z}_{i,i-1} \frac{v_{ii}}{\chi_{n-1}} - \mathbf{V}_{i-1}^{-1}\mathbf{v}_{i,i-1}, \quad i = 2, \ldots, p.$$

Let $\mathbf{l}_{ik}$ denotes a $i \times 1$ vector with entry 1 at the kth location and 0 elsewhere. Obvisouly, $t_{ij} = \mathbf{l}_{i-1,j}'\mathbf{t}_{i,i-1}$ and $t_{ij}^* = \mathbf{l}_{i-1,j}'\mathbf{t}_{i,i-1}^*$. Then,

$$(t_{ij}^*)_{1-\alpha} = \Big(\mathbf{l}_{i-1,j}'\mathbf{V}_{i-1}^{-1}\mathbf{z}_{i,i-1}^* \frac{v_{ii}}{\chi_{n-a_i}^*}\Big)_{1-\alpha} - \mathbf{l}_{i-1,j}'\mathbf{V}_{i-1}^{-1}\mathbf{v}_{i,i-1}$$

and

$$t_{ij} = \mathbf{l}_{i-1,j}'\mathbf{V}_{i-1}^{-1}\mathbf{z}_{i,i-1} \frac{v_{ii}}{\chi_{n-i\,1-\alpha}} - \mathbf{l}_{i-1,j}'\mathbf{V}_{i-1}^{-1}\mathbf{v}_{i,i-1}.$$

Therefore,

$$\begin{aligned} &P\{t_{ij} \leq (t_{ij}^*)_{1-\alpha} \mid \mathbf{t}, \tilde{\psi}\} \\ &= P\{\mathbf{l}_{i-1,j}'\mathbf{V}_{i-1}^{-1}\mathbf{z}_{i,i-1}\chi_{n-i}^{-1} \leq (\mathbf{l}_{i-1,j}'\mathbf{V}_{i-1}^{-1}\mathbf{z}_{i,i-1}^*\chi_{n-a_i}^{*-1})_{1-\alpha} \mid \mathbf{t}, \tilde{\psi}\}. \end{aligned} \tag{7.3.28}$$

For $i = 2, \ldots, p$, define $\tilde{\Psi}_i$, $\mathbf{E}_i$, and $\mathbf{T}_i$ be the upper and left $i \times i$ corner of matrices $\tilde{\Psi}$, $\mathbf{E}$, and $\mathbf{T}$, respectively. From (7.3.24), the two sides of the inequality in (7.3.28) can be written as

$$\mathbf{l}'_{i-1,j}\mathbf{E}_{i-1}^{-1}\tilde{\Psi}_{i-1}\mathbf{T}_{i-1}\mathbf{z}_{i,i-1}\chi_{n-i}^{-1} \text{ and } \mathbf{l}'_{i-1,j}\mathbf{E}_{i-1}^{-1}\tilde{\Psi}_{i-1}\mathbf{T}_{i-1}\mathbf{z}^*_{i,i-1}\chi^{*-1}_{n-a_i}. \tag{7.3.29}$$

which only involove parameters in matrices $\tilde{\Psi}_j$ and $\mathbf{T}_j$, i.e.,

$$P\{t_{ij} \le (t^*_{ij})_{1-\alpha} \mid \mathbf{t}, \tilde{\psi}\} = P\{t_{ij} \le (t^*_{ij})_{1-\alpha} \mid \mathbf{t}_j, \tilde{\psi}_j\}.$$

If $a_i = i$, the two terms in (7.3.29) have the same distribution. Thus, the coverage probability in (7.3.28) would be exactly $1-\alpha$. To prove necessity, choose $\mathbf{t} = (0,\ldots,0)$ and $\tilde{\psi} = (1,\ldots,1)$. Then $\mathbf{T} = \mathbf{I}$ and $\tilde{\Psi} = \mathbf{I}$. Consider $j = 1$, and then $\mathbf{l}'_{i-1,1}\mathbf{E}_{i-1}^{-1}\tilde{\Psi}_{i-1}^{-1}\mathbf{T}_{i-1} = (\chi_{n-1}^{-1}, 0, \ldots, 0)$. Clearly,

$$\begin{aligned} P\{t_{i1} \le (t^*_{i1})_{1-\alpha}\} &= P\{\chi_{n-1}^{-1}Z_{i1}\chi_{n-1}^{-1} \le \chi_{n-1}^{-1}(Z^*_{i1}\chi^{*-1}_{n-a_i})_{1-\alpha}\} \\ &= P\{\xi_{n-i} \le (\xi^*_{n-a_i})_{1-\alpha}\}, \end{aligned} \tag{7.3.30}$$

where $\xi_{n-i} = \frac{Z_{i1}}{\chi_{n-i}}$ and $\xi_{n-a_i} = \frac{Z^*_{i1}}{\chi_{n-a_i}}$, representing the constrcutive distributions of random variables following the t-distribution with degree of freedom $n-i$ and $n-a_i$ individually. Note that $\mathbf{1}_{i-1}$ is the unit vector with lenght of $i-1$. Because (7.3.30) equals $1-\alpha$, $a_i = i$.

Example 7.8. Consider the parameterization $\mathbf{D} = \text{diag}(d_1, \cdots, d_p)$ and $\mathbf{T}$ defined in (7.3.27), where $d_i = 1/\psi_{ii}^2$. We can write $\Sigma^{-1} = \mathbf{T}'\mathbf{D}^{-1}\mathbf{T}$.

Theorem 7.3.

(a) For any $\alpha \in (0,1)$, the posterior $1-\alpha$ quantile of d_i under the prior in (7.3.19) $\pi_{\mathbf{a}}$ has the expression

$$(d_i^*)_{1-\alpha} = \left(\frac{v_{ii}^2}{\chi^{2*}_{n-a_i}}\right)_{1-\alpha}. \tag{7.3.31}$$

(b) For given i, $i = 1,\ldots,p$, the frequentist coverage probability of the credible interval $(0, (d_i^*)_{1-\alpha}]$ is

$$P(d_i \le (d_i^*)_{1-\alpha} \mid \mu, \tilde{\psi}) = P\left\{\frac{1}{\chi^2_{n-i}} \le \left(\frac{1}{\chi^{2*}_{n-a_i}}\right)_{1-\alpha}\right\}, \tag{7.3.32}$$

which does not depend on $(\mu, \tilde{\psi})$ and equals $1-\alpha$ if and only if $a_i = i$.

Proof. From (7.3.22), part (a) is obvious. Combining (7.3.24), we have

$$(d_i^*)_{1-\alpha} = \left(\frac{\chi^2_{n-i}}{\chi^{2*}_{n-a_i}}\right)_{1-\alpha}\frac{1}{\psi_{ii}}.$$

Clearly,

$$P(d_i \le (d_i^*)_{1-\alpha} \mid \mu, \tilde{\psi}) = P\Big\{ \frac{1}{\psi_{ii}} \le \frac{1}{\psi_{ii}} \Big(\frac{\chi^2_{n-i}}{\chi^{2*}_{n-a_i}} \Big)_{1-\alpha} \mid \psi_{ii} \Big\}$$
$$= P\Big\{ \frac{1}{\chi^2_{n-i}} \le \Big(\frac{1}{\chi^{2*}_{n-a_i}} \Big)_{1-\alpha} \Big\},$$

which equals $1-\alpha$ if and only if $a_i = i$.

Example 7.9. Consider $\eta_i \equiv |\Psi_i| = \prod_{j=1}^{i} d_j$, the generalized variance of the upper and left $i \times i$ submatrix of Σ.

Corollary 7.1.

(a) For any $\alpha \in (0,1)$, the posterior $1-\alpha$ quantile of η_i under the prior $\pi_{\mathbf{a}}$ in (7.3.19) has the expression

$$(\eta_i^*)_{1-\alpha} = \Big(\prod_{j=1}^{i} v_{jj} \Big) \Big(\prod_{j=1}^{i} \frac{1}{\chi^{2*}_{n-a_j}} \Big)_{1-\alpha}.$$

(b) The frequentist coverage probability of the credible interval $(0, (\eta_i^*)_{1-\alpha}]$ is

$$P(\eta_i \le (\eta_i^*)_{1-\alpha} \mid \tilde{\psi}) = P\Big\{ \prod_{j=1}^{i} \frac{1}{\chi^2_{n-j}} \le \Big(\prod_{j=1}^{i} \frac{1}{\chi^{2*}_{n-a_i}} \Big)_{1-\alpha} \Big\},$$

which equals $1-\alpha$ if and only if $(a_1, \ldots, a_i) = (1, \ldots, i)$.

Proof. The proof is similar to that of Theorem 7.3.

Chapter 8
Bayesian Clinical Trials

Innovative clinical trial design is one of the currently most exciting and high impact frontiers in a Bayesian analysis. The increasingly complex nature of clinical study designs and the increasing pressures for efficient and ethical design naturally lead to Bayesian approaches. In this chapter we discuss some examples of specific research problems, including adaptive and sequential trial design, sample size choice determination for longitudinal studies, and subgroup analysis.

8.1 Application of a Bayesian Doubly Optimal Group Sequential Design for Clinical Trials

J. Kyle Wathen and Peter F. Thall

One of the big success stories of Bayesian inference for clinical trial design is the possibility to construct sophisticated adaptive and sequential designs. The sequential design proposed in Wathen and Thall (2008) is a typical example of such approaches. We review their design and illustrate it with an application to a trial with non-small cell lung cancer patients. A simulation study compares their method to a standard frequentist approach for sequential clinical trials.

8.1.1 A Non-Small Cell Lung Cancer Trial

Lung cancer is the leading cause of cancer-related death worldwide, with over one million new cases diagnosed and over 900,000 deaths from this disease each year. Approximately 75% to 80% of all lung cancers are non-small cell lung cancer (NSCLC). Patients with NSCLC that is metastatic, where the cancer cells originating in the lungs have invaded the lymphatic system, bones, or internal organs,

M.-H. Chen et al. (eds.), *Frontiers of Statistical Decision Making and Bayesian Analysis*, DOI 10.1007/978-1-4419-6944-6_8,

have a median progression-free survival (PFS) time of roughly four months. Once the disease has progressed, death follows very quickly, with reported median overall survival (OS) time in most studies between 5 and 7 months.

While NSCLC treatment typically involves a combination of chemotherapy, radiation, and localized surgery, the comparative benefits of these therapeutic modalities remain unclear. To address this issue, a randomized phase III trial for patients with metastatic NSCLC was organized to compare localized surgery or radiation therapy (LS/RT) to systemic chemotherapy (SC), the standard treatment. The primary outcome was PFS time, defined as the time from start of therapy to death or disease progression occurring in sites previously identified or in new sites not seen at the patient's baseline evaluation. Initially, a conventional design was planned. This was a two-sided group-sequential log rank testing procedure with O'Brien-Fleming (OF) boundaries (O'Brien and Fleming, 1979) up to two interim tests and a third, final test, overall size 0.05 and power 0.90 to detect a 100% increase (doubling) in median PFS time from 4 to 8 months. Using East version 5 (2007), the test cut-offs in terms of the standardized logrank statistic Z-score are ± 3.7103 when 30 events are observed, ± 2.5114 at 60 events and ± 1.993 at 89 events. The anticipated accrual rate was 2 to 4 patients per month, since in practice accrual to clinical trials can be quite variable. Table 8.1 gives trial durations and sample sizes with this design for a range of possible monthly accrual rates.

TABLE 8.1. Sample size and trial duration as functions of accrual rate for the conventional group sequential design with O'Brien-Fleming boundaries.

Accrual Rate	Trial Duration (Months)			Number of Patients		
(Patients/Month)	Min	Mean	Max	Min	Mean	Max
2.0	44.2	48.6	53.0	89	97	106
2.5	35.4	39.7	44.1	89	99	110
3.0	29.5	33.8	38.1	89	101	114
3.5	25.3	29.5	33.8	89	103	118
4.0	22.1	26.4	30.5	89	105	122

At the time the trial was designed, a previous NSCLC study of standard treatment in metastatic NSCLC had been published containing a Kaplan-Meier (KM) plot of PFS (Ciuleanu et al., 2008). Since this raised the concern that the assumption of proportional hazards (PH) would not be met, we developed an alternative design using the Bayesian Doubly Optimal Group Sequential (BDOGS) methodology proposed by Wathen and Thall (2008).

In this section we provide an overview of the BDOGS design and illustrate it by application to the lung cancer trial. We also describe how one may utilize information from available KM plots when constructing a BDOGS design. In Section 8.1.2 we review the essential features of the BDOGS design. In Section 8.1.3 we describe the application to the NSCLC trial, including simulation results comparing BDOGS to the conventional design using OF boundaries. We conclude with a brief discussion in Section 8.1.4.

8.1.2 Bayesian Doubly Optimal Group Sequential Designs

8.1.2.1 Hypotheses and Decision Criteria

When the primary outcome of a randomized clinical trial is patient survival or PFS time, the standard group sequential approach requires that the treatments have event time distributions satisfying the PH assumption. Under this assumption, interim decision boundaries for stopping and concluding superiority or possibly futility are constructed, often using the Lan-DeMets α-spending function approach (Lan and DeMets, 1983). The stopping boundaries are designed to maintain a desired overall false-positive probability, α^*, and power, β^*. At each interim analysis, a standardized normal statistic (Z-score), usually based on a log rank statistic, is calculated using the most recent data and compared to the decision boundaries to determine whether the trial should stop in favor of one of the treatments, stop for futility, or continue to enroll more patients and obtain additional information. If the PH assumption is not met, however, the properties of a conventional group sequential design may be quite different from what is assumed, and in particular the actual power may differ substantially from the nominal power (Wathen and Thall, 2008).

To simplify notation, we denote the LS/RT treatment by E and the SC treatment by S. Denote the median PFS times of E and S by θ_E and θ_S, $\delta = \theta_E - \theta_S$ with targeted improvement δ^*, and the observed data by $\boldsymbol{X}$. The goal of the trial is to test $H_0 : \delta = 0$ versus $H_1 : \delta \neq 0$. In contrast with conventional designs, which use Z-scores as test statistics and a single set of stopping boundaries, the BDOGS approach uses decision criteria based on posterior probabilities and applies sequentially adaptive Bayesian model selection. The BDOGS approach first defines a set of candidate models. Optimal boundaries are derived under each candidate model, and these are stored for future use. The candidate models are characterized in terms of the shapes of their event hazard functions. Under each model, the boundaries are optimal in the sense that they minimize the expected sample size computed as an equally weighted average under the null and alternative hypotheses. The trial is monitored group sequentially with up to K analyses and a maximum of N patients randomized fairly between the two treatments. At each interim analysis, BDOGS uses $\Pr(\delta > \delta^* \mid \boldsymbol{X})$ and $\Pr(\delta < -\delta^* \mid \boldsymbol{X})$ as decision criteria, which are analogous to, but very different from, the Z-statistics used in the conventional approach.

Based on the data at each interim decision during the trial, the optimal model, defined as that having highest posterior probability, is determined. The optimal decision boundaries under that model are used for that interim decision. Because this computation is repeated at each interim analysis, this allows the possibility that the optimal model, and consequently the boundaries, may change from one interim decision to the next as more data are obtained. By using adaptively chosen decision boundaries in this way, compared to conventional group-sequential methods the BDOGS design substantially reduces the sample size in most cases, and better maintains the targeted size and power when the PH assumption is not met.

8.1.2.2 Bayesian Model Selection

Bayesian model selection and model averaging have been used extensively in many settings (Madigan and Raftery, 1994; Kass and Raftery, 1995). However, Bayesian model selection has not been used in clinical trial design. Denote the set of J models under consideration by $\mathcal{M} = \{\mathcal{M}_1, \mathcal{M}_2, \ldots \mathcal{M}_J\}$, the prior probability of model $\mathcal{M}_\ell$ by $f(\mathcal{M}_\ell)$, and the prior odds of $\mathcal{M}_\ell$ to $\mathcal{M}_1$ by $\xi_\ell = f(\mathcal{M}_\ell)/f(\mathcal{M}_1)$, with $\xi_1 = 1$. Let $\boldsymbol{\psi}_\ell$ be the parameter vector and $\pi(\boldsymbol{\psi}_\ell \mid \mathcal{M}_\ell)$ the prior density of $\boldsymbol{\psi}_\ell$ under $\mathcal{M}_\ell$. For $\ell = 1, 2, \ldots, J$, the posterior probability of $\mathcal{M}_\ell$ given the current data $\boldsymbol{X}$ is

$$f(\mathcal{M}_\ell \mid \boldsymbol{X}) = \frac{f(\boldsymbol{X} \mid \mathcal{M}_\ell) f(\mathcal{M}_\ell)}{\sum_{r=1}^{J} f(\boldsymbol{X} \mid \mathcal{M}_r) f(\mathcal{M}_r)}, \tag{8.1.1}$$

where

$$f(\boldsymbol{X} \mid \mathcal{M}_\ell) = \int f(x \mid \boldsymbol{\psi}_\ell, \mathcal{M}_\ell) \pi(\boldsymbol{\psi}_\ell \mid \mathcal{M}_\ell) d\boldsymbol{\psi}_\ell$$

is the marginal likelihood.

Computing $f(\boldsymbol{X} \mid \mathcal{M}_\ell)$ can be very time consuming, especially if the dimension of each $\boldsymbol{\psi}_\ell$ is large. The BDOGS method is computationally intensive and in particular requires model selection to be done repeatedly during simulation, hence requires $f(\mathcal{M}_\ell \mid \boldsymbol{X})$ to be computed many times. Thus, the method requires a fast method for calculating the posterior model probabilities in (8.1.1). In addition, during the design phase when simulations are being run in order to obtain the design's operating characteristics, the efficiency of the method used to compute the posterior model probabilities is critical. Consequently, to ensure feasibility BDOGS uses an approximation of the Bayes factor given in Raftery (1996b) to compute the posterior probability of each $\mathcal{M}_\ell$ given by (8.1.1). The Bayes factor for model $\mathcal{M}_\ell$ versus $\mathcal{M}_1$ is the ratio of the posterior to prior odds,

$$B_{\ell,1} = \frac{f(\mathcal{M}_\ell \mid \boldsymbol{X})/f(\mathcal{M}_1 \mid \boldsymbol{X})}{f(\mathcal{M}_\ell)/f(\mathcal{M}_1)} = \frac{f(\boldsymbol{X} \mid \mathcal{M}_\ell)}{f(\boldsymbol{X} \mid \mathcal{M}_1)}, \tag{8.1.2}$$

with $B_{1,1} = 1$. Let $\mathcal{L}_\ell(\boldsymbol{X} \mid \boldsymbol{\psi}_\ell, \mathcal{M}_\ell)$ denote the likelihood, $\chi^2 = 2[\log \mathcal{L}_\ell(\boldsymbol{X} \mid \widehat{\boldsymbol{\psi}_\ell}, \mathcal{M}_\ell) - \log \mathcal{L}_0(\boldsymbol{X} \mid \widehat{\boldsymbol{\psi}_1}, \mathcal{M}_1)]$, n the number of observations, $\widehat{\boldsymbol{\psi}_\ell}$ the MLE of $\boldsymbol{\psi}_\ell$ under $\mathcal{M}_\ell$, and $p_\ell = \dim(\boldsymbol{\psi}_\ell)$. Raftery (1996b) gives the approximation

$$2 \log B_{\ell,1} \approx \chi^2 - (p_\ell - p_1) \log n, \tag{8.1.3}$$

where the notation $a_n \approx b_n$ means that $\lim_{n\to\infty}(a_n/b_n) = 1$. To compute $f(\mathcal{M}_\ell \mid \boldsymbol{X})$, the posterior model probabilities in (8.1.1) can be expressed in terms of Bayes factors, and the approximation of Raftery (1996b) exploited to obtain $B_{\ell,1}$ for $\ell = 2, 3, \ldots, J$. Combining (8.1.1) and (8.1.2), the posterior probability of $\mathcal{M}_\ell$ is

$$f(\mathcal{M}_\ell \mid \boldsymbol{X}) = \frac{B_{\ell,1} \times \xi_\ell}{\sum_{r=1}^{J} B_{r,1} \times \xi_r}. \tag{8.1.4}$$

Substituting (8.1.3) into (8.1.4) gives an approximate value of $f(\mathscr{M}_\ell \mid \boldsymbol{X})$ for $\ell = 2,\ldots,J$. The main computational requirements are obtaining the MLEs of $\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \ldots, \boldsymbol{\psi}_M$ under their respective models. While Raftery provides other approximations which are more accurate, BDOGS uses the slightly less accurate approximation in (8.1.3) to gain speed. Still, as the numerical illustration given below will show, the advantages afforded by BDOGS compared to the conventional approach are substantial.

8.1.2.3 Decision Boundaries and Utility Function

The goal of the BDOGS design is to test $H_0 : \delta = 0$ versus $H_1 : \delta \neq 0$ subject to size and power constraints without requiring the often incorrect assumption of proportional hazards. The utility function employed by BDOGS is the achieved sample size, which is random due to the group sequential structure. The expected utility is computed as the equally weighted average of the sample sizes under the null and alternative hypothesis. For brevity, we omit additional details of the decision-theoretic framework utilized by BDOGS, which are given in section 2 of Wathen and Thall (2008).

A standard approach to obtain the optimal decision procedure in a decision-theoretic design is backward induction (BI) (DeGroot, 1970). However, computational difficulties in implementing BI impose severe practical limitations on Bayesian optimal designs. Consequently, most Bayesian optimal designs using BI assume simple models (Berry and Ho, 1988; Lewis and Berry, 1994).

Alternative approaches to fully sequential BI have been proposed by many authors (Stallard, Thall, and Whitehead, 1999; Kadane and Vlachos, 2002; Stallard, 2003; Christen et al., 2004; Wathen and Christen, 2006). Carlin, Kadane, and Gelfand (CKG) (1998) proposed forward simulation (FS) as a practical alternative to BI. With FS, an optimal design is obtained by first simulating the trial repeatedly and storing the results. A given sequential decision procedure is applied to each simulated data set, and expected utilities are computed empirically from the simulated data. Since the simulation results have been stored, different decision procedures may be evaluated and a suitable search algorithm can be used to find the decision procedure that maximizes the expected utility. Storing the simulated data and the results of any time-consuming calculations eliminates the need to re-simulate the trial. BDOGS utilizes FS to obtain the optimal boundaries under each candidate model, subject to the size and power constraints. Without employing FS and storing the results of time consuming calculations it would not be possible to implement BDOGS. Specific details of the algorithms used in BDOGS are provided in Wathen and Thall (2008).

To facilitate computation of the optimal boundaries, BDOGS defines the decision boundaries in terms of two flexible monotone functions, each having three parameters. Let $\boldsymbol{\gamma} = (a_U,\ b_U,\ a_L,\ b_L, c_U,\ c_L)$ denote the decision boundary parameter vector, $\boldsymbol{X}_n$ the data for the first n patients and let $\mathrm{N}^+(\boldsymbol{X}_n)$ the number of treatment failures (events) in $\boldsymbol{X}_n$. BDOGS defines the boundary functions as

$$P_U(\boldsymbol{X}_n, a_U, b_U, c_U) = a_U - b_U \left(\frac{\mathrm{N}^+(\boldsymbol{X}_n)}{N} \right)^{c_U},$$
$$P_L(\boldsymbol{X}_n, a_L, b_L, c_L) = a_L + b_L \left(\frac{\mathrm{N}^+(\boldsymbol{X}_n)}{N} \right)^{c_L}, \tag{8.1.5}$$

with the requirement that $P_L(\boldsymbol{X}_{n_k}, a_L, b_L, c_L) \leq P_U(\boldsymbol{X}_{n_k}, a_U, b_U, c_U)$. If $P_L(\boldsymbol{X}_n, a_L, b_L, c_L) > P_U(\boldsymbol{X}_n, a_U, b_U, c_U)$ then BDOGS sets $P_L(\boldsymbol{X}_n, a_L, b_L, c_L) = P_U(\boldsymbol{X}_n, a_U, b_U, c_U)$. In these boundary functions, a_U and a_L define the initial decision boundaries before any patients are enrolled, b_U and b_L determine the final boundaries when all events have been observed, and c_U and c_L determine the rate at which P_U decreases and P_L increases with the number of failure events.

Denote the Bayesian posterior decision criteria $p_{E>S}(\boldsymbol{X}_n) = Pr(\delta > \delta^* \mid \boldsymbol{X}_n)$ and $p_{S>E}(\boldsymbol{X}_n) = Pr(\delta < -\delta^* \mid \boldsymbol{X}_n)$. BDOGS uses the decision boundaries and the two posterior probabilities to determine if a trial should stop, either for futility or superiority, or continue. The trial is conducted as follows:

1. **Superiority**: (a) If $p_{E>S}(\boldsymbol{X}_{n_k}) > P_U(\boldsymbol{X}_n, a_U, b_U, c_U) > p_{S>E}(\boldsymbol{X}_n)$, then terminate the trial and select E. (b) If $p_{S>E}(\boldsymbol{X}_n) > P_U(\boldsymbol{X}_n, a_U, b_U, c_U) > p_{E>S}(\boldsymbol{X}_n)$, then stop the trial and select S.
2. **Futility**: If $\max\{p_{E>S}(\boldsymbol{X}_n), p_{S>E}(\boldsymbol{X}_n)\} < P_L(\boldsymbol{X}_n, a_L, b_L, c_L)$, then stop the trial and conclude that neither treatment is superior to the other.
3. **Continuation**: If either (a) $P_L(\boldsymbol{X}_n, a_L, b_L, c_L) \leq p_{E>S}(\boldsymbol{X}_n)$, $p_{E>S}(\boldsymbol{X}_n) \leq P_U(\boldsymbol{X}_n, a_U, b_U, c_U)$ or (b) $\min\{p_{E>S}(\boldsymbol{X}_n), p_{E>S}(\boldsymbol{X}_n)\} \geq P_U(\boldsymbol{X}_n, a_U, b_U, c_U)$, then continue enrolling patients.

Part (b) of rule 3 is included to deal with cases where $\mathrm{Var}(\delta \mid \boldsymbol{X}_{n_k})$ is large and both $p_{S>E}(\boldsymbol{X}_{n_k})$ and $p_{E>S}(\boldsymbol{X}_{n_k})$ are large, although in practice both values being $\geq P_U$ rarely occurs. At the final analysis, if the superiority rule does not apply for either treatment then BDOGS concludes $\delta = 0$.

BDOGS utilizes an algorithm for calculating and storing calculations in conjunction with an efficient search algorithm to determine the numerical values of $\boldsymbol{\gamma}$ that maximize the expected utility under each model. For example, if there are $J = 5$ models under consideration then BDOGS determines optimal boundaries under each of the 5 models, and thus there would be 5 distinct $\boldsymbol{\gamma}$ vectors, one for each potential model.

8.1.3 Application of BDOGS to the Lung Cancer Trial

In Section 8.1.3.1, we describe how to use an available Kaplan-Meier plot to identify potential hazard distributions which can be used in a simulation study. In Section 8.1.3.2 we provide specific details for applying BDOGS to the lung cancer trial, and Section 8.1.3.3 summarizes a simulation study comparing the BDOGS and OF designs.

In the original proposal for the lung cancer trial, as in many studies, there was uncertainty about the patient accrual rate, with the anticipated range from 2 to 4 patients per month. This resulted in a maximum sample size ranging from 106 to 122 patients for the OF design. Since maintaining the targeted power is very important, we chose to develop a BDOGS design with maximum sample size of 122 patients, and we evaluated both designs under an assumed accrual of 2 patients per month.

8.1.3.1 Using Kaplan-Meier Plots to Identify Potential Survival Distributions

A previous study provided a KM plot of the PFS time distribution for NSCLC patients given the standard SC treatment (Ciuleanu et al., 2008). Since the raw PFS data from this study were not available, we used the published KM plot to identify potential PFS time distributions. The Weibull and Log-Normal (LN) distributions both were considered to be reasonable possible models since both have flexible hazard functions. The Weibull distribution allows a wide variety of hazards, including a constant, increasing or decreasing hazard. The Log-Normal allows non-monotone hazards that may increase and then decrease, or decrease and then increase.

Denote T = PFS time and $S(t) = \Pr(T > t)$. We visually extracted the estimates $\widehat{S}_{KM}(t) = \widehat{\Pr}(T > t)$ from the KM curve at each time point $t \in \mathscr{T} = \{0.5, 1.5, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 18\}$. The corresponding estimates $\widehat{S}_{KM}(t)$ were $\{0.975, 0.95, 0.70, 0.60, 0.50, 0.43, 0.375, 0.30, 0.25, 0.20, 0.175, 0.10, 0.05\}$. The time points were chosen so that straight lines connecting $(t, \widehat{S}_{KM}(t))$ pairs for consecutive $t \in \mathscr{T}$ gave a close fit to the KM plot. To determine numerical values for each distribution that provided the best fit, we used the sum of squared differences between the survival probability function for the parametric distribution and the estimated survival probability function obtained from the KM curves. Denote the survival function at time t for a LN or Weibull distribution with parameters a and b by $S(t|a,b)$. For each distribution, we solved for its parameters by minimizing

$$D(a,b) = \sum_{t \in \mathscr{T}} \left\{ S(t|a,b) - \widehat{S}_{KM}(t) \right\}^2.$$

For each parametric distribution, we performed a grid search to determine the values of a and b minimizing $D(a,b)$. For the LN where $\log(\mathrm{T}) \sim \mathrm{N}(\mu, \sigma^2)$, for $a = \mu$ and $b = \sigma^2$ we evaluated $D(a,b)$ over the domain $a \in [-100, 100]$ and $b \in (0,\ 100]$ with initial grid size $\Delta = 0.50$. Once an initial minimum was found we repeated the search locally with $\Delta = 0.10$ and finally with $\Delta = 0.01$. This gave minimizing values $a = 1.41$ and $b = 0.95$ with $D(1.41, 0.95) = 0.018$. For the Weibull distribution given by $S(t|a,b) = \exp(-t^a/b)$, performing the grid search in the same way over $(a,b) \in (0,\ 100]^2$ gave $a = 1.18$ and $b = 5.94$ with minimum value $D(1.18, 5.94) = 0.024$, a 33% increase over the minimum value achieved by the LN. As a simple alternative to the Weibull and LN distributions, we also evaluated the Exponential distribution with mean a. For the exponential, $D(a)$ was minimized at $a = 5.99$ with minimum

value $D(5.99) = 0.035$. Thus, the LN gave the best fit. Plots of the corresponding survival and hazard functions are displayed in Figure 8.1.

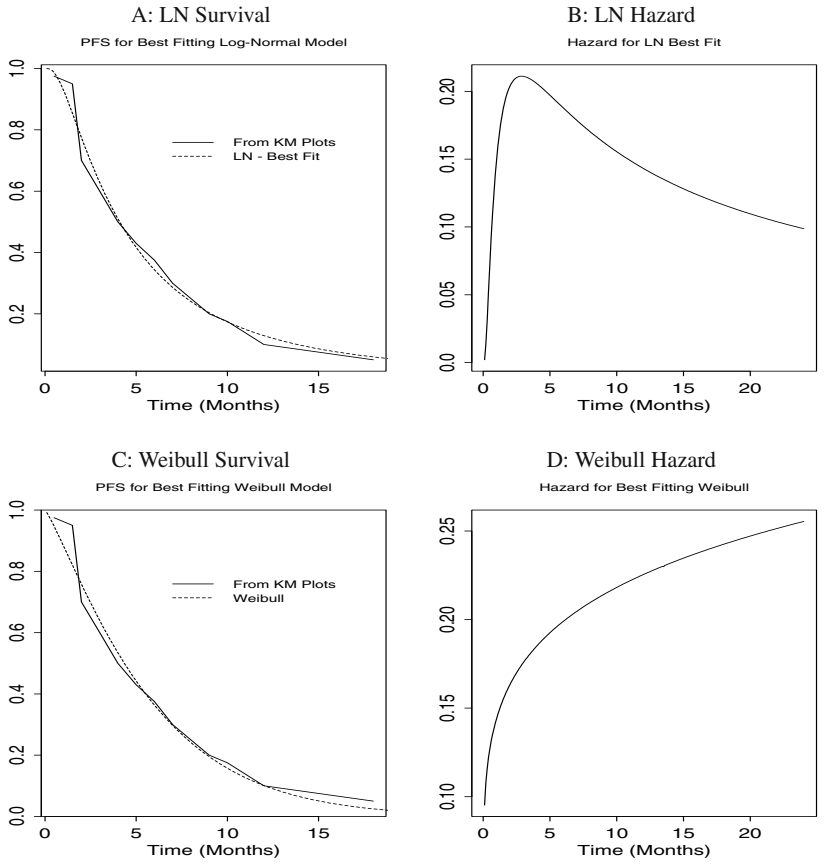


FIGURE 8.1. Survival functions and hazard functions for the best fitting Log-Normal (LN) distribution (A&B) and Weibull distribution (C&D). The solid lines in A & C represent the estimates obtained by visually extracting values from the available Kaplan-Meier plots.

8.1.3.2 A BDOGS Trial Design

By construction, BDOGS finds the optimal decision bounds of the form described in Section 8.1.2.3 under each of a variety of specified potential models, equivalently hazard functions. A key strength of the BDOGS procedure is its ability to adaptively switch decision boundaries based on the accruing information. In order to achieve a high degree of flexibility while still allowing implementation of the BDOGS method to be feasible, to design the NSCLC trial we included 5 potential

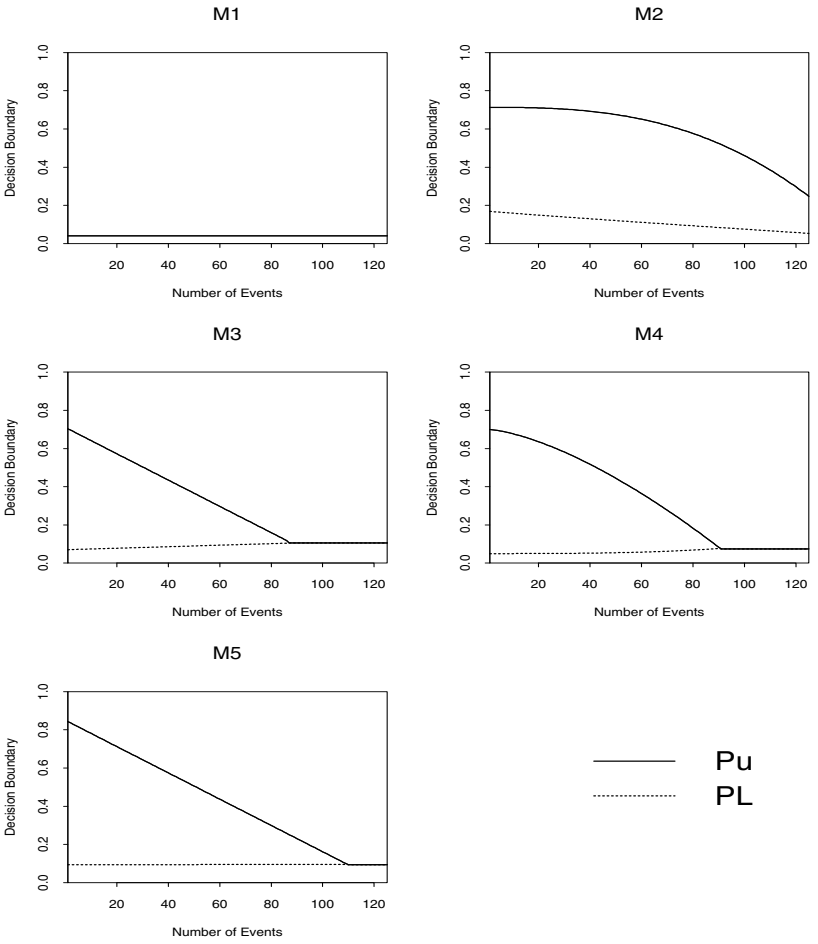


FIGURE 8.2. Stopping boundaries under the BDOGS design for each of the 5 candidate models, $\mathscr{M}_1$: increasing hazard, $\mathscr{M}_2$: decreasing hazard, $\mathscr{M}_3$: constant hazard, $\mathscr{M}_4$: initially increasing hazard followed by a slight decrease and $\mathscr{M}_5$: initially increasing hazard followed by a large decrease. The definitions of P_L and P_U are given in Section 8.1.2.3.

models, (1) $\mathscr{M}_1$: increasing hazard, (2) $\mathscr{M}_2$: decreasing hazard, (3) $\mathscr{M}_3$: constant hazard, (4) $\mathscr{M}_4$: initially increasing hazard followed by a slight decrease and (5) $\mathscr{M}_5$: initially increasing hazard followed by a large decrease. If significantly informative prior data were available, one could specify the prior model probabilities to reflect such knowledge. However, since we did not have sufficient information to assign different prior model probabilities, we took the conservative approach of assuming that the five models were equally likely, that is, $\mathrm{f}(\mathscr{M}_j) = 0.20$ for $j = 1, \ldots, 5$. Recall that with a BDOGS design the decisions are not based on Z-scores, but rather posterior probabilities using the decision boundaries given by (8.1.5). Since the de-

cision boundaries are model specific, we provide a graph of the boundaries under each model in Figure 8.2.

The initial OF design was proposed with up to two interim analyses and one final analysis. To accommodate a wider range of potential hazards and reduce the expected sample size, the BDOGS design was implemented allowing up to 6 interim analyses plus a final analysis. The analyses are conducted after observing 25, 50, 75, 87, 100, 112, and 122 events. Like the OF design, the BDOGS design has overall size 0.05 and power 0.90 to detect a 100% increase (doubling) in median PFS time from 4 to 8 months.

8.1.3.3 Simulations

To make the comparison of the OF and BDOGS designs fair, we increased the number of interim tests in the OF design to six in order to match the BDOGS design. In addition, both approaches monitored the trial at the the same interim numbers of events and had the same maximum sample size. Also to ensure fairness, for each iteration of the simulations we first generated the patient arrival times and simulated the event times under the assumed underlying distribution, and presented the two methods with the same patient data.

To assess robustness, we simulated the trial assuming seven different true PFS time distributions, defined in terms of their hazard functions. Aside from the case of a constant hazard, the hazard functions of the true distributions used in the simulations are plotted in Figure 8.3. The true hazard functions were the Log-Normal that provided the best fit to the KM plot (LN-Best Fit), the Weibull that provided the second best fit to the KM plot (W-Second Best Fit), a Weibull with decreasing hazard (WD), a Log-Normal with a hazard that increases then decreases slightly (LN-ID2), a Log-Normal with a hazard that increases followed by a large decrease (LN-ID3) and a Weibull with an increasing hazard (WI). The simulation results are summarized in Table 8.2 and 95% confidence intervals of the sample size distributions for BDOGS and OF under each true hazard are plotted in Figure 8.4.

When the true hazard was constant (Exp), both BDOGS and OF maintained the targeted false positive rate. However, in the null Exp case the mean sample was 53 for BDOGS compared to 80 for OF, and in the alternative Exp case when $\delta = 4$ was 65 for BDOGS compared to 87 for OF.

Under the distribution that provided the best fit to the historical KM curve, LN-BF, BDOGS had actual power 0.88 compared to 0.89 for OF. In terms of median sample size, BDOGS enrolled 39 patients in the null and 43 in the alternative LN-BF case, compared to 86 and 91 patients for OF. Thus, BDOGS provided over a 50% reduction in average sample size with a trivial drop in power. Similar results were obtained in the W-SBF case, where both BDOGS and OF maintained the desired false positive rate and power but BDOGS had a much smaller average sample size.

When the true hazard function is increasing followed by a substantial decrease (LN-ID3), OF has actual power 0.71 compared to 0.89 for BDOGS. The false-positive rate of BDOGS is 0.07. The average sample size for BDOGS is substantially

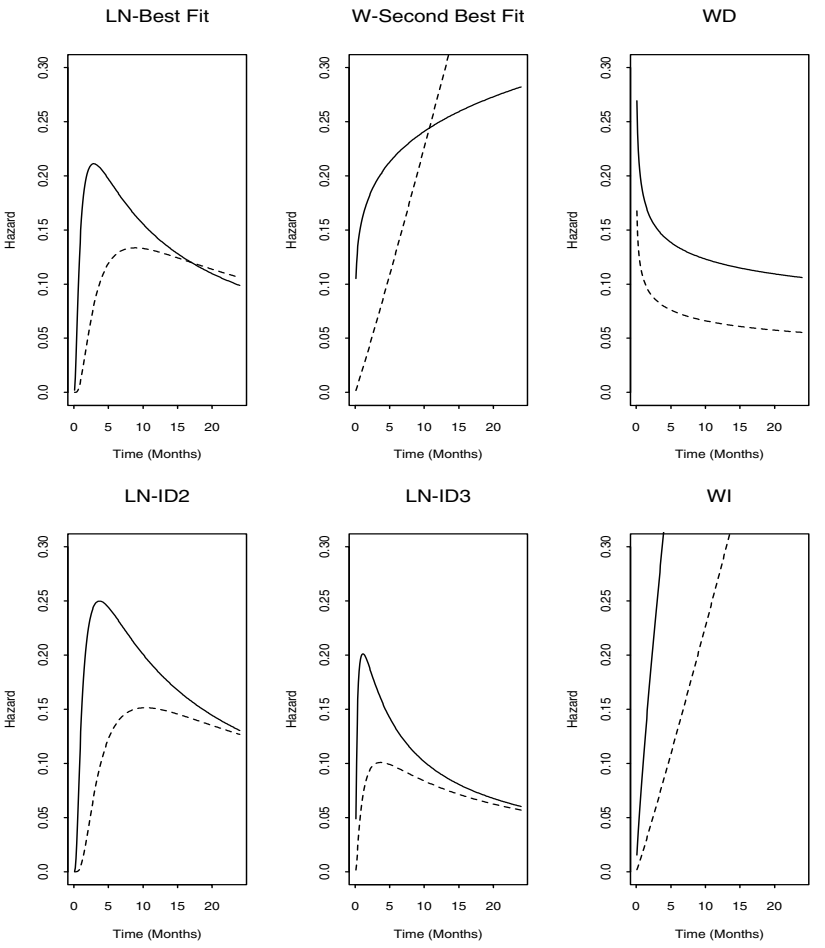


FIGURE 8.3. Hazard functions of the true distributions used in the simulation study: Log-Normal best fit (LN-Best Fit), Weibull second best fit (W-Second Best Fit), Weibull with a decreasing hazard (WD), Log-Normal with hazard that increases initially then decreases slightly (LN-ID2), Log-Normal with hazard that increases initially followed by a large decrease (LN-ID3), and Weibull with an increasing hazard (WI). The standard treatment (solid line) has median PFS 4 months. Under the alternative, the experimental treatment (dashed line) has median PFS 8 months.

smaller than that of OF, 66 compared to 85 under the null and 71 compared to 95 under the alternative. The main reason that BDOGS fails to maintain the desired size is that the initial model selection is performed after observing only 25 events. In general, if the first model selection in BDOGS is performed with a small amount of data the posterior model probabilities may favor incorrect models, resulting in an inflated false-positive rate.

TABLE 8.2. Simulation study to compare the Bayesian doubly optimal group sequential (BDOGS) design with the O'Brien-Fleming (OF) design. In all cases, the true null median PFS time is 4 months for the Standard treatment and the alternative PFS time is 8 months.

True		False		Sample Size, $\delta = 0$ ($\delta = 4$)					
Hazard	Method	Pos.	Power	Mean	2.5%	25%	50%	75%	97.5%
Proportional Hazards Assumption Met									
Exp	BDOGS	0.05	0.90	53(65)	30(35)	35(41)	58(63)	63(91)	93(117)
	OF	0.05	0.90	80(87)	56(61)	64(70)	85(91)	90(98)	111(118)
Robustness Study - Proportional Hazards Assumption Not Met									
LN-BF	BDOGS	0.05	0.88	49(54)	31(34)	35(39)	39(43)	62(67)	92(113)
	OF	0.05	0.89	81(86)	56(59)	65(69)	86(91)	91(98)	111(118)
W-SBF	BDOGS	0.02	0.90	47(44)	30(32)	34(37)	39(40)	60(44)	85(85)
	OF	0.05	0.90	79(80)	55(47)	63(64)	84(86)	89(93)	110(114)
WD	BDOGS	**0.07**	0.90	65(71)	31(36)	56(43)	63(68)	87(96)	114(122)
	OF	0.05	**0.81**	82(94)	58(63)	66(89)	87(96)	92(105)	113(122)
LN-ID2	BDOGS	0.04	0.90	41(46)	29(33)	33(38)	36(40)	41(45)	68(94)
	OF	0.05	0.97	80(77)	56(44)	64(64)	85(71)	89(91)	110(110)
LN-ID3	BDOGS	**0.07**	0.89	66(71)	31(36)	39(43)	64(68)	89(98)	122(122)
	OF	0.05	**0.71**	85(95)	59(63)	68(88)	90(97)	95(108)	116(122)
WI	BDOGS	0.01	0.91	33(39)	28(32)	32(36)	34(38)	36(41)	41(65)
	OF	0.01	0.99	78(63)	54(37)	61(60)	82(63)	87(66)	108(90)

When the true hazard is increasing (WI) then both BDOGS and OF maintain the desired size and power. However, BDOGS has a mean sample size that is less than half that of OF in the null case, 33 patients for BDOGS compared to 78 for OF. Under the alternative, BDOGS provides a reduction of 24 patients (39%) in mean sample size, from 63 to 39. The main reason for this substantial decrease in average sample size afforded by BDOGS is that, in this case, OF has a greatly inflated actual power of 0.99.

8.1.3.4 Logistics of Trial Conduct

With a conventional group sequential approach, interim analyses typically are conducted less frequently than with the BDOGS design. As the number of interim analyses increases, however, so does the logistical complexity of actually conducting the trial. This is an important issue in a multi-center trials. Logistical issues include the complexity and time involved to collect, clean, and prepare the data for analysis in a timely fashion. While such issues are a valid concern, the use of available technologies can greatly mitigate many potential problems. For example, for conducting the NSCLC trial a secure website has been developed that is similar to the ones utilized to conduct the adaptive trial described in Maki et al. (2007). The use of such a website greatly decreases the chance of errors in data collection, and also substantially reduces the logistical burden of collecting and tracking the trial data in real time.

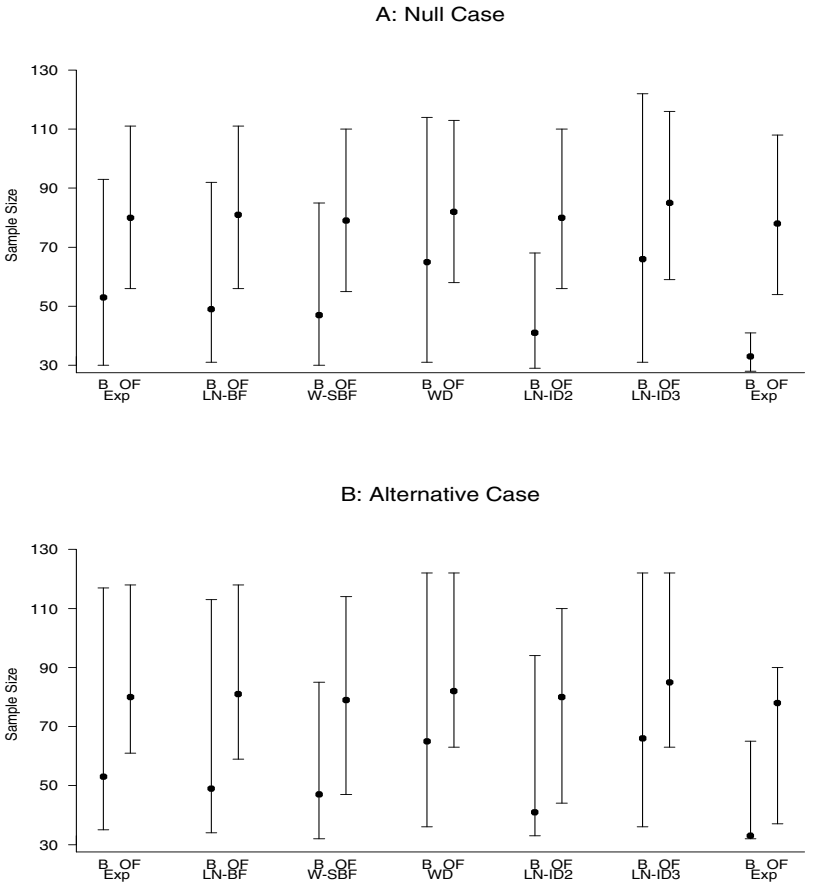


FIGURE 8.4. Sample size distributions for BDOGS (B) and O'Brien-Fleming (OF) designs under the A) Null case and B) Alternative case. For each line, the dot represents the mean and the endpoints are the 2.5% and 97.5% percentiles.

8.1.4 Discussion

BDOGS is a computationally intensive Bayesian methodology that provides a very attractive alternative to conventional group sequential designs for randomized clinical trials. Our simulation study in the context of the NSCLC trial illustrated the general facts that, compared to a conventional group sequential design using OF boundaries, BDOGS provides a substantial decrease in expected sample size while doing a better job of maintaining the targeted power. Under the distribution that best fit the historical data, BDOGS provided nearly a 40% reduction in expected sample

size while obtaining actual size and power that were comparable to the standard approach. A more extensive simulation study given by Wathen and Thall (2008) that included three different conventional group sequential designs gave very similar results.

The biggest potential problem with BDOGS arises if the first model selection is done too early in the trial. With very little data, the performance and accuracy of the model selection is limited and can increase the false positive rate. Under both the WD and LN-ID3 distributions, the actual size of the BDOGS design was 0.07, while in both cases the OF method maintained size 0.05. However, in these cases the OF method's power figures were greatly degraded from the nominal 0.90 to 0.81 under WD and 0.71 under LN-ID3, and moreover the sample sizes were much larger with OF compared to BDOGS. Recall that the first test is conducted when 25 events are observed. This was specified since the possibility of a low accrual rate was a concern, and an accrual rate of 2 patients per month would lead to the first test being conducted one year into the trial. While conducting the first test later would increase the probability of selecting the correct model at that look and thus reduce the overall false positive probability, it also would increase the sample size. Thus, the small increase in overall false positive rate with BDOGS seen in some cases may be viewed as an acceptable trade-off for the large decrease in expected sample size and greater power, compared to conventional methods, under most distributions.

A computer program, "BDOGS," for simulating the BDOGS method is available upon request from the first author.

8.2 Experimental Design and Sample Size Computations for Longitudinal Models

Robert E. Weiss and Yan Wang

In many studies the goal is to test a point null hypothesis against an alternative hypothesis. Longitudinal designs are extremely common, but appropriate methods for sample size specification are rare; exceptions are Lui and Cumberland (1992), Muller et al. (1992), Liu and Liang (1997), Hedeker, Gibbons, and Waternaux (1999), and Roy et al. (2007). Uncertainty in unknown parameters is not propagated. Previous Bayesian sample size methodology includes Spiegelhalter and Freedman (1986), Adcock (1997), Müller and Parmigiani (1995), Wang and Gelfand (2002), and De Santis (2007). In longitudinal data, a key additional issue includes the number and spacing of longitudinal measures.

The Bayesian tool for hypothesis testing is the Bayes factor (Berger, 1985). Weiss (1997) introduced the idea of choosing the sample size to guarantee that the Bayes factor is larger than a certain prespecified size. Let Y denote the data, let X be the covariates, also let H_k for $k = 0, 1$ be competing hypotheses with parameters θ_k, sampling distributions $f(Y|\theta_k, X, H_k)$ and priors $p(\theta_k|H_k)$. The prior predictive dis-

tributions of the data are $f(Y|H_k) = \int f(Y|\theta_k, X, H_k) p(\theta_k|H_k) d\theta_k$. The Bayes factor B_{01} for H_0 against H_1 is

$$B_{01} = \frac{f(Y|H_0)}{f(Y|H_1)}$$

and define the log Bayes factor $b_{01} = \log B_{01}$. Kass and Raftery (1995) suggest b_{01} greater than $+3$ or less than -3 constitute *strong* support for or against H_0, respectively; and $b_{01} > 5$ or $b_{01} < -5$ is said to constitute *very strong* support for H_0 or H_1, respectively.

Our goal is to select the sample size n and other design aspects so that the prior predictive probabilities $p_0(a_0) = P(b_{01} > a_0|H_0)$ and/or $p_1(-a_1) = P(b_{01} < -a_1|H_1)$ are suitably large for some $a_0, a_1 \geq 0$. We explore choices of $a_0 = a_1$ equal to 3 or 5. We use a combination of Monte Carlo simulation, algebraic calculations, and numerical integration to calculate $p(b_{01}|H_k)$, $k = 0, 1$ using a predictive prior (Weiss, Wang, and Ibrahim, 1997) based on data from a previous experiment. Section 8.2.1 discusses modeling issues. Section 8.2.2 outlines a general algorithm to simulate the distributions $p(b_{01}|H_k)$, $k = 0, 1$. Section 8.2.3 illustrates the problem of choosing a sample size for a complicated hierarchical repeated measures data random effects model based on a prior study. An expanded version of this work with formal proofs and computational formulas is available from the first author.

8.2.1 Covariates and Missing Data

The sampling density $f(Y|\theta_k, X, H_k)$ depends on covariates X whose distribution is uncertain except in special cases. Specifying known X is inappropriate because it underestimates uncertainty. Parametrically, we can model a sampling density $g(X|\phi)$ for covariates as depending on unknown parameters ϕ with prior $q(\phi)$. A randomized binary treatment indicator can be modeled as a Bernoulli random variable with probability of success equal to $\pi = 0.5$. When a prior sample x_j, $j = 1, \ldots, J$ of continuous m-dimensional covariates is available, a kernel density estimator may be used to estimate the distribution of the covariates. Under independence between X, ϕ and the hypotheses k and θ_k, then the Bayes factor B_{01} calculation does not involve $g(X|\phi)$ or $q(\phi)$. Longitudinal studies routinely feature missing data, as in the example reported below. Let $Y_{\text{obs}} \subset Y$ denote the observed subset of the complete data. Assuming the data are missing at random (MAR) (Little and Rubin, 2002), then the Bayes factor does not depend on the missingness pattern nor on the underlying parameters that control the missingness pattern.

8.2.2 Simulating the Predictive Distributions of the Bayes Factor

After specifying the necessary distributions, simulating $p(b_{01}|H_k)$ for a given sample size L_k involves the following steps. For each of $l \in \{1, 2, \ldots, L_k\}$ repeat

1. Sample $\phi^{(l)}$, the parameters of the X distribution from $q(\phi)$.
2. Sample the covariates $X^{(l)}$ from $g(X|\phi^{(l)})$.
3. Sample $\psi^{(l)}$, the parameters of the missingness distribution from its distribution $r(\psi)$.
4. Sample the missingness indicator variable $R^{(l)}$ from its distribution $p(R|\psi^{(l)}, X^{(l)})$. The variable R is a data structure of 0-1 indicator variables, 1 for missing, 0 for observed, with the same structure as the complete data vector Y.
5. Sample the unknown parameters $\theta_k^{(l)}$ from the prior $p(\theta|H_k)$.
6. Sample the data $Y_k^{(l)}$ from $f(Y|X^{(l)}, \theta_k^{(l)}, R^{(l)}, H_k)$.
7. Calculate the Bayes factor $b_{01,k}^{(l)} = \log[f(Y_{\text{obs},k}^{(l)}|H_0)/f(Y_{\text{obs},k}^{(l)}|H_1)]$.

It is usually possible to sample only the observed portion of the data vector Y_{obs} rather than the entire Y vector. It is also possible to reduce computations by reusing $\phi^{(l)}$, $X^{(l)}$, $\psi^{(l)}$, and $R^{(l)}$ for simulations under both H_0 and H_1 and for different designs. Having calculated samples from $p(b_{01}|H_k)$, we estimate the sum $P(b_{01} > a_0|H_0) + P(b_{01} < -a_1|H_1)$ or other summary statistic as desired, depending on the specific utility function used in designing the study.

For early calculations, we take L_k small, such as 100, which gives a standard error of approximately $0.05 = (0.5^2/100)^{1/2}$ for the estimated probability of interest. We increase L_k after tentatively bracketing the needed sample size.

8.2.3 Sample Size for a New Repeated Measures Pediatric Pain Study

We design a followup study to a previous longitudinal pediatric pain study (Weiss, 2005). The outcome is the log of the time in seconds that a child could keep his or her hand immersed in cold water, a proxy for pain tolerance. The main analysis had two important covariates, Coping Style (CS) and Treatment (TMT). Children's coping style can be attender (A) or distracter (D) with attenders paying attention to their arm and the experimental apparatus during each trial while distracters think about other things such as vacation or schoolwork. TMT was randomized to one of three counseling interventions: counseling to attend (A), to distract (D), or a null treatment (N).

If treatment has an effect, then CS and TMT are thought to interact. CS is observed and is not under the direct control of the investigator. The hypotheses are H_0: no treatment effect and H_1: a treatment effect which may be different for attenders and distracters. Weiss, Wang, and Ibrahim (1997) used a predictive prior and the 58 complete data cases to indicate that the data strongly supported H_0 against H_1. This was somewhat surprising, given that other analyses (Fanurik et al., 1993; Weiss, 1994) supported H_1. Thus there is interest in designing a followup study to discover whether the treatment intervention does indeed have an effect on pain tolerance. In the followup study we eliminate the N treatment.

8.2.3.1 Design Considerations

The previous study design had $m_1 = 3$ baseline observations followed by the counseling intervention, and then $m_2 = 1$ response observations for a total of 4 repeated measures. The four trials of the original study were given on two days, approximately two weeks apart. No effect due to practice, i.e., a time trend, or due to day has been found. In the followup, it is planned to have three trials per day with the intervention taking place before the fourth or fifth trial. Efficiency considerations might put the intervention before the third trial, but this is problematic, since the effect of the two week break between trial 3 and 4 on the intervention efficacy is unknown, is not of interest, and is unlikely to be ignorable. As long as the intervention takes place on day two, we are comfortable assuming that the mean structure is the same across trials before intervention, changes because of the intervention, and then remains constant again. The most reasonable design for the followup study will have $m_1 = 3$ pre-treatment trials and $m_2 = 3$ post-treatment trials; we investigate the effects of taking $m_1 = 4$ and $m_2 = 2$ and $m_1 = 2$ and $m_2 = 4$ as a form of sensitivity analysis. We call these designs the 3-3 design, the 4-2 design, and the 2-4 design.

For the new study design, we explore the ability to generate Bayes factors greater than 3 and 5 over a range of possible sample sizes. We propose a logistic regression methodology to estimate the utility of various intermediate sample sizes.

8.2.3.2 Sampling Density for Y and Prior Density for Parameters

The sampling distribution for the n_i vector of observations Y_i for a single case indexed by i is modeled using the random effects model

$$Y_i = X_i\alpha + Z_i\beta_i + \varepsilon_i, \tag{8.2.1}$$

where random effect $\beta_i \sim N_q(0, \sigma^2 D)$ and residual vector $\varepsilon_i \sim N_{n_i}(0, \sigma^2 I)$. The design matrix X_i for a completely observed case for the original study under H_1 is 4×8, with a column of ones for the intercept, a time-fixed indicator column of all zeros or ones for the effect of CS, and a 4×6 block of zeros, except for a single one in the fourth row to indicate which of the 6 TMT*CS groups the child belonged to. The Z_i matrices are n_i columns of ones in both the original and followup studies under both H_0 and H_1. Missing data within a case will cause rows of Y_i and corresponding rows of X_i and Z_i to be omitted. Under H_0, the X_i matrix is 4×2 with columns for the intercept and CS only; all columns for the treatment effect are omitted.

The pre-prior for the prior data is a flat prior, $p_0(\alpha, \sigma^2, D) \propto 1$, which will produce a proper posterior (Hobert and Casella, 1996). Data Y_{old} from all 64 children in the original study were used with this pre-prior to produce a prior $p(\alpha, \sigma^2, D | Y_{\text{old}}, H_k)$ for $k = 0, 1$ to design the future study. We made one modification to this prior. The prior for D was taken to be a gamma(c_0, c_1), where the mean c_0/c_1 was set equal to the sample mean of the Gibbs sample for D from the prior, and the variance c_0/c_1^2 was set equal to the variance of the Gibbs sample of D.

For the followup study, the two columns of X_i and the corresponding elements of α that refer to the N treatment are omitted. The length of Y_i will be 6 in the absence of missing data. The X_i matrix will have an initial column of ones and a second column of ones or zeros coding the CS value. Under H_1, the third through sixth columns will be all zeros except for m_2 ones in the last m_2 rows of whichever column of X_i is the indicator variable of the CS*TMT group that the case belongs to.

For brevity, technical details of the prior and Bayes factor calculation are given in the associated technical report. With some modifications, the prior follows the methods of Weiss, Wang and Ibrahim (1997), hereafter WWI. The data set that is used to form our prior is the data set analyzed in WWI. In WWI, the prior for σ^2 does not depend on D while our prior for σ^2 is directly a predictive prior for σ^2 derived from the prior data and dependent on D. Finally, our prior for D is based directly on the prior data as already described.

8.2.3.3 Covariate and Missing Data Distributions

There are two covariates in the study to be designed: CS and TMT, both binary. Of 64 children in the original study, 32 were observed to be distracters, and 32 were attenders. Starting with a uniform prior Beta$(1,1)$ for π_{CS}, the probability that a new child is a distracter, we will sample π_{CS} as a Beta$(33,33)$, and then for the n individuals in the followup study we sample CS_i as Bernoulli(π_{CS}). If the followup study were to be the same as the original, then the distribution of TMT is known to be multinomial$(1/3,1/3,1/3)$. Since we are eliminating the N treatment group, TMT is Bernoulli$(1/2)$.

In the original study the design called for 4 repeated measures per child for a total of $64 \times 4 = 256$ observations on $n = 64$ cases. However, 11 observations were missing on 6 children. Missingness appeared to be MCAR, related to things like school absence or illness and not to an inability to follow instructions or feelings about the experiment. A simple missingness model is one possibility, where π_{mis} the probability that an observation is missing has prior probability density Beta$(11+1, 256-11+1)$, and every observation might be deleted at random with probability $\pi_{\text{mis,obs}}$. This simple model violates prior knowledge plus missing observations tended to cluster; 11 observations on only 6 children is unlikely to occur by chance if observations are randomly missing.

The method we actually used was to consider that children divide into two groups, *missers* and *non-missers*. Non-missers never have missing data, while each observation on a misser is missing with probability π_{within}. The probability that a child is a misser is π_{misser}. The probability that a misser has no missing data is $(1-\pi_{\text{within}})^{n_i}$, where n_i is the designed number of observations for the child. For the prior study, $n_i \equiv 4$, and for the study under design, $n_i \equiv 6$. The prior data has 58 complete data cases, and 6 missers with a total of 11 missing observations. Including one extra unknown, l_0, the number of missers with zero missing observations allows us to produce a simple Gibbs sampler to draw samples from the posterior of

π_{within} and π_{misser}. Assuming flat Beta$(1,1)$ priors, the resulting posterior mean and standard deviation are 0.188 and 0.086 for π_{misser} and 0.226 and 0.085 for π_{within}.

8.2.3.4 Smoothing Simulation Results

Simulation sample sizes are never as large as we might desire, it is helpful to borrow strength from different simulations to estimate the design characteristics for different sample sizes. We did this by fitting a logistic regression model to the raw results (not shown) for each calculation such as $P(b_{01} > 3|H_0)$. For example, for the 3-3 design, and for the probability $P(b_{01} > 3|H_0)$ there were 7 simulation data points, with sample sizes $n = (20, 30, 37, 38, 39, 40, 50)$, and observed successes m_i equal to the number of times that $b_{01} > 3$ — these values were $m = (67, 71, 81, 78, 79, 88, 88)$, respectively, for the listed sample sizes. Then $m_i \sim \text{binomial}(L_i, \pi_i)$ with L_i equal to the number of simulations, usually 100, but for $n_i = 38, 39$, L_i was 1200 and 800. We modeled $\log(\pi_i/(1-\pi_i)) = \beta_0 + \beta_1 n_i$. The probabilities within a column do not vary greatly and a linear logistic regression can be expected to do a good job of interpolating and smoothing the results of the study. The resulting estimated values are given in Table 8.3 and substantially increase the effective sample size of most results often by a factor of 3 or 4.

8.2.3.5 Results

We simulated distributions $p(b_{01}|H_k)$ for the 3-3, 2-4, and 4-2 designs. Generally we increased n in steps of 10 starting from $n = 20$, searching for the point where the power was equal to 0.8. For the 3-3 design, we also investigated more carefully the power for sample sizes $n \in (37, 38, 39)$. Detailed results on the 4-2 design are omitted.

Results are reported in Table 8.3. The first column gives the sample size n, the second column is the simulation sample size. Column 3 gives $\psi_{0.05}(n)$, the lower 5% tail of the distribution $p(b_{01}|H_0)$ and column 4 is the power $P(b_{01} > \psi_{0.05}(n)|H_1)$, which is the probability under H_1 that b_{01} is greater than the cutoff in column 3. Columns 5-7 give the probabilities that the log Bayes factor is greater than 5, 3, and 0 if H_0 is true, and columns 8-10 give the probabilities that b_{01} is less than -5, -3, and 0 given that H_1 is true.

For the 3-3 design, the power achieves a level of 0.8 for a sample of size $n = 39$. We might expect the cutoff points $\psi_{0.05}(n)$ to be monotone in n, however, they are not. This is because of sampling variability in the calculations. The standard error of estimation of $\psi_{0.05}(n)$ from a sample of size 100 is approximately 0.9, and for a sample of size 1000, the standard error is approximately 0.3, so none of the inversions or equalities are too surprising. For the 2-4 design we appear to need less than 40 observations. The 4-2 design does appear to have less power than the other two, while the 3-3 and 2-4 are close in power.

TABLE 8.3. Smoothed results for the 3-3 and 2-4 designs.

3-3 design									
				$P(b_{01} > a\|H_0)$			$P(b_{01} < -a\|H_1)$		
n	L	$\psi_{0.05}(n)$	power	$a=5$	$a=3$	$a=0$	$a=5$	$a=3$	$a=0$
20	100	-2.00	0.40	0.29	0.64	0.89	0.25	0.35	0.61
30	100	-1.00	0.63	0.36	0.73	0.94	0.31	0.44	0.68
37	100	0.40	0.77	0.42	0.78	0.97	0.36	0.50	0.73
38	100	0.75	0.78	0.43	0.79	0.97	0.37	0.51	0.74
39	100	0.50	0.80	0.44	0.79	0.97	0.38	0.52	0.74
40	100	1.13	0.81	0.45	0.80	0.97	0.38	0.53	0.75
50	100	1.50	0.92	0.54	0.86	0.99	0.46	0.61	0.80
2-4 design									
				$P(b_{01} > a\|H_0)$			$P(b_{01} < -a\|H_1)$		
n	L	$\psi_{0.05}(n)$	power	$a=5$	$a=3$	$a=0$	$a=5$	$a=3$	$a=0$
20	100	-0.05	0.61	0.21	0.65	0.93	0.19	0.30	0.61
30	100	0.70	0.74	0.32	0.76	0.96	0.27	0.39	0.67
40	100	0.70	0.84	0.45	0.84	0.98	0.36	0.49	0.72
50	100	2.80	0.90	0.59	0.89	0.99	0.47	0.59	0.77

Inspection of Table 8.3 suggests that if we wanted to make $P(b_{01} > 3|H_0) + P(b_{01} < -3|H_1) \geq 1$, then the necessary sample size appears to be barely over 20 for the 3-3 and 2-4 designs, but will be close to 30 for the 4-2 design. If we try for $P(b_{01} > 5|H_0) + P(b_{01} < -5|H_1) \geq 1$, then we need an n of slightly over 50 for the 3-3 design, slightly under 50 for the 2-4 design, and approximately 60 for the 4-2 design.

Interpolation of Table 8.3 suggests that we need 39 cases for the 3-3 design to have power of 0.8, 34 cases for the 2-4 design and (not shown) fully 53 cases for the 4-2 design. The sample sizes needed to produce $P(b_{01} > 3|H_0) + P(b_{01} < -3|H_1) \geq 1$ are 21, 23, and 32 for the 3-3, 2-4, and 4-2 designs. Sample sizes to give $P(b_{01} > 5|H_0) + P(b_{01} < -5|H_1) \geq 1$ are 50, 48, and 60.

Approximately sixty children are available for the followup study. Sample size selection is driven by power, cost, subject availability, and other considerations. Our analysis shows we have sufficient power and reasonable probability of determining which hypothesis is correct.

Acknowledgments: This work was supported by grants GMS50011 and AI28697 from the National Institutes of General Medical Sciences and Allergy and Infectious Diseases of the NIH.

8.3 A Bayes Rule for Subgroup Reporting

Peter Müller, Siva Sivaganesan, and Purushottam W. Laud

Randomized clinical trials are carried out to establish the effectiveness of a treatment in a specified patient population. When a trial fails to show effectiveness for the original target population, the search for subgroups is a natural followup question. In general, subgroup analysis investigates whether a conclusion about effectiveness or lack of effectiveness in the overall population remains valid for subpopulations. Subpopulations are characterized by covariates, including age, prior treatment history, different disease subtypes, biomarkers, etc.

The main challenges related to subgroup analysis are concerns about data dredging and multiplicity issues when many potential subgroups are available and the search for subgroups is carried out in an unplanned fashion. Recent discussions of subgroup analysis appear among others in Cook, Gebski, and Keech (2004), Pocock et al. (2002), and Rothwell (2005). In this discussion we build on Sivaganesan, Laud, and Müller (2009) who introduced a Bayesian approach to subgroup analysis based on model selection. Let x_i, $i = 1,\ldots,I$, denote the list of reported covariates. For each covariate x_i they consider the family $\mathscr{M}_i$ of all subgroup models that can be described by x_i and define a probability models $p_i(M), M \in \mathscr{M}_i$. Posterior probabilities $p_i(M \mid data)$ are used to define an algorithm for subgroup reporting.

In this section we approach subgroup analysis as a Bayesian decision problem. In Section 8.3.1 we lay out the problem and introduce notation. In Section 8.3.2 we propose a Bayes rule for subgroup reporting. We use a simple extension of a 0/c utility function for hypothesis testing. We propose a few pragmatic simplifications to facilitate practical implementation. The resulting rule can be described in terms of posterior odds for subgroup models compared to the overall null model M_0 of no treatment effects and compared to the overall alternative M_1 of a common treatment effect in the entire patient population.

8.3.1 The Model Space

We consider a 2-arm clinical trial. Let y denote the outcome observed on each patient. Let θ generically denote a treatment effect. In a 2-arm trial with continuous outcomes, θ is the difference in mean outcomes across the two treatment arms. For a binary response $y \in \{0,1\}$, the treatment effect could be the difference in success probabilities under the two treatment arms, etc. The setup of the 2-arm trial is for ease of presentation only. The following discussion remains almost unchanged with any other sampling model or other designs.

We assume that the trial also records baseline covariates x_i, $i = 1,\ldots,I$ for each patient. Throughout we will use i to index covariates (not patients – we need no index for patients). For the moment we assume that all covariates are binary, $x_i \in$

$\{0,1\}$. The covariates define possible subgroups of patients. Each covariate defines five alternative models by the patterns of how treatment effects could differ across levels of x_i. Let

$$\gamma_i = (\gamma_{i0}, \gamma_{i1}) \in \{(0,0),(0,1),(1,0),(1,2),(1,1)\}$$

indicate these five patterns of subgroup effects, with $\gamma_{ij} = 0$ indicating no treatment effect for the subpopulation defined by $x_i = j$, and $\gamma_{ij} = 1$ and 2 denoting the first and second distinct non-zero treatment effect. For example, $\gamma_i = (1,2)$ indicates non-zero treatment effects for all patients, but with a different size effect for subgroups defined by $x_i = 0$ versus $x_i = 1$, and $\gamma_i = (1,0)$ indicates no treatment effect for patients with $x_i = 1$, and a non-zero treatment effect for patients with $x_i = 0$. We use $M_{i\gamma_i}$ to denote a model with treatment effects grouped as indicated by γ_i. The models indexed by $\gamma_i = (0,0)$ and $\gamma_i = (1,1)$ are special. The model $M_0 \equiv M_{i,(0,0)}$ is the overall null hypothesis of no treatment effect. The model $M_1 \equiv M_{i,(1,1)}$ is the overall alternative hypothesis of a common treatment effect for the entire population. Let $\Gamma = \{(0,1),(1,0),(1,2)\}$ denote the possible subgroup arrangements distinct from M_0 and M_1. In summary the family of all possible models is

$$\mathscr{M} = \{M_0, M_1, M_{i\gamma_i},\ i = 1,\ldots,I,\ \gamma_i \in \Gamma\}.$$

Conditional on M and the covariates we assume a sampling model for the observed outcomes, generically $p(y \mid M, x)$. Details of the sampling model are not required for the upcoming discussion.

Subgroup analysis refers to inference about $M \in \mathscr{M}$. In the following section we describe a Bayes rule, and a simplified approximate Bayes rule to implement inference about M. In the upcoming discussion we need to assume that the model includes a probability model $p(M)$ over the model space. We will define a specific model $p(M)$ later, in Section 8.3.3, only. But an important assumption implicit in treating M as a random variable is that only one subgroup model M is correct.

8.3.2 Subgroup Selection as a Decision Problem

8.3.2.1 A Bayes Rule for Reporting Subgroups

The problem of reporting subgroups is naturally described as a Bayesian decision problem. Let δ denote the desired decision. Possible decisions are to report an overall treatment effect, $\delta = M_1$; to report no evidence for any treatment effects, $\delta = M_0$; or to report subgroup effects. When reporting subgroup effects we allow to report multiple subgroups, one for each covariate i. Thus reporting subgroups involves the identification of a set of covariates, $A_I \equiv \{i_1,\ldots,i_m\} \subset \{1,\ldots,I\}$ together with a subgroup model index $\gamma_i \in \Gamma$ for each covariate. Let $\Gamma_I = \{\gamma_i,\ i \in A_I\}$ and let $A = (A_I,\ \Gamma_I)$ denote the pair of covariate indices and list of subgroups for each chosen covariate. Reporting subgroups is thus characterized as $\delta = A$. In summary

$$\delta \in \{M_0, M_1, A\} \equiv \mathscr{D} \text{ with } A = (A_I,\, \Gamma_I).$$

Note that the action space $\mathscr{D}$ differs from the model space $\mathscr{M}$ because we allow to report multiple subgroups simultaneously, but the probability model $p(M)$ allows only for one true subgroup model at a time.

A utility function $u(\delta, M, y)$ represents the investigator's relative preferences for the alternative actions under an assumed true model M and data y. Let $n_A = |A_I|$ denote the number of reported subgroups when $\delta = A$. We assume a natural generalization of a traditional 0/c utility function for testing problems:

$$u(\delta, M, y) = \begin{cases} u_0\, I(M = M_0) & \text{if } \delta = M_0, \\ u_1\, I(M = M_1) & \text{if } \delta = M_1, \\ u_2\, I(M \in \Gamma_I) - (n_A - 1) & \text{if } \delta = A. \end{cases} \tag{8.3.1}$$

In short, we realize a reward when the correct model is reported, and we pay a price for reporting more than one subgroup. Adding 1 in each row to include a price for reporting M_0 and M_1 as well would only shift the utility function by 1 and leave the optimal decision unchanged. Like in many decision problems, the specific choice of u includes some arbitrary and simplifying choices. In particular, we assume that the data enter the utility function only indirectly through the decision rule $\delta = \delta(y)$. This is typical for inference problems.

Let $p(M)$ denote a probability model over the model space and let

$$U(\delta, y) \propto \sum u(\delta, M, y)\, p(M \mid y)$$

denote the posterior expected utility. The optimal decision δ^* is the action with maximum expected utility. It is easy to show that the optimal decision under (8.3.1) can be characterized as follows. Assume $\delta^*(y) = A^*$. If $i \in A_I^*$, i.e., we report subgroups for covariate i, then the reported subgroups for i are the subgroups with highest posterior probability,

$$\gamma_i^* = \arg\max_{\gamma \in \Gamma} \{p(M_{i\gamma} \mid y)\},$$

and

$$A_I^* = \{i :\ p(M_{i\gamma_i^*} \mid y) \geq 1/u_2\},$$

i.e., we report subgroups for all covariates that include a model $M_{i\gamma_i^*}$ with posterior probability greater than $1/u_2$. Let $\Gamma_I^* = \{\gamma_i^*,\, i \in A_I^*\}$. In summary, if $\delta^* \notin \{M_0, M_1\}$ then it must be $A^* = (A_I^*, \Gamma_I^*)$. Thus, to determine the Bayes rule δ^* we only need to compare expected utilities for $\delta = M_0$, M_1 and A^*:

$$U(\delta, y) = \begin{cases} u_0\, p(M_0 \mid y) & \text{if } \delta = M_0, \\ u_1\, p(M_1 \mid y) & \text{if } \delta = M_1, \\ u_2 \sum_{i \in A_I^*} p(M_i \gamma_{*i} \mid y) - (n_{A^*} - 1) & \text{if } \delta = A^*. \end{cases}$$

We can now see the Bayes rule. Write $\overline{p}(M)$ as short for $p(M \mid y)$, let $M_i^* = M_i \gamma_i^*$ denote the maximum posterior subgroup model with covariate x_i, let $M^* =$

$\arg\max\{\overline{p}(M_i^*)\}$ denote the highest posterior probability subgroup model and let $i^* = \arg\max_i \overline{p}(M_i^*)$ denote the index of the covariate that defines M^*. Then

$$\delta^* = \begin{cases} M_1 & \text{if } \frac{\overline{p}(M_1)}{\overline{p}(M_0)} > \frac{u_0}{u_1} \text{ and } \frac{\overline{p}(M_1)}{\overline{p}(M^*)} > \frac{u_2}{u_1} + \sum_{A_I^* \setminus i^*} \frac{u_2 \overline{p}(M_i^*) - 1}{u_1 \overline{p}(M^*)}, \\ A^* & \text{if } \frac{\overline{p}(M_1)}{\overline{p}(M^*)} < \frac{u_2}{u_1} + \sum_{A_I^* \setminus i^*} \frac{u_2 \overline{p}(M_i^*) - 1}{u_1 \overline{p}(M^*)} \text{ and } \frac{\overline{p}(M_0)}{\overline{p}(M^*)} < \frac{u_2}{u_0} + \sum_{A_I^* \setminus i^*} \frac{u_2 \overline{p}(M_i^*) - 1}{u_0 \overline{p}(M^*)}, \\ M_0 & \text{otherwise.} \end{cases}$$

Note that $u_2\, \overline{p}(M_i^*) > 1$ for all $i \in A_I^*$, i.e., the terms in the sums are all strictly positive (although some can be very small).

8.3.2.2 Simplified Rule

In the interest of ease of implementation we now depart from a strictly decision theoretic implementation, and take the form of the Bayes rule δ^* as a motivation for a slightly simplified rule. We drop the sum over $A^* \setminus i^*$ in the conditions for reporting A^*, i.e., we are slightly more conservative in reporting subgroups. Then

$$\delta^* = \begin{cases} M_1 & \text{if } \frac{\overline{p}(M_1)}{\overline{p}(M_0)} > \frac{u_0}{u_1} \text{ and } \frac{\overline{p}(M_1)}{\overline{p}(M^*)} > \frac{u_2}{u_1}, \\ A^* & \text{if } \frac{\overline{p}(M_1)}{\overline{p}(M^*)} < \frac{u_2}{u_1} \text{ and } \frac{\overline{p}(M_0)}{\overline{p}(M^*)} < \frac{u_2}{u_0}, \\ M_0 & \text{otherwise.} \end{cases}$$

and finally, we replace A^* by $\{M_{i\gamma_i^*} : \overline{p}(M_{i\gamma_i^*}) > \frac{1}{u_2} \max\{\overline{p}(M_1)u_1, \overline{p}(M_0)u_0\}\}$. This enables us to describe the final rule in terms of thresholds on odds $\overline{p}(M_1)/\overline{p}(M_0)$, $\overline{p}(M_i^*)/\overline{p}(M_0)$ and $\overline{p}(M_i^*)/\overline{p}(M_1)$ only. Let $t_0 = u_0/u_1$ and $t_1 = u_1/u_2$. Noting that $\overline{p}(M^*) < x \Leftrightarrow \overline{p}(M_i^*) < x\ \forall i$ we get

$$\delta^* = \begin{cases} M_1 & \text{if } \frac{\overline{p}(M_1)}{\overline{p}(M_0)} > t_0 \text{ and } \frac{\overline{p}(M_i^*)}{\overline{p}(M_1)} < t_1 \text{ for all } i, \\ A^* & \text{if for some } i\text{: } \frac{\overline{p}(M_i^*)}{\overline{p}(M_0)} > t_0 t_1 \text{ and } \frac{\overline{p}(M_i^*)}{\overline{p}(M_1)} > t_1 \\ & \text{and } A^* = \{i : \text{ above holds}\}, \\ M_0 & \text{otherwise.} \end{cases} \tag{8.3.2}$$

This is almost the rule proposed in Sivaganesan, Laud, and Müller (2009). A similar rule is descibed there in terms of probabilities (rather than odds). The main difference to (8.3.2) is that $t_0 t_1$ is replaced by t_1. In Sivaganesan, Laud, and Müller (2009) the rule is defined without reference to a decision problem, and only justified by its frequentist properties. An important feature of δ^* is that the rule is described in terms of posterior odds $\overline{p}(M_0)/\overline{p}(M_1)$ and $\overline{p}(M_i^*)/p(M_j)$ $(j = 0, 1)$ only.

8.3.3 Probability Model

For a practical implementation of the proposed rule we still need to specify a probability model $p(M)$ over the space of all models $\mathscr{M}$. We proceed in steps. Let $\mathscr{M}_i = \{M_i\gamma_i,\ \gamma_i \in \Gamma\} \cup \{M_0, M_1\}$ denote the subspace defined by groupings based on covariate x_i, including the overall null and alternative. We define probability models $p_i(M)$ for $M \in \mathscr{M}_i$, $i = 1,\ldots,I$. We construct p_i such that $\pi_0 \equiv p_i(M_0)$ and $\pi_1 \equiv p_i(M_1)$ are common across i. Eventually we will piece the sub-models together to

$$p(M) = \begin{cases} \pi_0 & \text{for } M = M_0, \\ \pi_1 & \text{for } M = M_1, \\ p_i(M)\,\frac{1}{I} & \text{for } M = M_i\gamma,\ \gamma \in \Gamma. \end{cases} \tag{8.3.3}$$

We propose this construction of the full model $p(M)$ from models for the subspaces $\mathscr{M}_i$ because we find it easier to think about subgroups defined by one covariate at a time. We even further simplify the construction by using the same model for each p_i. The model $p_i(M)$ for each covariate is defined as a zero-enriched Polya urn, indexed with parameters (p, α). By a slight abuse of notation we write $p(\gamma_i)$ as short for $p_i(M_i\gamma_i)$ and factor $p(\gamma_i) = p(\gamma_{i0})\,p(\gamma_{i1} \mid \gamma_{i0})$ (The first factor refers to the probability of all models with the treatment effect in the subpopulation defined by $x_i = 0$ being characterized by γ_{i0}, and similarly for the second factor.)

We start the definition of p_i by $p(\gamma_{i0} = 0) = p(\gamma_{i1} = 0 \mid \gamma_{i0}) = p$, i.e., the probability of no treatment effect for any subpopulation is p, independently of the other subpopulations. Conditional on $\gamma_{i1} \neq 0$, we have $p(\gamma_{i1} = 1 \mid \gamma_{i0} = 0, \gamma_{i1} \neq 0) \equiv 1$, by definition of the notation. Finally, we define

$$p(\gamma_{i1} = g_1 \mid \gamma_{i0} = 1, \gamma_{i1} \neq 0) = \begin{cases} 1/(\alpha+1) & g_1 = 1, \\ \alpha/(\alpha+1) & g_1 = 2. \end{cases} \tag{8.3.4}$$

Model (8.3.4) is the Polya urn (Blackwell and MacQueen, 1973) for two observations. Since we allow an additional probability for the special cluster $\gamma_{ij} = 0$ we refer to the model as zero-enriched Polya urn. This description also clarifies the interpretation of the model parameters (p, α). The first, p is the marginal probability of no treatment effect in any subgroup defined by the covariate. The parameter α determines the relative probability of the treatment effect for one subgroup being different from the treatment effect for another subgroup with non-zero treatment effect; α are odds of different treatment effect versus same treatment effect.

Model (8.3.4) can straightforwardly be extended for categorical covariates with more than $S = 2$ levels. The interpretation of the parameters (p, α) remains unchanged. This is the model used in Sivaganesan, Laud, and Müller (2009).

When substituting $p(\gamma_i)$ back into (8.3.3) for the definition of $p(M)$ we have to modify the definition for covariates with $S > 2$ levels. Let $\pi_{S0} \equiv p_i(M_0)$ for covariates with S levels, and similarly for π_{S1}. For covariates with $S > 2$ levels we replace $p_i(M)$ in the last line of (8.3.3) by $p_i(M)\frac{1-\pi_0-\pi_1}{1-\pi_{S0}-\pi_{S1}}$, with $\pi_0 = \pi_{20}$ and $\pi_1 = \pi_{21}$, as before.

8.3.4 A Dementia Trial

Kovach et al. (2006) describe a two-arm double-blind randomized trial for late-stage dementia. The experimental therapy is Serial Trial Intervention (STI). The outcome of interest is the improvement in Discomfort-DAT (Discomfort-Dementia of Alzheimer's Type scale) from baseline to after intervention. The study enrolled 112 patients, with 55 randomized to control and 57 randomized to treatment. For each patient the investigators recorded two baseline covariates. The first covariate is an indicator x_1 for FAST (Functional Assessment of Staging of Dementia) score ≥ 7 ($x_1 = 1$) versus < 7 ($x_1 = 0$). The second covariate is an indicator for presence ($x_2 = 1$) versus absence ($x_2 = 0$) of vocalization in behavioral symptoms initiating treatment (MVOCAL). Subgroup sample sizes defined by these two covariates vary between 19 and 66. The same data was analyzed in Sivaganesan, Laud, and Müller (2009) for possible subgroup effects. To facilitate comparison we use the same sampling model and the same priors.

Let y_0 and y_1 denote the outcome of a patient under control and under STI, respectively. Conditional on an assumed model $M_{i\gamma_i}$ we assume for a patient with covariate $x_i = j$ the sampling model

$$y_0 \mid x_i = j \sim N(\mu_{0j}, \sigma^2) \text{ and } y_1 \mid x_i = j \sim N(\mu_{0j} + d_j, \sigma^2)$$

under control and treatment, respectively.

Here d_j is the treatment effect in the subgroup defined by $x_i = j$. The unknown parameters are $\mu_0 = (\mu_{0j},\ j = 0,\ldots,S-1)$, $(d_j,\ j = 0,\ldots,S-1)$ and σ^2. Let $K = \max\{\gamma_j,\ j = 0,\ldots,S-1\}$ denote the number of distinct non-zero treatment effects. We assume a hierarchical prior

$$d_j \mid \eta_{\gamma_j} \sim N(\eta_{\gamma_j}, \varepsilon\sigma^2)$$

with

$$\eta_k \mid g \sim N(0, g\sigma^2),\ k = 1,\ldots,K$$

for all distinct non-zero treatment effects and $\eta_0 \equiv 0$. Here $\varepsilon > 0$ is a small constant. The model is completed with a non-informative prior

$$p(g, \mu_0, \sigma^2) \propto \frac{1}{(1+g)^2}\,\frac{1}{\sigma^2},\ \ g > 0.$$

Such "mixture-g priors" have been proposed as reasonable non-informative priors for linear models (Berger, 2006; Liang et al., 2008).

Still to allow easy comparison, we use similar cutoffs as in Sivaganesan, Laud, and Müller (2009). They use cutoffs $c_0 = c_1 = 0.76$ and alternatively $c_0 = 0.86$, $c_1 = 0.5$ for probabilities $\overline{p}(M_1)/(\overline{p}(M_1) + \overline{p}(M_0))$ and for $\overline{p}(M_{i\gamma_i})/(\overline{p}(M_{i\gamma_i}) + \overline{p}(M_1))$, respectively. The cutoffs c on the probabilities are equivalent to cutoffs $t = \frac{c}{1-c}$ on the corresponding odds. There is no separate cutoff for comparing $M_{i\gamma_i}$ versus M_0 in the approach by Sivaganesan, Laud, and Müller (2009).

These cutoffs were determined in Sivaganesan et al. to achieve a desired type I error (TIE) rate and false subgroup rate (FSR). As usual, error rates are defined as (frequentist) error rates under repeat experimentation under an assumed truth. Let $p_f(A \mid M)$ denote the probability of a decision A under repeat experimentation under an assumed true model M. We define

$$TIE = p_f(M_0 \text{ not selected } \mid M_0)$$

and

$$FSR = p_f(\text{some } M_{i\gamma} \text{ selected } \mid M_1).$$

See Sivaganesan, Laud, and Müller (2009). for a more extensive discussion of alternative error rates. As done in that paper, we used simulation to calculate an average FSR, averaged over the normal distribution for the overall effect size with mean 0 and standard deviation equal to that of the data.

TABLE 8.4. Posterior probabilities of models for Kovach et al. (2006) data.

	Model							
	M_0	$M_{1(1,0)}$	$M_{1(0,1)}$	$M_{1(1,2)}$	$M_{2(1,0)}$	$M_{2(0,1)}$	$M_{2(1,2)}$	M_1
Prior Prob.	3/8	1/16	1/16	3/32	1/16	1/16	3/32	3/16
Posterior Prob.	4×10^{-5}	3×10^{-4}	0.032	0.054	10^{-4}	0.605	0.305	2.2×10^{-3}

TABLE 8.5. Subgroup effect model as selected by the decision rule in Section 8.3.2.2 for Kovach et al. (2006) data.

c_0	c_1	TIE	Model(s) Selected	Average FSR
0.76	0.76	0.03	$M_{1(1,2)}, M_{2(0,1)}$	0.13
0.86	0.50	0.05	Same as above	0.33

The posterior probabilities of models using the Kovach et al. (2006) data are given in Table 8.4. The selected models and the corresponding error rates are given in Table 8.5. Recall that the covariates are x_1=FAST and x_2=MVOCAL. The selected models are $M_{1(1,2)}$ and $M_{2(0,1)}$. Model $M_{1(1,2)}$ indicates non-zero and distinct effects in the subgroups defined by FAST. Model $M_{2(0,1)}$ refers to a model with no treatment effect when MVOCAL$= 0$ and a non-zero effect when MVOCAL$= 1$. For comparison, in Sivaganesan, Laud, and Müller (2009), either $M_{2(0,1)}$, or both $M_{1(1,2)}$ and $M_{2(0,1)}$ were selected depending on the chosen cutoffs or error rates.

8.3.5 Discussion

We have outlined a practicable decision theoretic approach to subset selection. The final rule is described in terms of simple thresholds on posterior odds of subgroup models relative to the overall null and alternative models.

The strengths of the proposed approach are the principled nature and the simplicity of the solution. The nature of the proposed solution as a Bayes rule makes it easy to adjust for problem-specific variations. For example, if the relative utilities (or losses) were to differ from (8.3.1), one could easily follow the same argument to find alternative solutions. For example, (8.3.1) implies that the loss of not reporting the correct model is the same, no matter which wrong model is reported.

Many limitations remain. Like in any decision theoretic approach, the solution depends on the often arbitrary choice of the relative utilities u_0, u_1, and u_2. The problem is mitigated by validation of frequentist operating characteristics like TIE and FSR, which can help to calibrate these relative utilities and the implied thresholds in the decision rule.

Finally, we note that although the problem involves massively multiple comparisons, commonly used false discovery rate (FDR) control is not meaningful. Since the model $p(M)$ implicitly allows only one true model, any reported set A of subgroup models would trivially include at least $|A| - 1$ false discoveries.

Acknowledgments: We thank Christine Kovach, PhD, RN of the University of Wisconsin-Milwaukee and Brent Logan, PhD of the Medical College of Wisconsin for providing advice and the data from their study. This research was initiated during a program at SAMSI (Statistical and Applied Mathematical Sciences Institute, NC).

Chapter 9
Bayesian Methods for Genomics, Molecular and Systems Biology

Inference for high throughput genomic data has emerged as a major source of challenges for statistical inference in general, and Bayesian analysis in particular. This chapter discusses some related current research frontiers. The chapter highlights how specific strengths of the Bayesian approach are important to model such data. Bayesian inference provides a natural paradigm to exploit the considerable prior information that is available about important biological pathways. Another strength of Bayesian inference that leads to research opportunities with phylogenomic data is the natural ease of simultaneous modeling and inference on multiple related processes.

9.1 Bayesian Modelling for Biological Annotation of Gene Expression Pathway Signatures

Haige Shen and Mike West

Studies in high-throughput genomics often generate multiple gene expression *signatures*—lists of genes with associated numerical measures of change in gene expression relative to an experimental condition or outcome. A biological or environmental design factor in a controlled experiment generates a signature of response to that factor (Huang et al., 2003b; Bild et al., 2006; Chen et al., 2008), while evaluation of expression related to a specific clinical outcome may generate a signature of the outcome in disease studies (West et al., 2001; Huang et al., 2003a; Rich et al., 2005; Seo et al., 2007). Indeed, the concept of gene expression signatures as characterizing pathway status has emerged as central in Bayesian analyses in genomics and emerging systems biology in cancer and other areas. The development of sparse ANOVA and latent factor models to improve estimation of expression signatures in both experimental (*in vitro*) and observational (*in vivo*) contexts has been substantially motivated by this view (e.g., West, 2003; Lucas et al., 2006; Wang

M.-H. Chen et al. (eds.), *Frontiers of Statistical Decision Making and Bayesian Analysis*, DOI 10.1007/978-1-4419-6944-6_9,

et al., 2007; Carvalho et al., 2008; Merl et al., 2009b). Recent applied studies of deregulated pathways in cancer as well as other contexts, including basic biological pathway discovery and evaluation studies, studies of drug responsiveness and prognostic/predictive risk profiling, reflect this (e.g., Chen et al., 2008; Chang et al., 2009; Lucas, Carvalho, and West, 2009; Merl et al., 2009a).

Interpretation of identified gene expression signatures relies in part on comparison with biological databases that contain lists of putatively pathway-specific genes. The gene lists themselves, without real-biology connectivities explicitly described, simply represent biological pathways through the sets of genes named as participating in the biological processes the pathways play roles in. A core challenge is to assess the signatures against these databases to suggest potential pathway interpretations. Our focus here is a formal, novel Bayesian approach to this problem.

Identification of this problem led to the non-Bayesian gene set enrichment analysis (GSEA) method (Subramanian et al., 2005) and follow-on approaches (Newton et al., 2007). These methods aim to measure aggregate association between a full list of genes ranked by their association with an outcome — also referred to as a *phenotype* — and one or more given sets of genes. The underlying idea is to assess whether or not a specified "pathway" gene set is enriched with genes that score highly in association with the experimental outcome. Based on non-Bayesian testing and (sample or gene) randomization methods, these methods tend to lead to false positives, have difficulties in dealing with small sized gene sets, rely on an assumption that pathway database gene lists are error-free, and are restricted in applications to simple contexts of genes up/down regulated. On the latter point, we are particularly interested in understanding potential biological pathways underlying estimated latent factors applied to observational data sets (e.g., Lucas et al., 2006; Carvalho et al., 2008; Lucas et al., 2009; Merl et al., 2009b) and these existing methods simply do not apply to the forms of information summaries produced in such analyses.

Our Bayesian *probabilistic pathway annotation* (PROPA) model presented here addresses these broader questions. PROPA provides: (a) probabilistic assessments of phenotype-pathway concordance in terms of marginal likelihoods and posterior probabilities; (b) an ability to assessment of experimental results against many biological pathways simultaneously and in comparison with each other; (c) adaptation to uncertainties and potential errors in both experimentally defined gene-phenotype association measures *and* in biological databases; and (d) a general theoretical framework that allows the specific method to be extended to incorporate other forms of genomic data. Item (c) here also leads to an ability to suggest refinements to pathway gene lists. Simulation and breast cancer genomics examples illustrate these points. A core component of the annotation analysis involves evaluation of marginal likelihoods in models with high-dimensional parameters. For this, we develop a novel extension of variational methods (e.g., Jordan et al., 1999; McGrory and Titterington, 2007) that, in addition to proving extremely effective in the pathway annotation analysis, is of broad interest and potential use in Bayesian model evaluation.

9.1.1 Context and Models

9.1.1.1 Notation and Framework

A biological study investigates the changes in gene expression on p genes due to an experimental *factor*. We are interested in (a) which genes are related to this factor in terms of the expression change, and (b) how does this factor relate to known, published gene lists representing annotation of biological pathways? The experiment leads to measures of association of the genes, in terms of expression changes, with the experimental factor; these are inputs to annotation analysis. We define terminology and notation as follows:

- $\mathscr{G} = \{1,\ldots,p\}$, the full list of genes; in human studies, $p \sim 20-25{,}000$.
- A *pathway* is, simply, any specific subset of genes from $\mathscr{G}$.
- $\mathscr{F}$, an unknown list of genes whose expression changes are truly related to the experimental factor; we call $\mathscr{F}$ the *factor pathway* to give it a definite name.
- $\Pi = \{\pi_g,\ g = 1,\ldots,p\}$, a set of numerical measures of association of each of the genes, in terms of the expression change, with the experimental factor pathway $\mathscr{F}$.
- $\mathscr{A}$, a generic label for a *biological pathway*; $\mathscr{A}$ is a simply an unknown list of genes. $\mathscr{A}_j,\ j = 1,\ldots,m$, a full set of known biological pathways.
- A, a generic label for a list of genes in a published, annotated biological database, putatively linked to a true, unknown biological pathway $\mathscr{A}$. We call A a *reference gene list* for pathway $\mathscr{A}$. $A_j,\ j = 1,\ldots,m$, the set of reference gene lists corresponding to pathways $\mathscr{A}_j$.

We use the Molecular Signatures database (MSigDB C2 collection) (Broad Institute, 2007) to obtain $m \approx 1000$ reference gene sets A_j. These are, of course, incomplete and typically error-prone; A_j provides incomplete and noisy information on the pathway $\mathscr{A}_j$.

Based on the expression experiment, Π is known *data* to be used in assessing concordance of the unknown, underlying experimental factor pathway $\mathscr{F}$ with candidate biological pathways $\mathscr{A}_j,\ j = 1,\ldots,m$. We do this with models that compute the $j = 1,\ldots,m$ posterior probabilities

$$Pr(\mathscr{F} = \mathscr{A}_j|\Pi, A_1,\ldots,A_m) \propto Pr(\mathscr{F} = \mathscr{A}_j|A_1,\ldots,A_m)p(\Pi|A_1,\ldots,A_m,\mathscr{F} = \mathscr{A}_j).$$

Focus here on the likelihood terms $p(\Pi|A_1,\ldots,A_m,\mathscr{F} = \mathscr{A}_j)$ as j moves across all the pathways; this is the overall measure from the experimental predictions Π that underlies pathway assessment, and can be applied whatever the chosen values of the $Pr(\mathscr{F} = \mathscr{A}_j|A_1,\ldots,A_m)$.

Here Π may include essentially any measures, such as test statistics or other summaries of a statistical analysis of the experimental data. Different measures should be modelled differently within the overall framework. Here, our example measures are probabilities of differential expression. In designed experiments, π_g will be a posterior probability of differential expression of gene g related to an experimental

intervention. In observational studies, π_g will be a posterior probability of a non-zero regression coefficient or loading on a latent factor in a sparse factor model of expression data (West, 2003; Lucas et al., 2006; Seo et al., 2007; Carvalho et al., 2008). So $\pi_g \in [0,1]$ and larger values indicate stronger association with $\mathscr{F}$; typically very many of the π_g will be very small, while those for genes associated with $\mathscr{F}$ will be larger.

9.1.1.2 Statistical Model

Focus on a single, generic biological pathway $\mathscr{A} = \mathscr{A}_1$ and its reference gene list $A = A_1$, and consider relevant statistical models for the core component $p(\Pi|A_1,\ldots, A_m, \mathscr{F} = \mathscr{A}_j)$.

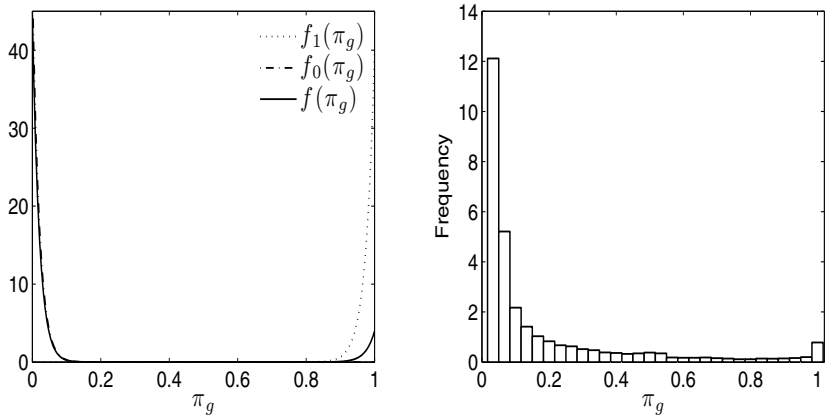
Model for data Π assuming known pathway membership of genes. We first assume that π_g, given an associated pathway $\mathscr{A}$ and its reference gene set A, is independent of the other $\{\pi_g\}_{k\neq g}$. Note that this *does not* assume lack of interaction or co-regulation among genes; that dependence should already be accounted for in the analysis that led to the Π. If $\mathscr{F} = \mathscr{A}$, then π_g will likely be higher for $g \in \mathscr{A}$ than for $g \notin \mathscr{A}$, suggesting models of the form

$$(\pi_g|g \in \mathscr{A}, \mathscr{F} = \mathscr{A}) \sim f_1(\pi_g) \quad \text{and} \quad (\pi_g|g \notin \mathscr{A}, \mathscr{F} = \mathscr{A}) \sim f_0(\pi_g), \qquad (9.1.1)$$

where f_0, f_1 are densities on $[0,1]$ with f_1 favoring high values of π_g and f_0 favoring lower values. A natural choice is beta densities: $f_1(\pi) \equiv f_1(\pi|\alpha_1) = \text{Be}(\alpha_1, 1)$ and $f_0(\pi) \equiv f_0(\pi|\alpha_0) = \text{Be}(1, \alpha_0)$ with $\alpha_0, \alpha_1 > 1$ (Figure 9.1(a)). This picture is consistent with histograms of π_g values generated in sparse factor analyses (e.g., Carvalho et al., 2008; Wang et al., 2007-present); see Figure 9.1(b) as an example. We have explored model robustness to the assumed form in other examples, including simulation examples, and have no major concerns about the beta forms being overly restrictive, though other forms will be relevant in analyses with other definitions of Π. We use independent reference priors for the α parameters, viz $p(\alpha_r) \propto \alpha_r^{-1}$ for $1 < \alpha_r$ $(r = 0, 1)$.

Model for pathway membership of genes. We do not know which genes are in $\mathscr{A}$; the reference gene set A provides data. If $g \in A$, that suggests $g \in \mathscr{A}$ although g may be a false-positive in the published list. Also, reference gene lists are subject to revision as new biological information arises, so genes $g \notin A$ may be members in future; hence, there may be false-negatives, i.e., genes $g \in \mathscr{A}$ but $g \notin A$.

Introduce indicators $z_1, \ldots, z_p$ such that, when $\mathscr{F} = \mathscr{A}$, $z_g = 1$ if $g \in \mathscr{A}$, and 0 otherwise. Call z_g the *pathway membership indicator* of gene g. We need probabilities over the z_g; A provides relevant information. Assume conditionally independent Bernoulli models $Pr(z_g = 1|\beta_g) = \beta_g$, so that marginalization of (9.1.1) with respect to z_g yields the implied prior data distribution as a mixture of $f_1(\pi_g|\alpha_1)$ and $f_0(\pi_g|\alpha_0)$ weighted by β_g and $1-\beta_g$; see Figure 9.1(a). To complete the model requires priors for the β_g, which we take as $(\beta_g|g \in A, \mathscr{F} = \mathscr{A}) \sim \text{Be}(\phi_A r_A, \phi_A(1 - r_A))$ and $(\beta_g|g \notin A, \mathscr{F} = \mathscr{A}) \sim \text{Be}(\phi_B r_B, \phi_B(1 - r_B))$ with specified means $r_A, r_B \in$



(a) (b)

FIGURE 9.1. (a) $f(\pi_g)$ is a mixture of $f_1(\pi|\alpha_1) = \text{Be}(\alpha_1, 1)$ and $f_0(\pi|\alpha_0) = \text{Be}(1, \alpha_0)$; in this example, $f_0(\pi_g)$ is very close to $f(\pi_g)$. (b) Histogram of π_g of thousands of genes from a real expression data analysis.

$(0,1)$ and $\phi_A, \phi_B > 0$. Marginalizing over the β_g, we see that r_A is the *á priori* true positive probability for genes $g \in A$, while r_B is the false negative probability for $g \in \mathscr{A}$. Specification of r_A should depend on the expectation of the quality of reference gene sets. In a pathway gene set database, genes sets are curated from a variety of sources. We adopt a generic view that a published gene set A is a fairly good representation of the true pathway gene set $\mathscr{A}$ but allow for errors, so take r_A relatively large, e.g., 0.7. For r_B, note that the number of genes in A, typically tens to a few hundreds, will usually be small compared to the full gene list $\mathscr{G}$, and a reasonable value of r_B should be at least less than the ratio of the number of signature genes (genes with high probabilities of association with $\mathscr{F}$) to the total number of genes, e.g., 0.005. The specification of r_A and r_B is empirical and to some extent allows flexibility. The impact of r_A and r_B specification is demonstrated in an example below. The ϕ_A and ϕ_B constrain the variation range of the prior for the β_g around these means, and relatively small values provide robustness.

Annotated databases are incomplete and error prone. We can explore this using posterior pathway membership probabilities for each gene g, namely

$$\pi_g^* = Pr(g \in \mathscr{A} | \Pi, A, \mathscr{F} = \mathscr{A}), \tag{9.1.2}$$

with respect to the pathway $\mathscr{A}$. This is exemplified below. In any one example, there may well be genes *known* to lie in a specific biological pathway but that are not activated under a specific experimental condition. Such genes will be treated as false-positive members of a reference gene set in our model; they may, of course, appear differently under other experimental conditions.

9.1.1.3 Marginal Likelihood for Pathway Assessment

We use $\alpha_{0:1}$, $\beta_{1:p}$ and $z_{1:p}$ to denote $\{\alpha_0,\alpha_1\}$, $\{\beta_1,\ldots,\beta_p\}$ and $\{z_1,\ldots,z_p\}$, respectively, extending the use of this concise notation to other quantities as needed. The full model likelihood $p(\Pi|A,\mathscr{F}=\mathscr{A})$ can be expressed as

$$\begin{aligned} &p(\Pi|A,\mathscr{F}=\mathscr{A}) \\ &= \int_{\alpha_{0:1}}\int_{\beta_{1:p}}\sum_{z_{1:p}}\mathscr{L}(\alpha_{0:1},z_{1:p})\prod_{g=1}^{p}p(z_g|\beta_g)p(\beta_g|A,\mathscr{F}=\mathscr{A})p(\alpha_{0:1})d\beta_{1:p}d\alpha_{0:1} \end{aligned} \tag{9.1.3}$$

with $\mathscr{L}(\alpha_{0:1},z_{1:p})=\prod_{g=1}^{p}f_1(\pi_g|\alpha_1)^{z_g}f_0(\pi_g|\alpha_0)^{1-z_g}$. We can integrate analytically over $\beta_{1:p},\alpha_{0:1}$ reducing the computation to summation over the 2^p values $z_{1:p}$; see Section 9.1.5.1 where we derive

$$p(\Pi|A,\mathscr{F}=\mathscr{A})=\sum_{z_{1:p}}p(\Pi,z_{1:p}|A,\mathscr{F}=\mathscr{A}) \tag{9.1.4}$$

and where the quantity $p(\Pi,z_{1:p}|A,\mathscr{F}=\mathscr{A})$ can be evaluated at any chosen $z_{1:p}$. The sum above is a difficult numerical problem addressed in Section 9.1.2.2.

Another reduced form that is theoretically attractive but practically of little value results from marginalization over $z_{1:p}$ and $\beta_{1:p}$ conditional on $\alpha_{0:1}$, namely

$$p(\Pi|A,\mathscr{F}=\mathscr{A})=\int_{\alpha_{0:1}}p(\Pi|\alpha_{0:1},A,\mathscr{F}=\mathscr{A})p(\alpha_{0:1})d\alpha_{0:1}. \tag{9.1.5}$$

The integrand here can be evaluated, but only practicably when p is small; see Section 9.1.5.1.

9.1.2 Computation

9.1.2.1 MCMC Posterior Simulation

The following conditional distributions are immediate; in each, only the conditioning quantities required to specify the distribution are mentioned.

First, α_0 and α_1 are conditionally independent with truncated gamma conditionals; specifically, the two distributions are $\mathrm{Ga}\Big(\alpha_0|\sum_{g=1}^{p}(1-z_g),-\sum_{g=1}^{p}(1-z_g)\log(1-\pi_g)\Big)$ and $\mathrm{Ga}\Big(\alpha_1|\sum_{g=1}^{p}z_g,-\sum_{g=1}^{p}z_g\log\pi_g\Big)$, subject to $1<\alpha_r$ $(r=0:1)$.

Second, the β_g are conditionally independent with beta distributions $\mathrm{Be}(a_g,b_g)$ depending on z_g. For $g\in A$, $a_g=z_g+\phi_A r_A$ and $b_g=(1-z_g)+\phi_A(1-r_A)$; for $g\notin A$, $a_g=z_g+\phi_B r_B$ and $b_g=(1-z_g)+\phi_B(1-r_B)$.

Third, the z_g are conditionally independent with probabilities on $z_g=1$ of

$$\rho_g = \beta_g \alpha_1 \pi_g^{\alpha_1 - 1} / \left(\beta_g \alpha_1 \pi_g^{\alpha_1 - 1} + (1 - \beta_g) \alpha_0 (1 - \pi_g)^{\alpha_0 - 1} \right).$$

The posterior pathway membership probability π_g^* is the posterior mean of ρ_g.

Efficient code for this MCMC evidences generally fast mixing and rapid convergence across many examples. The rather low dependence among the z_g, induced by lack of knowledge of the $\alpha_{0:1}$, suggests swift convergence is to be expected even though $p \approx 20 - 25,000$.

9.1.2.2 Marginal Likelihood Computation: General Strategy

A core methodological issue is the evaluation of the determining marginal likelihood of (9.1.3), and sets of such quantities $p(\Pi | A_1, \ldots, A_m, \mathscr{F} = \mathscr{A}_j)$ in the practical context of assessing evidence for and against $\mathscr{F} = \mathscr{A}_j$ for a number or many pathways $j = 1, \ldots, m$.

In very small, unrealistic examples we can use quadrature methods to compare with other approximations. We do this in the simulated example in Section 9.1.3, simply applying direct quadrature to the two-dimensional integral form of (9.1.5). Even with p very small, this method is limited since it requires evaluation of integrands on the density scale and quickly runs into floating-point overflow problem. Quadrature is simply not relevant for real applications.

The reduced version of (9.1.4) has a closed form but involves summing over all 2^p values of $z_{1:p}$ so that numerical approximations are needed. Since we use MCMC, then methods of marginal likelihood computation using MCMC outputs are attractive. Having experimented with multiple such methods (Newton and Raftery, 1994; Chib, 1995), all found to be inapplicable due to either the floating-point overflow problem or difficulties in proposing good density functions to approximate the joint posterior distribution of model parameters, we adapted mean-field variational methods (VM) (Jordan et al., 1999; Corduneanu and Bishop, 2001; McGrory and Titterington, 2007). The VM approach naturally solves the floating-point overflow problem by using a summation of logarithmic terms to approximate the log marginal likelihood. Our studies confirm the utility of this approach, especially in this high-dimensional context. A VM method yields a *lower bound* on the target value of the marginal likelihood; our extensions include an *upper bound* so we can bracket the actual value.

For *any* two densities $q_L(z_{1:p}), q_U(z_{1:p})$ with the same support as $p(z_{1:p} | \Pi, A, \mathscr{F} = \mathscr{A})$, manipulating Jensen's inequality easily yields

$$L(q_L) \leq \log(p(\Pi | A, \mathscr{F} = \mathscr{A})) \leq U(q_U),$$

where, for any such density $q(z_{1:p})$, the quantities

$$L(q) = \sum_{z_{1:p}} q(z_{1:p}) \log[p(\Pi, z_{1:p} | A, \mathscr{F} = \mathscr{A}) / q(z_{1:p})] \tag{9.1.6}$$

and

$$U(q) = \sum_{z_{1:p}} p(z_{1:p}|\Pi, A, \mathscr{F} = \mathscr{A}) \log[p(\Pi, z_{1:p}|A, \mathscr{F} = \mathscr{A})/q(z_{1:p})] \qquad (9.1.7)$$

bound the log marginal likelihood; see Section 9.1.5.2 for technical details.

The VM concept is to choose parametric *variational densities* $q_L(z_{1:p})$ and $q_U(z_{1:p})$ to optimize these bounds. If each depends on a free parameter that can be varied, the computational problem is optimizing these *variational* parameters. The closer a variational density is to $p(z_{1:p}|\Pi, A, \mathscr{F} = \mathscr{A})$, the better will be the bound. Mean-field VM methods use factorized variational densities. This is natural here since the z_g have low dependence under the posterior. Thus we use q_L, q_U of the form

$$q(z_{1:p}|\gamma_{1:p}) = \prod_{g=1}^{p} \gamma_g^{z_g} (1-\gamma_g)^{1-z_g}, \qquad (9.1.8)$$

where the $\gamma_{1:p}$ are vectors of variational parameters to be chosen.

9.1.2.3 Marginal Likelihood Computation: A Variational Method

We refer to the implementation of the above ideas that rely on the MCMC analysis as *Monte Carlo variational approximation*. Full details appear in Ji, Shen, and West (2009); essential results for the PROPA model are noted here, with more details in Section 9.1.5.2

Upper Bound Optimization. With q_U of the form in (9.1.8), it is trivially seen that the global minimum value of the upper bound in (9.1.7) is achieved at $\gamma_{1:p} = \bar{z}_{1:p} = E(z_{1:p}|\Pi, A, \mathscr{F} = \mathscr{A})$, i.e., by setting the latent indicators $z_{1:p}$ equal to their posterior means. These means are estimated at values $\bar{z}_{1:p}$ based on the MCMC output $\{z^i_{1:p},\ i = 1, \ldots, I\}$ and the Monte Carlo approximation to the optimal upper bound is simply

$$\bar{U} = I^{-1} \sum_{i=1}^{I} \{\log p(\Pi, z^i_{1:p}|A, \mathscr{F} = \mathscr{A}) - \log q(z^i_{1:p}|\bar{z}_{1:p})\}.$$

This is easily computed and, assuming MCMC convergence, $\bar{U}$ converges almost surely to the true global minimum upper bound of the log marginal likelihood.

Lower Bound Optimization. Existing mean-field, lower bound variational methods typically build on the Monte Carlo EM algorithm (Celeux and Diebolt, 1992; Chan and Ledolter, 1995). By combining with a stochastic approximation step, convergence of a stochastic version of EM was established under mild conditions in Delyon, Lavielle, and Moulines (1999). This inspired the novel variational method (Ji, Shen, and West, 2009) that is applied here; full algorithmic details are in Section 9.1.5.3.

The global optimizing value of $\gamma_{1:p}$ satisfies the set of p equations $f_g(\gamma_{1:p}) = 0$, where for each $g = 1, \ldots, p$,

$$f_g(\gamma_{1:p}) = \sum_{z_{1:p}} (z_g - \gamma_g)\,[1 + \log q(z_{1:p}|\gamma_{1:p}) - \log p(\Pi, z_{1:p}|A, \mathscr{F} = \mathscr{A})]\,. \quad (9.1.9)$$

An iterative procedure successively approximates the solution to these equations using Monte Carlo and stochastic approximation; the former allows us to estimate $f_g(\gamma_{1:p})$ by Monte Carlo over $z_{1:p}$ at any value of $\gamma_{1:p}$, while the latter applies to successively update estimates of the optimizing vector $\gamma_{1:p}$. As detailed in Section 9.1.5, an iterative algorithm uses these ideas to define a sequence of $\gamma_{1:p}$ vectors that converges with probability one to $\gamma^*_{1:p}$ satisfying $f_g(\gamma^*_{1:p}) = 0$, $g = 1, ..., p$; a finite run of the algorithm provides an iterative approximation to this optimizing value. By further Monte Carlo sampling $z^h_{1:p} \sim q(z_{1:p}|\gamma^*_{1:p})$, $(h = 1, \ldots, H)$, we can then also evaluate a consistent estimate of the optimal lower bound,

$$\bar{L} = H^{-1} \sum_{h=1}^{I} \{\log p(\Pi, z^h_{1:p}|A, \mathscr{F} = \mathscr{A}) - \log q(z^h_{1:p}|\gamma^*_{1:p})\}. \quad (9.1.10)$$

9.1.3 Evaluation and Illustrations

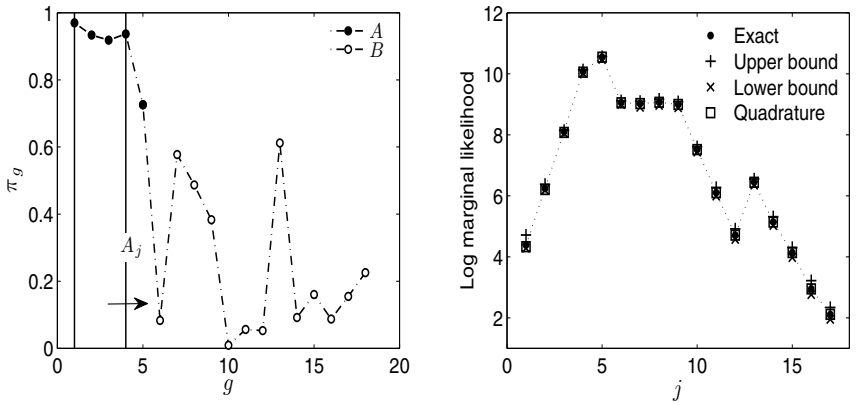


FIGURE 9.2. (a) π_g in the simulated data set with $p = 18$. Genes in reference set A_j are genes $1, \ldots, j$ for each $j = 1, \ldots, 17$. (b) Log marginal likelihood for each of the 17 pathways $\mathscr{A}_j$.

A "Small" Simulated Example $(p = 18, m = 17)$**.** In a synthetic example to fix ideas and demonstrate the marginal likelihood approximation, association probabilities on $p = 18$ genes (Figure 9.2) show that the first five genes are likely members of $\mathscr{F}$, several genes with very low π_g are not likely to be in $\mathscr{F}$, while four genes with π_g near 0.5 are uncertain. Consider $m = 17$ biological pathway reference gene

sets, $A_1,\ldots,A_{17}$, constructed as in Figure 9.2(a); reference set A_j is the first j genes in the ordered list of 18 genes. Analyses use $r_A = 0.8$, $r_B = 0.1$ and $\phi_A = \phi_B = 8$.

The log marginal likelihood (shifted and scaled to $[0,1]$ in Figure 9.2(b)) increases over $j = 1,\ldots,5$ to a peak at $j = 5$, suggesting pathways $\mathscr{A}_4$ and $\mathscr{A}_5$ are supported by the data Π. This is consistent with the simulation design in that the first few genes are the signature genes of $\mathscr{F}$, having high π_g values. The marginal likelihood across the remaining reference gene sets is also reasonable given the values of the π_g. Figure 9.2(b) shows that the Monte Carlo variational upper and lower bounds agree well with the exact marginal likelihood values and quadrature based approximations using (9.1.5). In this "tiny p" example, exact and quadrature computations are feasible, and demonstrate the accuracy of the upper and lower bound approximations. The spread between upper and lower bounds are small on the log likelihood scale ($\sim 0.05 - 0.2$) and certainly good enough to distinguish the different pathways/models.

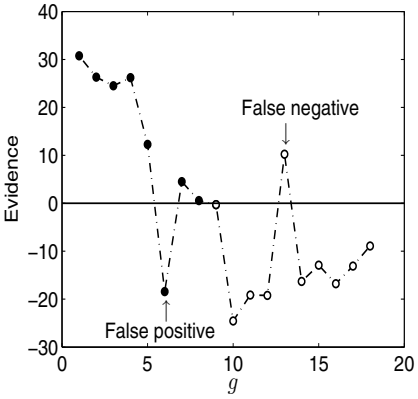


FIGURE 9.3. Pathway $\mathscr{A}_8$ membership evidence for each gene g, in terms of log base 10 Bayes' factors for:against $g \in \mathscr{A}_8$, in analysis of simulated data with $p = 18$.

The MCMC provides estimates of posterior pathway membership probabilities π_g^* of (9.1.2) to aid false-positive/false-negative assessments. Focus on pathway $\mathscr{A}_8$; reference set A_8 is exactly the first 8 genes. For each g and each reference gene set, compute π_g^* and convert to the corresponding *evidence* dB scale, i.e., the log base 10 Bayes' factors on $g \in A_8$ versus $g \notin \mathscr{A}_8$; see Figure 9.3. Genes in A_8 but with low π_g, and genes not in A_8 but with high π_g, might be regarded as false positives and false negatives, respectively. Gene $g = 6$, a member of gene set A_8, has membership evidence close to -20dB, strongly suggesting it is not a member of the true pathway $\mathscr{A}_8$ (false positive). Gene 13 is not a member of A_8, but it has membership evidence greater than 10dB, which is substantial evidence that this gene is in fact a member of $\mathscr{A}_8$ (false negative).

Marginal Likelihood Approximation with Real Data: "Large" p = 19,645. With realistically large p, the convergence of the iterative lower bound optimization can

be slow. A slight modification of the bounding approach to address this is a compromise strategy with a pseudo-optimal lower bound; specifically, a bound as in (9.1.10) but now with $\gamma^*_{1:p}$ replaced by $\bar{z}_{1:p}$, the MCMC posterior mean of $z_{1:p}$. This uses the same variational density as in the optimal *upper* bound approximation. The rationale is that, when the factorized density q is a good approximation of the posterior for $z_{1:p}$, the optimal variational densities for upper and lower bounding will be similar; this has been seen in multiple examples. The pseudo-optimal lower bound is always less than the optimal, but is massively more attractive computationally when p is large.

We demonstrate this with real data on $p = 19,645$ genes and with $m = 15$ pathways whose reference gene sets come from the MSigDB C2 collection. The Π are probabilities of association between genes and a gene expression signature representing genes related to the responses of human mammary epithelial cells to lactic acidosis (Chen et al., 2008; Merl et al., 2009a) Analysis assumes $r_A = 0.7$, $r_B = 0.005$, $\phi_A = 8$, and $\phi_B = 3$. Figure 9.4(a) shows upper and pseudo-optimal lower bounds of log marginal likelihoods for the 15 pathway gene sets. The distances between pairs of bounds are clearly small enough for practical usage; see Figure 9.4(b).

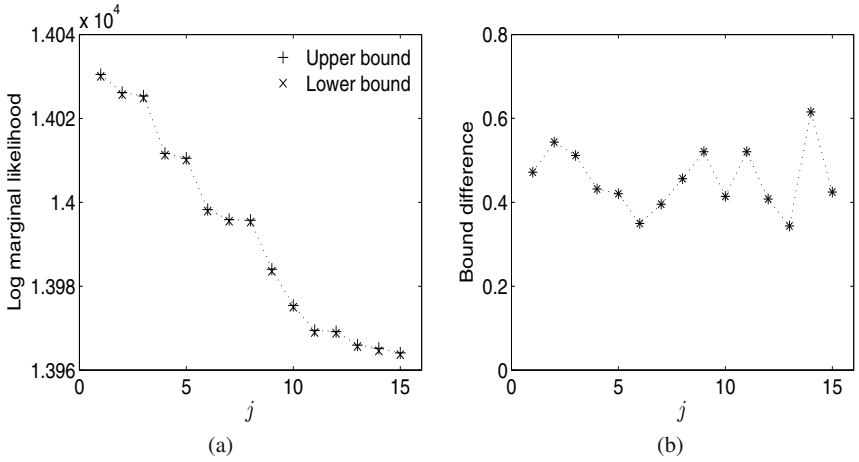


FIGURE 9.4. (a) Upper and quasi-lower bounds of log marginal likelihoods for each pathway $j = 1 : m, m = 15$ in study with $p = 19,645$ genes. (b) Upper minus lower bound for each pathway $j = 1 : 15$.

9.1.4 Applications to Hormonal Pathways in Breast Cancer

9.1.4.1 ER Pathway

About two-thirds of diagnosed breast cancers show over-expression of ER, the estrogen-receptor gene. ER status (high/low) is a key prognostic factor in breast cancer (Moggs and Orphanieds, 2001; Deroo and Korach, 2006). Our prior study of 153 primary breast tumor samples (Carvalho et al., 2008) records expression data and protein assay-based ER$+/-$ status from immunohistochemical (IHC) staining. Analysis using BFRM (Wang et al., 2007) generated association probabilities $\Pi = \pi_{1:p}$ as well as the sign of association between expression and ER$+/-$ status for $p = 8,764$ genes (unique Entrez gene IDs). The π_g, displayed in Figure 9.5(a), show that a substantial number of genes apparently associate with the experimental factor pathway $\mathscr{F}$, here known to be ER related. Figure 9.5(b) shows PROPA upper and pseudo-optimal lower bounds on log marginal likelihoods for the $m = 956$ MSigDB pathway gene sets. For almost all the pathways, the distance between upper and lower bound is very small and hence the evidence is reliably evaluated.

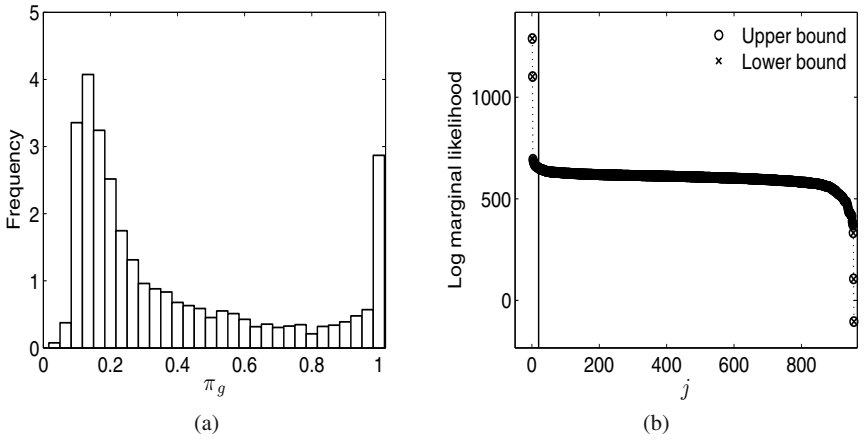


FIGURE 9.5. Breast cancer ER status study. (a) Histogram of association probabilities π_g. (b) Log marginal likelihood upper ($\circ$) and lower ($\times$) bounds for pathway gene lists $j = 1,\ldots,956$ sorted in decreasing order; the line demarks the "top 20" pathways.

Table 9.1 summarizes the top 20 pathway gene sets. The first two are breast tumor ER$-/+$ signatures defined by experimental microarray studies in Van't Veer et al. (2002), clearly validating the PROPA results. PROPA identifies several other pathway gene sets with defined links to breast tumor ER status. Patients with ER$-$ tumors generally have poorer prognoses than those with ER$+$ tumors, and there are several well-known risk-related signatures linked to this that involve intersecting gene sets (Maynard et al., 1978; Van't Veer et al., 2002); these are well-represented

TABLE 9.1. Summary of top ER-related pathways identified by PROPA.

Rank	Pathway	Size	Log(ML): UB	Log(ML): LB	UB-LB
1	BRCA ER Neg	692	1289.58	1289.15	0.43
2	BRCA ER Pos	380	1102.39	1101.81	0.58
3	Flechner Kidney Transplant Rej/Up	81	694.82	694.20	0.62
4	BRCA Prognosis Neg	69	684.14	683.41	0.73
5	Caries Pulp Up	186	680.36	679.93	0.43
6	Cancer Undifferentiated Meta Up	65	676.26	675.72	0.55
7	UVB NHEK3 C7	50	670.12	669.43	0.68
8	CIS XPC Up	131	666.99	666.30	0.69
9	Serum Fibroblast Cell cycle	83	665.70	665.10	0.60
10	Li Fetal VS WT Kidney DN	157	662.48	661.74	0.74
11	Vantveer Breast Outcome/Down	58	661.27	660.69	0.58
12	Frasor ER Up	29	661.21	660.70	0.50
13	Tarte Plasma Blastic	295	660.86	660.28	0.58
14	BRCA Prognosis Pos	26	657.90	657.28	0.62
15	IFNA HCMV 6hrs Up	52	657.72	657.22	0.50
16	Caries Pulp High Up	83	656.37	655.75	0.62
17	Lindstedt Dend 8h vs 48h Up	64	654.77	654.11	0.66
18	Vantverr Breast Outcome/Up	20	654.63	654.02	0.60
19	Zhan MM Molecular Classi Up	34	653.06	652.42	0.64
20	Becker Tamoxifen Resistant DN	48	652.40	651.72	0.68

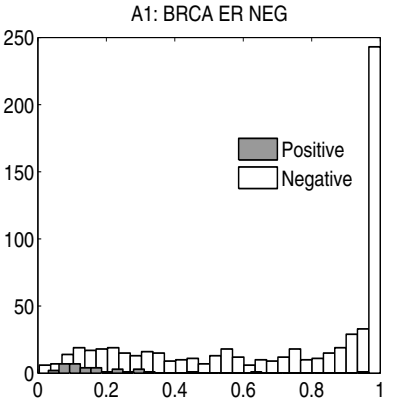


FIGURE 9.6. Breast cancer ER status study. Histogram of π_g for $g \in A_1$ of Table 9.1. Gray/white indicates genes whose expression levels are positively/negatively correlated with ER status, respectively.

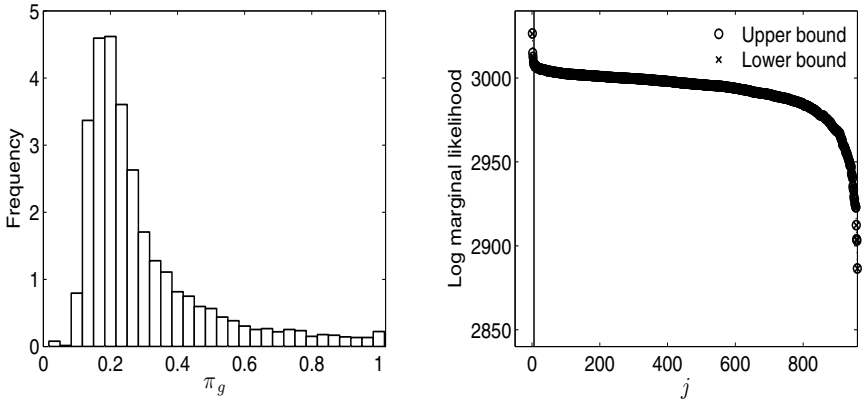
among the highly scoring pathway gene sets, including A_4, A_{11}, A_{14} and A_{18} in Table 9.1. Further, ER− breast tumors tend to be less well differentiated that ER+s, consistent with the novel PROPA identification of the undifferentiated cancer signature A_6. Figure 9.6 shows histogram of the π_g for $g \in A_1$. Genes in A_1 have expression associations with ER that are generally concordant.

9.1.4.2 ErbB2 Pathway

ErbB2 is an epidermal growth factor receptor for which high levels of activity represents a substantial cancer risk factor. About 20-25% of breast cancers have overexpression of ErbB2, primarily due to gene amplification; this is the major cause of ErbB2 pathway deregulation in breast cancers (Ménard et al., 2003; Badache and Gonçalves, 2006). Immunohistochemistry assays of protein levels measure ErbB2 status $(+/-)$ on 146 of the primary breast tumor samples in Carvalho et al. (2008) together with expression data. Analysis using Bayesian factor regression modelling method (BFRM) (Wang et al., 2007) generated posterior probabilities $\Pi = \pi_{1:p}$, as well as the sign of association between gene expression and ErbB2 status for the set of $p = 8,764$ unique genes corresponding to Entrez gene IDs. The π_g are displayed in Figure 9.7(a) and show rather few genes are associated with the experimental factor pathway $\mathscr{F}$, here known to be the ErbB2 related. We re-curated the MSigDB gene lists to align with gene names based on the Entrez human gene database. Since the database does not include signatures explicitly linked to ErbB2, we curated two additional gene sets from the literature: first, a *molecular portrait* set of several genes in chromosome 17 linked to ErbB2 over-expression related to amplification (Perou et al., 2000; Sørlie et al., 2001); second, genes differentially expressed with-versus-without over-expression of the ErbB2 protein measured in data from tumors and cell lines, from Bertucci et al. (2004).

PROPA generated upper and pseudo-optimal lower bounds on log marginal likelihoods for each of the 958 gene sets appear in Figure 9.7(b) The decrease in marginal likelihoods notably diminishes after the first four or five pathways, suggesting stronger evidence of association between these few pathways and ErbB2. These pathways (Table 9.2) include the two ErbB2 signatures, ranked 1 and 4. These two curated gene sets are also identified by GSEA in the top "up-regulated" list but are ranked 4 and 6 by GSEA. Random set-based methods just fail to identify these two gene sets. The small numbers of genes in these sets limits GSEA and random set methods. In contrast, PROPA generally demonstrates good sensitivity and specificity when transcriptional evidence of phenotype-pathway association is relatively weak in the sense of small numbers of genes in the reference gene lists.

Table 9.3 gives information on the ErbB2 molecular portrait reference gene set A_1. This includes pathway membership inference via the π_g^* values and their corresponding log Bayes' factors as well as the initial π_g. Six genes in the chromosomal regions 17q11-q12 and 17q21 have relatively high probabilities π_g of positive association with the experimental breast tumor ErbB2 factor pathway $\mathscr{F}$. The posterior membership probabilities of these genes confirm their membership in the molecular



(a) (b)

FIGURE 9.7. Breast cancer ErbB2 status study. (a) Histogram of association probabilities π_g. (b) Log marginal likelihood upper ($\circ$) and lower bounds ($\times$); pathways are sorted in order of decreasing upper bound and the vertical line indicates the top 6 pathways.

TABLE 9.2. Summary of top ErbB2-related pathways identified by PROPA.

Rank	Pathway	Size	Log(ML): UB	Log(ML): LB	UB-LB
1	ERBB2 overexpression cluster genes	9	3026.53	3026.29	0.24
2	Human Tissue Kidney	11	3014.87	3012.41	2.46
3	Croonquist IL6 Starve Up	31	3012.64	3012.59	0.05
4	ERBB2 gene expression signature	24	3009.77	3009.73	0.04
5	HDACI Colon Colon Cur16HRS DN	8	3008.42	3007.66	0.76
6	MMS Human Lymph Low 4HRS DN	16	3007.84	3007.81	0.03

TABLE 9.3. Genes in the ErbB2 molecular portrait gene set.

g	Symbol	Description	Gene ID	CHR. Loc.	π_g	π_g^*	Log(BF)	Corr.
1	STARD3	START domain containing 3	10948	17q11-q12	0.99	1.00	10.07	+
2	GRB7	growth factor receptor-bound protein 7	2886	17q12	0.99	1.00	10.07	+
3	THRAP4	thyroid hormone receptor associated protein 4	9862	17q21.1	0.96	0.99	6.09	+
4	ERBB2	v-erb-b2 oncogene homolog 2	2064	17q11.2-q12	0.94	0.99	4.34	+
5	TRAF4	TNF receptor-associated factor 4	9618	17q11-q12	0.90	0.92	1.60	+
6	FLOT2	flotillin 2	2319	17q11-q12	0.88	0.72	0.12	+
7	PCGF2	polycomb group ring finger 2	7703	17q12	0.57	0.00	−16.19	+
8	MMP15	matrix metallopeptidase 15	4324	16q13-q21	0.34	0.00	−30.78	+
9	SMARCE1	SWI/SNF related regulator of chromatin	6605	17q21.2	0.21	0.00	−42.19	−

portrait biological pathway $\mathscr{A}_1$. The other three genes with relatively low association probabilities are inferred by PROPA as false positive genes. Notably, gene MMP15 is located at 16q13-q21. It was included in the ErbB2 portrait gene set by a gene clustering analysis based on microarray data; we conclude that MMP15 should *not* be designated a member of the ErbB2 pathway. Several others genes not listed (G6PC, ERAL1, OMG, RPL19, CRKRS) are located in the regions 17q11-q12 and 17q21, and each has positive correlation with ErbB2 status. The Bayes' factors for pathway membership on these genes are greater than 34 dBs, indicating very strong if not decisive evidence for these genes being false negatives, i.e., they *are* members of the ErbB2 pathway.

9.1.5 Theoretical and Algorithmic Details

9.1.5.1 PROPA Model Marginal Likelihood

Refer to the marginal likelihood function shown in (9.1.3). Integrating out $\beta_{1:p}$ and $\alpha_{0:1}$ results in

$$\begin{aligned} &p(\Pi, z_{1:p}|A, \mathscr{F} = \mathscr{A}) \\ &= c(\Pi, z_{1:p}) \prod_{g=1}^{p} \left[\left(\frac{r_A}{\pi_g}\right)^{z_g} \left(\frac{1-r_A}{1-\pi_g}\right)^{1-z_g} \right]^{I(g\in A)} \left[\left(\frac{r_B}{\pi_g}\right)^{z_g} \left(\frac{1-r_B}{1-\pi_g}\right)^{1-z_g} \right]^{I(g\notin A)}, \end{aligned}$$

where $c(\Pi, z_{1:p}) = \gamma_{1:p}(\nu_1)\gamma_{1:p}(\nu_0)\lambda_1^{-\nu_1}\lambda_0^{-\nu_0}(1-\Psi(1;\nu_0,\lambda_0))(1-\Psi(1;\nu_1,\lambda_1))$ with $\nu_1 = \sum_{g=1}^{p} z_g$, $\nu_0 = \sum_{g=1}^{p}(1-z_g)$, $\lambda_1 = -\sum_{g=1}^{p}(z_g \log \pi_g)$, $\lambda_0 = -\sum_{g=1}^{p}(1-z_g)\log(1-\pi_g)$, and Ψ are gamma cdfs. Then the marginal likelihood is $p(\Pi|A, \mathscr{F} = \mathscr{A}) = \sum_{z_{1:p}} p(\Pi, z_{1:p}|A, \mathscr{F} = \mathscr{A})$, where each summand can be evaluated.

The alternative expression derived by summation over the $z_{1:p}$ and integration over $\beta_{1:p}$ conditional on $\alpha_{0:1}$ is

$$p(\Pi|A, \mathscr{F} = \mathscr{A}) = \int_{\alpha_{0:1}} p(\Pi|\alpha_{0:1}, A, \mathscr{F} = \mathscr{A}) p(\alpha_{0:1}) d\alpha_{0:1},$$

where $p(\Pi|\alpha_{0:1}, A, \mathscr{F} = \mathscr{A}) = \prod_{g=1}^{p} p(\pi_g|\alpha_{0:1}, A, \mathscr{F} = \mathscr{A})$. The terms here are

$$p(\pi_g|\alpha_{0:1}, A, \mathscr{F} = \mathscr{A}) = \begin{cases} r_A f_1(\pi_g|\alpha_1) + (1-r_A) f_0(\pi_g|\alpha_0), & g \in A, \\ r_B f_1(\pi_g|\alpha_1) + (1-r_B) f_0(\pi_g|\alpha_0), & g \notin A. \end{cases}$$

9.1.5.2 Marginal Likelihood Upper and Lower Bound Theory

For model M and data D, marginal likelihood in a general form is

$$p(D|M) = \int_{\Theta} p(\theta, D|M)\mathrm{d}\theta,$$

with $\theta = \{\theta_1, \ldots, \theta_K\} \in \Theta$ representing model parameters. In PROPA, data D is Π, model M is specified with $A, \mathscr{F} = \mathscr{A}$, and parameters are $z_{1:p}$ in the reduced form.

For any density function $q(\theta;\gamma)$ parameterized by $\gamma = \{\gamma_1, \ldots, \gamma_J\} \in \Gamma$ and with the same support as the posterior for θ, $p(\theta|D,M)$, Jensen's inequality

$$\log p(D|M) \geq \int_{\Theta} q(\theta;\gamma) \log \frac{p(\theta, D|M)}{q(\theta;\gamma)}\mathrm{d}\theta$$

provides a lower bound for log marginal likelihood. Maximization of this lower bound corresponds to minimization of the Kullback-Leibler divergence of model parameter posterior density $p(\theta|D,M)$ from the variational density $q(\theta;\gamma)$.

If $q(\theta;\gamma) = p(\theta|D,M)$, we can rewrite

$$\log p(D|M) = \int_{\Theta} p(\theta|D,M) \log p(D,\theta|M)\mathrm{d}\theta - \int_{\Theta} p(\theta|D,M) \log p(\theta|D,M)\mathrm{d}\theta.$$

Combining this expression with Gibbs' inequality

$$-\int_{\Theta} p(\theta|D,M) \log p(\theta|D,M)\mathrm{d}\theta \leq -\int_{\Theta} p(\theta|D,M) \log q(\theta;\gamma)\mathrm{d}\theta$$

leads to

$$\log p(D|M) \leq \int_{\Theta} p(\theta|D,M) \log \frac{p(\theta, D|M)}{q(\theta;\gamma)}\mathrm{d}\theta,$$

which provides an upper bound on the log marginal likelihood.

9.1.5.3 Monte Carlo Variational Algorithm

The Monte Carlo variational method using stochastic approximation to generate estimates of the lower bound of marginal likelihoods in the PROPA model has the key steps below. The resulting algorithm is easy to implement, and its convergence can be guaranteed as described, in more general contexts, in Ji, Shen, and West (2009). In essentials here, it is first easy to see that the global, lower bound optimizing value of $\gamma_{1:p}$ satisfies $f_g(\gamma_{1:p}) = 0$, $g = 1, \ldots, p$ for the function defined in (9.1.9). The method is based on the observations that:

1. $f_g(\gamma_{1:p})$, $g = 1, \ldots, p$ is an expectation with respect to $z_{1:p} \sim q(z_{1:p}|\gamma_{1:p})$. Monte Carlo averaging can efficiently estimate this expectation at any value of $\gamma_{1:p}$; in our model this simply involves generating repeat Monte Carlo sample of p independent Bernoulli variates; and
2. the resulting Monte Carlo estimate of $f_g(\gamma_{1:p})$, $g = 1, \ldots, p$ can be used to derive updated values of $\gamma_{1:p}$ using stochastic approximation (Robbins and Monro, 1951).

The algorithmic implementation of these ideas is as follows:

- Begin at iterate $t = 0$ with values of $\gamma_{1:p} = \bar{z}_{1:p}$, the approximate posterior means from the MCMC posterior sample.
- At any later iterate $t \geq 1$ based on current values $\gamma_{1:p}^{(t-1)}$, generate a random sample of $z_{1:p}$ from $q(z_{1:p}|\gamma_{1:p}^{(t-1)})$;
- Compute the implied Monte Carlo estimate of $f_g^{(t-1)}(\gamma_{1:p}^{(t-1)})$, $g = 1, \ldots, p$ replacing the sum in (9.1.9) with the Monte Carlo average over the samples of $z_{1:p}$;
- Update via the stochastic approximation form

$$\gamma_{1:p}^{(t)} = \gamma_{1:p}^{(t-1)} + s^{(t)} f_{1:p}^{(t-1)}(\gamma_{1:p}^{(t-1)}),$$

  where $s^{(t)}$ is a chosen sequence of weights whose sum over $t \geq 1$ diverges but for which the sum of squared values is finite, e.g., $s^{(t)} = c/t$ for some constant $c > 0$.

This is an example of a general algorithm for which it can be shown (Robbins and Monro, 1951; Ji, Shen, and West, 2009) that $\gamma_{1:p}^{(t)}$ converges with probability one to $\gamma_{1:p}^*$ satisfying $f_g(\gamma_{1:p}^*) = 0$, $g = 1, \ldots, p$, providing an iterative approximation of the lower bound optimizing value. Terminate the iterates at some finite step assuming $\gamma_{1:p}^* \approx \gamma_{1:p}^{(t)}$, draw a final, large Monte Carlo sample $z_{1:p}^h$, $(i = 1, \ldots, H)$, from $q(z_{1:p}|\gamma_{1:p}^*)$, and then evaluate the Monte Carlo estimate of lower bound

$$\bar{L} = H^{-1} \sum_{h=1}^{I} \{\log p(\Pi, z_{1:p}^h | A, \mathscr{F} = \mathscr{A}) - \log q(z_{1:p}^h | \gamma_{1:p}^*)\}.$$

This is a consistent estimate of the optimal lower bound assuming the stochastic approximation estimate has converged (Ji, Shen, and West, 2009).

9.1.6 Summary Comments

PROPA is a formal, fully Bayesian framework for matching experimental signatures of structure or outcomes in gene expression — represented in terms of weighted gene lists — to multiple biological pathway gene sets from curated databases. In the setting here, gene weights are explicit gene-factor phenotype association probabilities. The analysis delivers estimated marginal likelihood values over pathways for each factor phenotype, allowing quantitative assessment and ranking of pathways putatively linked to the phenotype as well as gene-specific posterior membership probabilities. We develop a novel Monte Carlo variational method for estimating marginal likelihoods for model comparisons, and evaluate and illustrate the model with simulated and cancer genomic data.

For the future, there is a key need for improved quality of biological pathway databases, an area that PROPA can contribute to as we have exemplified. Open methodological issues include specification of model priors across pathways, and

the use of alternative, multiple numerical summaries of the relationships between genes and experimental phenotypes. Advances in these areas will enhance the contributions of Bayesian reasoning in biological pathway studies. Software for practitioners is also key; current PROPA code, with examples, is freely available at the URL below.

Acknowledgments: We are grateful to Ashley Chi and Chunlin Ji for discussions and input, and to Quanli Wang for computational contributions. This work was partly supported by NSF (DMS-0102227, DMS-0342172) and NIH (U54-CA-112952-01). Any opinions, findings, and conclusions or recommendations expressed in this work are those of the authors and do not necessarily reflect the views of the NSF or NIH.

Software: http://www.stat.duke.edu/research/software/west/propa

9.2 Bayesian Methods for Network-Structured Genomics Data

Stefano Monni and Hongzhe Li

One of the main problems in biological research is the identification of genetic variants such as single nucleotide polymorphisms (SNPs) or gene expression levels that are responsible for a clinical phenotype such as disease status. The problem can in general be formulated as a variable selection problem for regression models. To deal with high-dimensionality, many statistical methods have been developed, including Lasso (Tibshirani, 1996) and its many extensions such as fused lasso (Tibshirani et al., 2005), adaptive lasso (Zou, 2006), group lasso (Yuan and Lin, 2006), SCAD (Fan and Li, 2001), the elastic net (Zou and Hastie, 2005), LARS (Efron et al., 2004), and the Dantzig selector (Candes and Tao, 2007). These methods are mainly based on the idea of regularization. Alternatively, variable selection has also been developed and extensively studied in a Bayesian framework, especially for linear or generalized linear models (George and McCulloch, 1993, 1997; George, 2000). Hans, Dobra, and West (2007) developed shotgun stochastic search in regression with many predictors in order to make the Bayesian variable selection procedures applicable and feasible to the analysis of genomic data. Bayesian formulations of some regularized procedures are also available: a Bayesian Lasso, for example, has been developed recently in (Park and Casella, 2008). Many of these methods have also been employed to analyze genomic data, especially microarray gene expression data in order to identify the genes that are related to a certain clinical or biological outcome.

One limitation of all these popular approaches is that often the methods are developed purely from computational or algorithmic points without utilizing any prior biological knowledge or information and thus important structures of the data may be ignored. For many complex diseases, especially for cancers, a wealth of biological knowledge (e.g., pathway information) is available as a result of many years of intensive biomedical research. This large body of information is now pri-

marily stored in databases on different aspects of biological systems. Some well-known pathway databases include KEGG, Reactome (www.reactome.org), BioCarta (www.biocarta.com), and BioCyc (www.biocyc.org). Of particular interest are gene regulatory pathways that provide regulatory relationships between genes or gene products. These pathways are often interconnected and form a web of networks, which can then be combined and represented as a graph, the vertices of which are genes or gene products and the edges representations of inter-gene regulatory relationships of some kind. This information is a useful supplement to the standard numerical data collected from an experiment. Incorporating the information from these graphs into a data analysis is a non-trivial task, which is generating increasing interest. In genome-wide association studies, the SNPs are often in linkage disequilibrium (LD) and are therefore dependent. Li, Wei, and Maris (2009) introduced the idea of weighted LD graphs based on the pair-wise r^2 statistics between the SNPs. The problem we encounter is that the predictors are constrained on a graph and the challenge we face is to incorporate these constraints in the regression analysis. Motivated by a Gaussian Markov random field prior on the regression coefficients, Li and Li (2008) proposed a network-constrained regularization procedure to incorporate the network-structure information into the analysis, and demonstrated gain in sensitivity in identifying the relevant genes. In the Bayesian context, Li and Zhang (2008) proposed a variable selection for Gaussian linear models with structured covariates using an Ising prior and a Gibbs sampling. Tai and Pan (2009) put forward a similar approach using several different Markov random field priors. In this section we consider a Bayesian variable selection method that takes account of the fact that the covariates are measured on a graph for both linear Gaussian and probit models. Because prior distributions model our *a priori* knowledge of the data, the network structure is introduced in a very natural way at the level of prior probabilities. We consider here an Ising prior, as in Li and Zhang (2008). An Ising model was also used for network-based analysis in Wei and Li (2007). In addition, we implement an MCMC sampler for estimating the posterior probabilities that a variable is selected that is based on the Wolff algorithm (Wolff, 1989). This algorithm was introduced to eliminate the critical slowing down of local updating schemes in Ising models, and is extremely natural in this problem, as we hope will be clear.

The rest of the section is organized as follows. In Section 9.2.1, we formulate the problem in the context of Bayesian variable selection and describe the models, the prior probability distributions, and the algorithm used for inference. In Section 9.2.2, we report the results of some applications of the method to simulated data sets and to a real data set. Finally, we make some comments and present some discussions.

9.2.1 Bayesian Variable Selection with a Markov Random Field Prior

From a statistics viewpoint, we are interested in the problem of Bayesian variable selection in the case in which the data enjoy a graphical representation. Namely,

variables have pairwise relations, which are represented as edges in a graph whose nodes represent the variables. We assume the network to be simple and undirected, i.e., that the relations are among pair of distinct variables and are symmetric (if the variable i is related to j, then j is related to i). If one is able to assess the relative strength of the pairwise interactions, one can furnish the edges with a quantitative label (a weight) that measures such strengths. When such an assessment is not possible, the only information an edge encodes is the existence of the interactions. Both situations are possible and will be taken into account in our model.

To fix notation, let $X = (X_1, \ldots, X_p)$ be the vector of p-covariates and Y the binary or continuous outcome. Each variable is measured on N samples. We denote by $\mathbf{Y} = (y_1, \ldots, y_N)^T$ the vector of responses, by $\mathbf{X} = (x_{ij})$ the $N \times p$ matrix of covariate values, and by $\mathbf{x}_i^T = (x_{i1}, \ldots, x_{ip})$, with the super-script T being transposition, the i^{th} row of the covariate matrix, that is, the values of the covariates for the i^{th} sample. Finally, we let (G_{ij}) be the adjacency matrix of the network. For unweighted networks

$$G_{ij} = \begin{cases} 1 \text{ if} X_i \text{and} X_j \text{are related}, & i, j = 1, \ldots, p, \\ 0 \text{ otherwise.} \end{cases}$$

The assumption that the network is simple and undirected is tantamount to (G_{ij}) being symmetric and having zeros along the diagonal.

In our approach the network structure will be taken into consideration in the choice of prior distributions and in the Markov chain used for the inference. The rest of the formalism is quite common, and will be sketched here to make the exposition self-contained. We first describe the models used to relate the outcome Y to the covariates when Y is binary or continuous. We then detail the inferential strategy.

9.2.1.1 Likelihood and Prior Distributions

Binary outcomes can be modeled in many ways. Here, we consider a probit model. This choice allows us to write marginalized quantities in a manageable form. In this model, the responses are assumed to be independent samples of Bernoulli distributions

$$Y_i | \beta, \mathbf{X} \sim \mathsf{Bernoulli}(\mu_i) \quad i = 1, \ldots, N. \tag{9.2.1}$$

The probability μ_i of success ($y_i = 1$) is related to a linear combination of the covariates (linear predictor) by the following relation:

$$\mu_i = \Phi(x_i^T \beta),$$

where Φ is the cumulative distribution function of the standard normal distribution. Alternatively, if the outcome is continuous, we consider instead a Gaussian linear model

$$Y_i = y_i | \beta, \mathbf{X}, \sigma^2 \sim N(x_i^T \beta, \sigma^2 I). \tag{9.2.2}$$

We assume that some predictors have negligible coefficients β, which will then be considered zero. Each model will thus be labeled by a vector of latent binary variables $\gamma = (\gamma_1, \ldots, \gamma_p)^T$, with each component γ_i being 1 (0) if the corresponding variable X_i is (not) present in the model: namely, $\gamma_i = 0$ if and only if $\beta_i = 0$. Accordingly, we denote by $p_\gamma = \sum_i \gamma_i$ the number of variables, by $\mathbf{X}_\gamma$ the matrix $N \times p_\gamma$ obtained from $\mathbf{X}$ by removing any column i such that $\gamma_i = 0$. We leave explicit the intercept β_0 in the linear predictor so that $\mathbf{X}_\gamma$ will in fact be the $N \times (1 + p_\gamma)$ matrix of covariates, with the first column being a vector of 1, with an abuse of notation.

The expressions (9.2.1) and (9.2.2) are the likelihood functions for our models, which share the parameters β. The normal model has the additional parameter σ^2, the residual variance.

We now specify the prior distributions. For the regression coefficients, we choose the commonly used prior

$$\beta_\gamma | \gamma \sim N(m_\gamma, \Sigma_\gamma), \tag{9.2.3}$$

where m_γ is the mean and Σ_γ is the covariance matrix. The prior distribution of the parameters γ will instead be non-standard. Indeed, the γ_i are generally chosen independent, *e.g.* samples from a multivariate Bernoulli distribution, with probabilities $w_i = P(\gamma_i = 1)$ either predetermined or (usually Beta) random samples. Here, instead, we do not make the latter assumption as we want to take account of the network structure. This is the first difference with the usually proposed Bayesian variable selection models. Namely, we want a probability measure that enjoys the Markov property, that is, we assume the conditional probability that a variable i is in the model to depend only on its neighbors. In addition, we impose the stronger requirement that the probability that a variable is selected be greater if its neighbors are also selected. These conditions are satisfied by the following distribution

$$\pi(\gamma | \mathbf{J}) \propto \exp\left\{ \sum_{i<j} G_{ij} \delta(\gamma_i, \gamma_j) J_{ij} \right\} \cdot \rho^{-\sum_i \gamma_i}, \tag{9.2.4}$$

where δ is the Kronecker delta, ρ and $J_{ij} \geq 0$ are non-negative real numbers. The omitted normalization constant is the sum over all γ configurations. One may have recognized the form of the prior (9.2.4) as defining an Ising model. The parameter ρ is chosen greater than one so as to penalize large models. As for the interaction terms J_{ij}, the simplest model is that with all set equal to a constant J_0. If the network is a weighted network, J_{ij} can be chosen equal to the weights. A particular interesting case is that in which the correlation structure of the covariates is used to define $\mathbf{J}$: $J_{ij} \propto \left|\mathrm{Corr}(X_i, X_j)\right|$. With this choice, variables that are linked in the network are a priori forced to be simultaneously inside the model (or outside the model) with a probability that is higher for variables that are more highly correlated. It would also be interesting to consider the case where J_{ij} are random samples from a distribution $\pi(\mathbf{J})$ (that is, a random Ising model). There are some computational difficulties associated with this situation. For example, the dependence of the normalization constant of the prior (9.2.4) on J makes it difficult to find a prior distribution that leads to a conditional distribution completely available in its analytic form.

9.2.1.2 Posterior Distributions

Once the likelihood and the prior distributions are specified, we can apply the Bayes' formula to obtain the posterior probability. Since the main goal of our analysis is to determine which variables enter the model, we can do away with the sampling of the regression coefficients, and average over them to compute marginalized posterior probabilities. For continuous responses, we use the following values for the parameters of the prior distribution (9.2.3) of the regression coefficients

$$m_\gamma = \mathbf{0}, \quad \Sigma_\gamma = \tau\sigma^2(X_\gamma^T X_\gamma + \lambda I_\gamma)^{-1}, \tag{9.2.5}$$

and assume the variance σ^2 to be a random variable distributed according to the law

$$\pi(\sigma^2) \propto \frac{1}{\sigma^2}.$$

The prior for the regression coefficients with parameters (9.2.5) reduces to that of Smith and Kohn (1996) when $\lambda = 0$, which is related to Zellner's g-prior (Zellner, 1986). We fix $\tau = N$. For possible implications of the values of τ in model selection, we refer the reader to Chipman, George, and McCulloch (2001). The constant λ in the covariance matrix is introduced so that L_γ can be computed even when the number of selected variables p_γ is larger than the sample size N.

With these choices, the posterior distribution is

$$p(\gamma, \mathbf{J}|\mathbf{Y}) \propto p(\mathbf{Y}|\gamma, \mathbf{J}, \mathbf{X})\pi(\gamma|\mathbf{J}),$$

where

$$\begin{aligned} p(\mathbf{Y}|\gamma, \mathbf{J}, \mathbf{X}) &= \int \mathrm{d}\sigma^2 \mathrm{d}\beta \, p(\mathbf{Y}|\gamma, \beta, \mathbf{J}, \mathbf{X}, \sigma^2)\pi(\sigma^2)\pi(\beta_\gamma|\gamma) \\ &\propto (Y^T L_\gamma Y)^{-N/2} \frac{1}{\sqrt{\det R_\gamma}}, \end{aligned}$$

with L_γ and R_γ being the matrices

$$L_\gamma = I - \frac{\tau}{\tau+1} X_\gamma \left(X_\gamma^T X_\gamma + \frac{\lambda}{\tau+1} I \right)^{-1} X_\gamma^T,$$

and

$$R_\gamma = I + \tau X_\gamma^T X_\gamma \left(X_\gamma^T X_\gamma + \lambda I \right)^{-1}.$$

For the binary case, we follow Albert and Chib (1993) and introduce N Gaussian latent variables Z_i $i = 1, \ldots, N$ in terms of which the responses $\mathbf{Y}$ are recovered via the relation: $Y_i = I(Z_i > 0)$, with I being the indicator function. As a consequence, one now needs to consider the joint posterior probability of the parameters of the model and of the latent variables, which is

$$p(\mathbf{Z},\beta,\gamma,\mathbf{J}|\mathbf{Y}) \propto \prod_{i=1}^{N} f(Y_i|Z_i) f(Z_i|\beta_\gamma,\gamma)\pi(\beta_\gamma|\gamma)\pi(\gamma|\mathbf{J})$$

with

$$f(Z_i|\beta_\gamma,\gamma) = \frac{1}{\sqrt{2\pi}} \exp\Big\{-\frac{1}{2}(Z_i - \mathbf{x}_{i\gamma}^T\beta)^2\Big\}, \tag{9.2.6}$$

and

$$f(Y_i|Z_i) = P(Y_i = y_i|Z_i) = I(z_i > 0)\delta(y_i,1) + I(z_i \le 0)\delta(y_i,0).$$

The marginalized joint distribution of γ and $\mathbf{Z}$

$$p(\mathbf{Z},\gamma,\mathbf{J}|\mathbf{Y}) = p(\mathbf{Z}|\gamma,\mathbf{J},\mathbf{Y})\pi(\gamma|\mathbf{J})$$

is expressed in term of the marginalized distribution of $\mathbf{Z}$, which we now compute. Choosing the values (9.2.5), with $\sigma^2 = 1$, for the parameters of the prior distribution (9.2.3) of the regression coefficients, we have

$$\begin{aligned} & p(\mathbf{Z}|\gamma,\mathbf{Y},\mathbf{J}) \\ \propto & \int \prod_{i=1}^{N} f(Y_i|Z_i) f(Z_i|\beta_\gamma,\gamma)\pi(\beta_\gamma|\gamma)\mathrm{d}\beta_\gamma \\ \propto & \frac{\exp\{-\frac{1}{2}Z^T L_\gamma Z\}}{\sqrt{\det R_\gamma}} \prod_{i=1}^{N} \Big\{ I(z_i > 0)\delta(y_i,1) + I(z_i \le 0)\delta(y_i,0) \Big\}, \end{aligned} \tag{9.2.7}$$

with the matrices L_γ and R_γ as above.

9.2.1.3 Markov Chain Monte Carlo Inference

We have determined the marginalized posterior probabilities up to a normalizing constant, which can not be computed. To deal with this problem and identify high-probability models, we consider a Metropolis algorithm for γ. In the binary case, we additionally draw Z from the conditional distribution which can be read out from (9.2.7). It is in the Metropolis algorithm where we make use of the network structure. Namely, we apply the algorithm devised by Wolff (1989). We randomly select a variable, i, and construct a cluster of nodes $Cl(i)$ around it iteratively and stochastically. Each neighbor j of each node k in the cluster is added to the cluster with probability $p_{kj} = G_{kj}\delta(\gamma_k,\gamma_j)\lambda_{kj}$. The cluster $Cl(i)$ initially contains only the vertex i and is iteratively grown until no neighbor is available to be added to the cluster. $Cl(i)$ is therefore composed of nodes that have all the same gamma values as i. Each proposed move is $\gamma \to \gamma'$ with

$$\gamma'_k = \begin{cases} \gamma'_k + 1 \mod 2 & \text{if } k \in \mathrm{Cl(i)} \\ \gamma'_k & \text{otherwise.} \end{cases}$$

It is clear that if the randomly chosen variable i has no neighbors, it is the only one that is added to (if $\gamma_i = 0$) or removed from (if $\gamma_i = 0$) the present model to obtain the proposed model. In our implementation, we alternate a proposal to add variables to the model with a proposal to remove variables from it. The proposed configuration γ' is accepted with probability $F(z) = \min\{1, z\}$, where

$$z = \frac{p(\mathbf{Z}|\gamma', \mathbf{J}, \mathbf{Y})}{p(\mathbf{Z}|\gamma, \mathbf{J}, \mathbf{Y})} \cdot \rho^{-\Sigma_i(\gamma_i' - \gamma_i)} \tag{9.2.8}$$

for discrete $\mathbf{Y}$, and

$$z = \frac{p(\mathbf{Y}|\gamma', \mathbf{J}, \mathbf{X})}{p(\mathbf{Y}|\gamma, \mathbf{J}, \mathbf{X})} \cdot \rho^{-\Sigma_i(\gamma_i' - \gamma_i)} \tag{9.2.9}$$

for continuous $\mathbf{Y}$. For the above relations (9.2.8, 9.2.9) to hold true, one must choose the proposal probability $\lambda_{ij} = 1 - \exp(-J_{ij})$ because of equation (9.2.4) and the detailed balance condition. For vanishing values of J_{ij} the algorithm reduces to a single-variable updating, as in this case the network is effectively a collection of isolated vertices. Larger values of J_{ij} favor larger clusters, and for sufficiently large values, variables in the same connected component of the network will have the same values of γ. The parameter ρ instead discourages large models. Thus, the choice of J and the choice of ρ affect the realizations of the model. When the network is very complex, it may seem preferable to stop the construction of the cluster $Cl(i)$ about the randomly selected variable i to include only its neighbors up to some distance. For example, one can add to the cluster only the nearest neighbors j of i with probability $\lambda_{ij} = 1 - \exp(-J_{ij})$ without iterating this procedure any further. In this case, the equations (9.2.8) and (9.2.9) do not hold true anymore. Indeed, the first factor of the prior (9.2.4) would give a contribution to the ratios z that would only be partially canceled by the kernel of the proposed move. Ideally, and more naturally, the same goal could be reached by modelling J with a distribution that decreases rapidly with the distance.

The advantage of this collective updating algorithm over single updating algorithms is that very few steps are generally necessary to go from one configuration to an independent one. The Wolff algorithm can be viewed as a one-cluster variant of the Swendsen-Wang algorithm (Swendsen and Wang, 1987), which was applied in variable selection in Nott and Green (2004), and has the advantage of being more easily implemented.

9.2.2 Numerical Examples

We present in this subsection some applications of the method to simulated and real data.

9.2.2.1 Simulated Regulatory Network

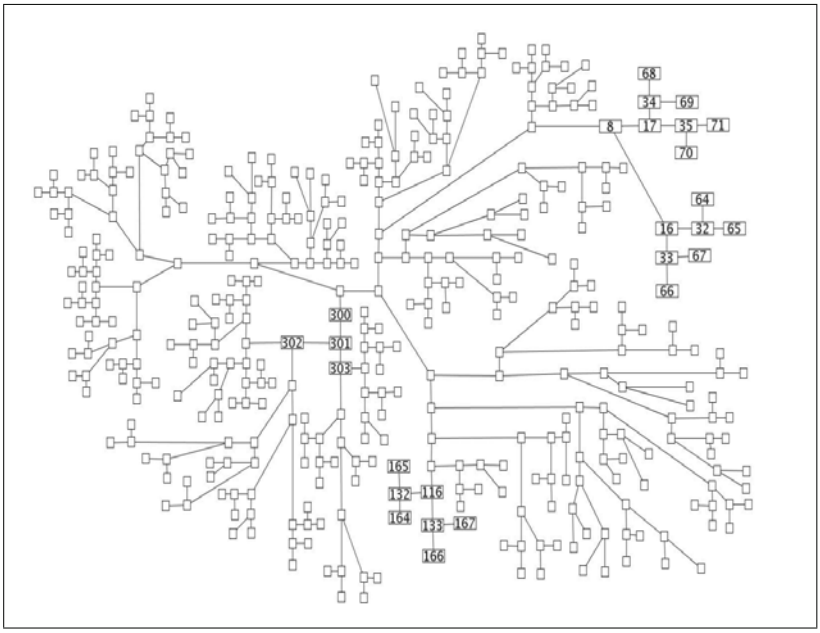


FIGURE 9.8. Tree-structured network used for the first simulated data sets. Rectangular nodes represent the relevant variables.

We have considered a simulation with $p = 399$ covariates, one continuous outcome Y, and the network represented in Figure 9.8. The rectangular nodes represent the variables entering the simulated model, i.e., the variables that are related to the response. The regression coefficients were assigned values according to two different schemes. The assignment was completely random in a first set of simulations, with values drawn uniformly in the interval $I = [-2, -0.5] \cup [0.5, 2]$. In a second set of simulations, the values were chosen in the same interval I, but constrained in such a way that the top node in each group of defining variables had larger values than the rest. For both simulations, the variables X were drawn from a multivariate normal distribution with variance-covariance matrix $\text{Corr}(X_i, X_j) = 0.3^{|i-j|} + G_{ij}0.2^{|i-j|}$. This choice of variance-covariance matrix insures that neighboring variables are a bit more correlated than non-neighboring variables, although the added correlation may be very small. The outcomes were sampled from a normal distribution centered about the linear predictor and with variance σ^2 such that the noise-to-signal ratio (NSR) for the data had values 0.1, 0.3, 0.4, 0.5, 1. We present the results of runs carried out using one of the data sets simulated in one of the two simulation schemes. Similar conclusions are valid for the other data. As one can expect, the lower the noise-to-signal ratio, the easier it is to select the true model. In the easi-

est case NSR = 0.1, all variables were selected but two (variable 300 and variable 302), and two false positives were identified using as a criterion that the posterior probability that a variable i is in the model, $P(\gamma_i = 1)$, is greater than 0.5. Increasing the latter value to 0.75, the two false positives disappear. The same results are obtained in this case if the network structure is ignored, *viz.* if one considers a network of isolated vertices. For NSR = 1, which is the hardest case, only one variable is selected (variable 302) when no network structure is used, while the network helps select few other variables, but at the same time some false positives. This is a pattern that was verified for other values of the noise-to-signal ratio as well: employing the network structure detects more variables of the true model at the expense of introducing some false positives.

TABLE 9.4. Results of simulations using the network represented in Figure 9.8 or without using the network structure for two different noise-to-signal ratios (NSR).

	NSR = 0.4	NSR = 0.3
with network	8, 16, 17, 32, 33, 116 132, 301, 303 (356)	8, 16, 32, 116, 132 164, 300, 301, 302 (90,131,138, 168,261,298)
without network	8, 16, 17, 32, 116 132, 301 (122, 322)	8, 16, 132, 301 (83)

Table 9.4 summarizes the results for NSR = 0.4 and NSR = 0.3. In Table 9.4, the true model consists of variables $8, 16, 17, 32-35, 64-71, 116, 132, 133, 164-167, 300-303$ and the variables listed have a posterior probability of being present in the model greater than or equal to 0.5. Figure 9.9 shows the plots of the true positive rate versus the false positive rate for four values of NSR, which give further illustration of the advantage of employing the network structure. We observed that in general the areas under the ROC curves are higher when the network structures are utilized in the prior distribution and in the MCMC inferences.

9.2.2.2 Simulation Based on a KEGG Regulatory Network

We also considered a data set with a discrete outcome and a more complicated network, with $p = 400$ nodes, which is represented, with the exception of some isolated nodes, in Figure 9.10. This network is a subset of a real KEGG network (Kanehisa and Goto, 2002) that was used in Wei and Li (2007, 2008). We sampled the coupling J for each edge from an exponential distribution, and, starting from a random assignment of values, we decided if a variable was in the model or not using the conditional distributions obtained from (9.2.4)

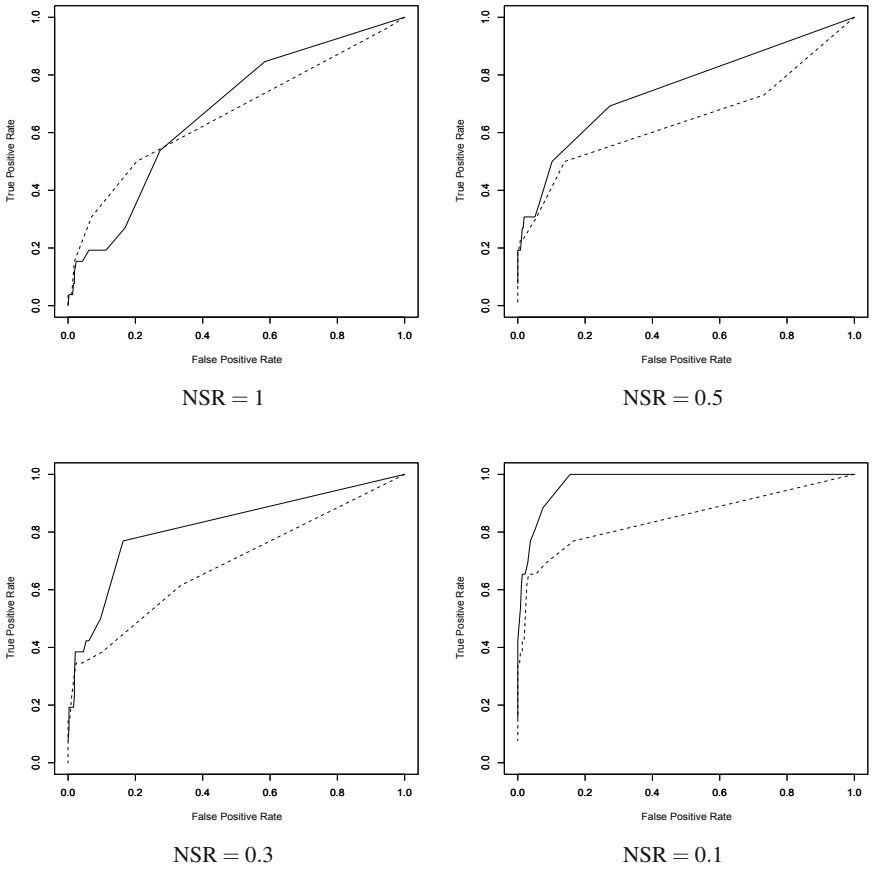


FIGURE 9.9. Regulatory network (tree) example. True positive rates vs. false positive rates for different values of the noise-to-signal ratio (NSR), using the network structure (solid lines) and without using the network structure (dashed lines).

$$P(\gamma_i = 1|\cdots) = K \cdot \exp\Big\{\sum_{j\neq i} G_{ij}\delta(\gamma_j,1)J_{ij}\Big\}\rho^{-1}$$
$$P(\gamma_i = 0|\cdots) = K \cdot \exp\Big\{\sum_{j\neq i} G_{ij}\delta(\gamma_j,0)J_{ij}\Big\}$$

with $K^{-1} = \exp\{\sum_{j\neq i} G_{ij}\delta(\gamma_j,1)J_{ij}\}\rho^{-1} + \exp\{\sum_{j\neq 1} G_{ij}\delta(\gamma_j,0)J_{ij}\}$. The variables selected are depicted as black nodes in Figure 9.10. We then sampled X_{ik} from a normal distribution with variance-covariance matrix $\mathrm{Cov}(X_i,X_j) = G_{ij}/2, i \neq j$ for $k = 1,\ldots,N = 200$, and β, from a uniform distribution in the interval $[-5,-2]\cup$

[2,5], rather than take them from a multi-variate distribution (see (9.2.3)). Finally, for each i, we drew Z_i from (9.2.6) and took as Y_i its sign.

The variables identified by the algorithm are the square and rectangular nodes in Figure 9.10. The two shapes refer to two values of the posterior probabilities used as the criteria for identifying the selected variables, with the rectangles referring to a posterior probability of at least 0.5 and the squares of at least 0.4. We note that all the variables in the models that are isolated nodes have been omitted from Figure 9.10. Three of these variables entered the simulated models and one was correctly identified with a posterior probability greater than 0.5. Other runs gave similar results, with some variations in the variables selected in each true cluster.

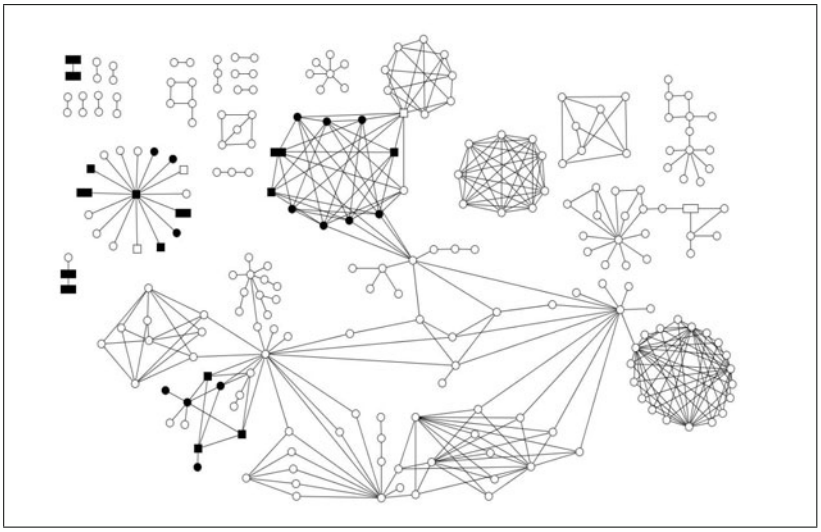


FIGURE 9.10. A subset of the KEGG network used for the second simulation. The black nodes are the relevant variables. Rectangles and squares indicate the selected variables based on a posterior probability of 0.5 and 0.4 or greater, respectively.

9.2.2.3 Application to Real Data

Aging of human brain is one of the most complex biological processes. It is a cause of cognitive decline in the elderly and a major risk factor in age-based degenerative diseases such as Alzheimer's. For this reason, uncovering the genetic underpinning of brain-aging has become the focus of recent research. Indeed, there have been a number of efforts to collect genetic data from brain tissue of individuals of different ages. In particular, Lu et al. (2004) gathered the transcriptional profiling of the human frontal cortex from 30 persons of age ranging from 26 to 106, using the Affymetrix HG-U95Av2 oligonucleotide arrays. In this section, we present the

results of an analysis carried out using these data. Specifically, we set out to identify which genes and which pathways were related to brain aging. To do this, we supplemented Lu's data with pathways information acquired from the KEGG data bases. We first constructed a (non-connected) 1668-node network by combining 33 KEGG regulatory pathways (Kanehisa and Goto, 2002), and then considered only those genes on the U95Av2 chip and those nodes in the network that overlapped and for which data were available on the entire cohort of 30 patients ($N = 30$). This resulted in $p = 1302$ genes and a network with 5258 edges. Our method could have also been applied to the entire genes on the U95Av2 chip by treating those genes as additional isolated nodes for which no pathways information was available. We log-transformed, centered and standardized the data. As responses we used the logarithm (in base 10) of the age.

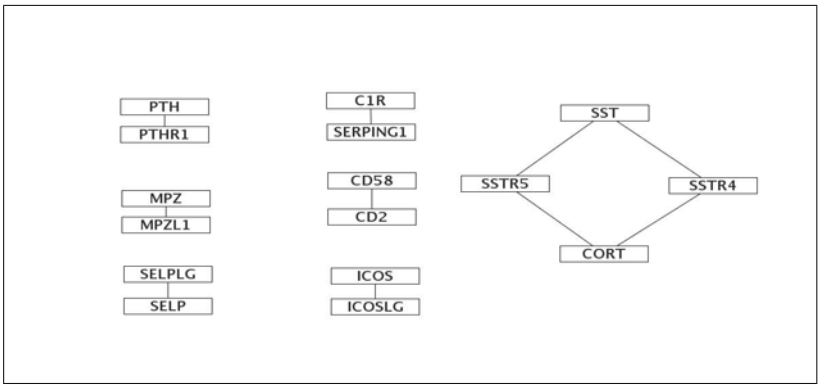


FIGURE 9.11. Subnetwork of the KEGG network whose vertices represent variables with a posterior probability of 0.5 in the real data analysis. Isolated nodes are not represented.

For this analysis, we fixed $J_{ij} = J \cdot \left|\text{Corr}(X_i, X_j)\right|$, so as to favor highly correlated variables that are connected in the network to be jointly in the model. The constants J and ρ were chosen so as to allow very high acceptance rates and reasonable model size. We have considered variables that have a posterior probability of being in the model greater than 0.5: $P(\gamma_i) \geq 0.5$. With this criterion, 44 variables were selected. Figure 9.11 depicts the subnetwork composed of vertices among this set, except for isolated vertices. There are a few interesting observations from these identified subnetworks. First, we identified a small subnetwork with 4 genes including Somatostatin gene (SST) and its receptors (SSTR4 and SSTR5) and another gene cortistatin (CORT) that also binds to the same receptors as SST. Somatostatin is an important regulator of endocrine and nervous system function (Yacubova and Komuro, 2002). Because its levels change with age, it is likely that age-related changes are affected or affect SST (Reed et al., 1999). A role for SST in Alzheimer's disease has also been proposed (Saito et al., 2005). Another interesting pair of genes, the complement component 1 inhibitor gene (SERPINGS) and the complement com-

ponent 1 (C1R), was also reported to be related to aging related phenotypes. For example, Ennis et al. (2008) identified an association between the SERPING1 gene and age-related macular degeneration using a two-stage case-control study. The selenium transport protein, selenoprotein P (SELP), and its ligand (SELPLG), are essential for neuronal survival and function and were reported to be associated with Alzheimer's pathology in human cortex (Bellinger et al., 2008).

9.2.3 Discussion and Future Direction

Motivated by the application of incorporating prior pathway and network structure information into the analysis of genomic data, we have considered Bayesian variable selection for both linear Gaussian models and probit models when the covariates are measured on a graph. In our approach, a flexible Markov random field prior that takes account of the graph structure is employed and a Markov chain sampler based on the Wolff algorithm is used. Our simulations indicate that incorporating the graph structure can lead to increased sensitivity in identifying the relevant variables. The algorithm performs better for continuous than for binary outcomes, as in the latter case sampling of the Gaussian latent variables $\mathbf{Z}$ is required. This chapter focuses on how to utilize the prior genetic pathway and network information in the analysis of genomic data in order to obtain a more interpretable list of genes that are associated with the genotypes. An equally important topic is how to construct these pathways and networks. One area of intensive research in the last several years has been on estimating sparse Gaussian graphical models based on gene expression data (Li and Gui, 2006; Peng et al., 2009). Although such models built from gene expression data can provide some information on how genes are related at the expression level, they hardly correspond to any of the real biological networks. The future will likely see more research on how to build meaningful biological networks by integrating various types of genomic data. This leads to great challenges due to both the complexity of the real biological networks and the high-dimensionality of the genomic data. Again, utilizing the prior network information in the framework of Bayesian analysis can lead to better network inference (Mukherjee and Speed, 2008). Alternative to the Gaussian graphical models, Bayesian networks provide more detailed information on causal relationship among genes based on various types of genomic data. However, the computation is even more challenging given the fact that a very large model space has to be explored and novel MCMC methods are required (Ellis and Wong, 2008). Finally, as more and more biological networks are accumulated, statistical methods for analysis of these large graphs are also needed. Some interesting problems include the identification of network modules and network motifs. Here as well, Bayesian approaches seem to provide important solutions to these problems (Berg and Lassig, 2004, 2006; Monni and Li, 2008).

Acknowledgments: This research was supported by NIH grants R01ES009911 and R01CA127334. We thank Professor Ming-Hui Chen and other editors for inviting us to contribute a section to this book in honor of Professor James O. Berger.

9.3 Bayesian Phylogenetics

Erik W. Bloomquist and Marc A. Suchard

Molecular phylogenetics aims to reconstruct and infer the evolutionary history relating N organisms or taxa from molecular data such as DNA or protein sequences that these organisms have shared through their history (Felsenstein, 2003). Additional goals include: the detection of selection, or the evolutionary success of novel variation, the inference of population genetic demographic histories, and the modeling of RNA virus epidemics. In all these studies, the molecular sequence typically encountered can be abstracted into a matrix Y of size $N \times S$. The N rows of Y represent contiguous stretches of molecular information from a specific taxa or organism. The S columns, or phylogenetic sites, Y_s in Y represent homology statements about the sequence characters. In particular, the statements imply that each Y_{sn} in the column Y_s derives from the same common ancestral character some time in the past. Each element Y_{sn} in Y draws from an alphabet $\mathscr{A}$ whose structure depends upon the particular application. For example, in the case of genetic DNA, $\mathscr{A} = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$, where A, C, G, and T signify the four different nucleotide bases that compose DNA. Other common alphabets include the arbitrary set of M distinct morphological traits, the set of amino acids of size 20, and the set of codons of, generally, size 61.

Under the most basic likelihood-based framework, phylogenetic researchers model the column vector Y_s in Y at each site $s = 1, \ldots, S$ as an independent draw from a Markov process $f(\mathrm{Y}_s \,|\, \tau, \mathrm{D})$ where τ is a bifurcating tree topology and D is a continuous-time Markov chain (CTMC). The CTMC $\mathrm{D} = \{\mathrm{D}(t)\}$ dictates the character substitution process in time along the bifurcating tree τ. The CTMC D exists on the state space $\mathscr{A}$ and is governed by the irreducible instantaneous rate matrix Q and stationary distribution π, i.e., $\mathrm{D} = (\mathrm{Q}, \pi)$. Over the past 50 years, numerous models for D have appeared, typically following a progression of more flexible parameterizations for Q (Galtier, Gascuel, and Jean-Marie, 2005). Regardless of these parameterizations, however, researchers rely on the Chapman-Kolmogorov equations to derive the finite-time transition probability matrix $\mathrm{P}(t) = \exp\{\mathrm{Q}t\}$, where $\mathrm{P}_{ij}(t) = p(\mathrm{D}(t) = j \,|\, \mathrm{D}(0) = i)$ equals the probability of the CTMC starting in state $i \in \mathscr{A}$ and ending in state $j \in \mathscr{A}$ at time t. In most situations, the instantaneous rates in Q and time t are confounded, necessitating specific restrictions on Q. Once deriving $\mathrm{P}(t)$, researchers imagine an unobserved character drawn from π at the root of τ, and then the CTMC D acting conditionally independently along each branch of τ to arrive at the probability $f(\mathrm{Y}_s \,|\, \tau, \mathrm{D})$ of observing Y_s, where each element Y_{sn} associates with an external tip on τ (Felsenstein, 1981). Due to the independence of each site Y_s, the full data likelihood $f(\mathrm{Y} \,|\, \tau, \mathrm{D})$ expands into $\prod_{s=1}^{S} f(\mathrm{Y}_s \,|\, \tau, \mathrm{D})$. Suchard and Rambaut (2009) provide a gentle introduction to the calculation of $f(\mathrm{Y}_s \,|\, \tau, \mathrm{D})$ and the derivation of $\mathrm{P}(t)$. Excellent work by Felsenstein (2003) provides a comprehensive overview of molecular evolution in book-length form.

Adopting f as a probabilistic framework for Y, Bayesian methods for posterior inference on τ and D began to flourish in the middle 1990s due to availability of in-

expensive computing and the widespread introduction of Markov chain Monte Carlo (MCMC) sampling techniques in statistics (Huelsenbeck et al., 2001). By adopting relatively vague and initially independent priors for $p(\mathrm{D})$ and $p(\tau)$ phylogenetic researchers soon realized that posterior inference of τ and D

$$p(\tau, \mathrm{D} \mid \mathrm{Y}) = \frac{f(\mathrm{Y} \mid \tau, \mathrm{D}) \times p(\tau) \times p(\mathrm{D})}{\int_{\mathrm{D}} \int_{\tau} f(\mathrm{Y} \mid \tau, \mathrm{D}) \times p(\tau) \times p(\mathrm{D}) \, d\mathrm{D} \, d\tau}, \tag{9.3.1}$$

provided numerous advantages over alternative methods for phylogenetic inference, including methodological, applied, and philosophical considerations (Holder and Lewis, 2003). As an example, the marginal posterior distribution of τ

$$p(\tau \mid \mathrm{Y}) \propto \int_{\mathrm{D}} f(\mathrm{Y} \mid \tau, \mathrm{D}) \times p(\mathrm{D}) \times p(\tau) \, d\mathrm{D}, \tag{9.3.2}$$

allows for easily interpretable probability statements about the topology τ (Sinsheimer, Lake, and Little, 1996) as opposed to the more traditional bootstrapping methodology (Felsenstein, 1985). Moreover, a Bayes factor test B_{10} allows for the direct comparison of non-nested hypotheses H_1 and H_0 regarding the space of bifurcating trees $\mathscr{T}$. In practice, these Bayes factor tests are extremely difficult to estimate, and research continues on more accurate and faster estimation techniques (Suchard, Weiss, and Sinsheimer, 2005). For example, in a recent breakthrough, Lartillot and Phillipe (2006) adopt path sampling to provide more sound estimates of B_{10} than the more commonly used harmonic mean estimator of Newton and Raftery (1994).

Due to the relative ease of inference under the Bayesian paradigm, investigators have made numerous advances beyond the basic statistical model just presented. As a prime example, researchers now allow for variation in the substitution rate D across τ, avoiding the pitfalls of a "strict molecular clock." In particular, researchers extend the parametrization of D in terms of scalar rate multipliers along each branch of τ and model the prior dependence between D and τ via $p(\mathrm{D}, \tau) = p(\mathrm{D} \mid \tau) \, p(\tau)$. In the first instance of this idea under the Bayesian framework, Thorne, Kishino, and Painter (1998) model the "rate of evolution of the rate of evolution" (Gillespie, 1991) using autocorrelated lognormal distributions for the rate multiplier portion of $p(\mathrm{D} \mid \tau)$. Building from this work, Drummond et al. (2006) relax the autocorrelation assumption and instead adopt a discrete, Dirichlet-like, rate model for $p(\mathrm{D} \mid \tau)$, with a nod towards heterochronous data, i.e., data sampled at different points in time. These methods provide vastly improved model fit for empirical datasets. Nevertheless, relaxation techniques remain poorly studied, especially on issues such as overparameterization and identifiability (Rannala, 2002). As such, Bayesian modeling of rate variation will likely remain a major endeavor for the foreseeable future.

In addition to temporal variation along τ, investigators also allow for spatial variation in D along Y. For example, Pagel and Meade (2004) allow the substitution model D_s at each site s to multinomially draw from a latent set of any J arbitrary substitution models $\{\mathrm{D}^{(1)}, \ldots, \mathrm{D}^{(J)}\}$, where these J models may have different rates or even different parameterizations. In another such effort, Huelsenbeck and Suchard

(2007) allow the site-specific substitution process D_s at each site s to have its own overall rate of substitution, but then nonparametrically cluster these rates using the Dirichlet process. Taking a joint modeling approach, Blanquart and Lartillot (2008) combine ideas from Pagel and Meade (2004) and Thorne, Kishino, and Painter, (1998) to jointly model spatial variation in D along Y and temporal variation in D along τ.

These spatial mixture models control for heterogeneity of D along Y. Nevertheless, these models still treat each site Y_s as an exchangeable draw, ignoring the possibly strong linear dependence in the site-specific processes. In many cases, this dependence is of paramount biological interest (Minin et al., 2005). To correct this, two general approaches have been adopted. The first approach relies on a hidden Markov model (HMM) to handle the linear dependence along Y, i.e., investigators assume each site s has its own substitution model D_s, but use a first-order HMM $p(D_1, \ldots, D_S) = \prod_{s=2}^{S} p(D_s \,|\, D_{s-1})\, p(D_1)$ to explicitly model the dependence among these S models (Lehrach and Husmeier, 2009). Avoiding the use of a HMM, Suchard et al. (2003b) adopt a multiple change-point (MCP) that assumes a random partitioning structure ρ along Y that divides D_s into homogeneous segments. The partitioning structure ρ not only has random breakpoint locations, but a random number of breakpoints as well.

More flexible and sophisticated Bayesian models, such as those presented above, have resulted in deeper biological understanding. Nevertheless, researchers have also advised caution in these new modeling endeavors (Huelsenbeck et al., 2002). For example, Mossel and Vigoda (2005) demonstrate that, under specific circumstances, if Y derives from a mixture of two generative distributions, a simplified MCMC sampler can take an exponential amount of time to converge to the stationary distribution. In another such instance, Suzuki, Glazko, and Nei (2002) find that Bayesian posterior inference of τ can tend to give strong posterior support for groupings of taxa that are not present in the true tree. An uniformed evolutionary biologist might view these cautionary tales as reasons to avoid Bayesian inference. However, the unsuspected success of Bayesian modeling in providing biologically realistic models supports the opposite view. In particular, Bayesian inference of molecular evolution provides a rich research endeavor in which the field is only beginning to understand. To alleviate any lingering concern in the phylogenetic community, the development of reference priors (Berger and Bernardo, 1989) is an important avenue of research; Ferreira and Suchard (2008) offer one such example of a conditional reference prior for D.

The models and research highlighted above have transformed molecular phylogenetic research (Holder and Lewis, 2003). The Bayesian revolution has also led to numerous well-developed and freely available software packages that implement many of these advances, e.g., MrBayes and BEAST. Yet, the true power of the Bayesian paradigm–its ability to jointly and simultaneously model the relationships among generative processes–remains relatively unexplored. In the next two subsections, we touch on areas where simultaneous and joint modeling under a Bayesian framework has generated significant success.

9.3.1 Statistical Phyloalignment

The preceding section implicitly assumes that the data likelihood describes a fixed and known alignment Y whereby each column in Y represents a homologous site. In practice however, we do not observe this alignment Y directly. Instead, the raw genetic data consist of separate strings of molecular characters from $\mathscr{A}$ that must first be aligned for homology.

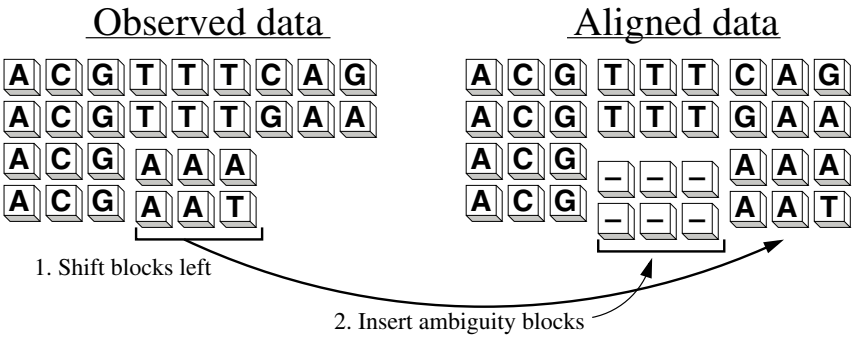


FIGURE 9.12. Statistical phyloalignment. The left hand side displays the raw observed sequence data, and the right hand side displays the aligned homologous data.

Figure 9.12 provides a simple example of this alignment step. Due to the small scale of this alignment, the process can be accomplished by eye in two steps. But when dealing with highly divergent molecular data from whole genomes, especially data affected by the insertion and deletion (indels) of characters over time, aligning for homology becomes extremely challenging (Wong, Suchard, and Huelsenbeck, 2008).

Currently, investigators solve this alignment dilemma in two steps. First, an analyst will enter their raw molecular sequence data into a bioinformatics algorithm that attempts to find an "optimal" alignment Y from the raw data. Then conditioning on this alignment Y, the analyst draws inference on τ and D by considering Y as the observed data. Numerous bioinformatics tools and techniques exist to complete the first step. Historically, ClustalW has the most widely used, but in recent years, a wide variety of purportedly more accurate alignment algorithms and tools have appeared (Edgar and Batzoglou, 2006). Unfortunately, the choice of alignment method can significantly affect the alignment and resulting inference (Wong, Suchard, and Huelsenbeck, 2008). Moreover, depending on the discipline, Morrison (2009) finds that 26% to 78% of applied researchers manually edit the alignment Y after using a computerized algorithm, suggesting that these bioinformatics tools remain far from ideal.

Critically, conditioning on a single alignment can also lead to statistical overconfidence in the second stage of analysis (Lake, 1991). Statisticians have been long

aware of this problem with two-stage analyses. A classic simile occurs with the single imputation strategy for inference of missing data (Rubin, 1978). In many longitudinal and health-related studies, missing observations due to drop-out and non-responses plague investigators and make analysis using standard statistical techniques much more difficult. To combat this issue, investigators commonly adopt imputation to predict the missing values, so that standard statistical techniques can be applied. The techniques used to predict the missing values are statistically valid; nevertheless, this imputation approach ignores uncertainty in the missing responses since only a single point estimate is predicted. As a result, inferential statements tend to be overconfident. To correct this problem, Rubin (1978) instead suggests that multiple values for missing observations should be predicted to properly account for uncertainty.

Uncertainty due to missing data directly relates back to homology alignment and subsequent phylogeny estimation. Just as we must make inference on the missing values to use standard statistical inferential techniques, we must make inference on the raw data Ψ to form an alignment Y. Hence, if we ignore uncertainty in the alignment Y by conditioning on a single "probable alignment," we fall into the same trap as single imputation. Realizing this, statisticians and biologists have just recently begun to estimate Y, τ, D, and an indel model Λ jointly from the raw molecular data Ψ (Lunter et al., 2005; Redelings and Suchard, 2005). so that proper inference can be made. Before moving to these models, we first outline the definition of the observed raw molecular data. Under an abstract formulation, the raw data Ψ consists of N strings of letters $\Psi_1, \ldots, \Psi_N$ from $\mathscr{A}$, where each string has length $|\Psi_n|$. As an example, in Figure 9.12 $\Psi_1 =$ ACGTTTCAG, $\Psi_2 =$ ACGTTTGAA, $\Psi_3 =$ ACGAAA, and $\Psi_4 =$ ACGAAT.

Using this formulation, joint modeling aims to find the posterior distribution of τ, D, Λ, and the alignment Y given the raw data Ψ. The parameters in Λ characterize the random indel processes along τ and Y. The posterior distribution equals

$$p(\tau, \mathrm{D}, \Lambda, \mathrm{Y} \,|\, \Psi) \propto f(\Psi \,|\, \mathrm{Y}, \tau, \mathrm{D}, \Lambda) \times p(\mathrm{Y} \,|\, \tau, \Lambda) \times p(\mathrm{D}, \tau, \Lambda), \tag{9.3.3}$$

where the first term $f(\Psi \,|\, \mathrm{Y}, \tau, \mathrm{D}, \Lambda)$ conveniently computes the same as $f(\mathrm{Y} \,|\, \tau, \mathrm{D})$ (Redelings and Suchard, 2005). The distribution $p(\mathrm{Y} \,|\, \tau, \Lambda)$ provides a prior on the alignment Y and has multiple formulations (Thorne, Kishino, and Felsenstein, 1992; Redelings and Suchard, 2007). Notice that joint inference allows one to treat the alignment Y as a random quantity that can be integrated out of the posterior distribution

$$p(\tau, \mathrm{D}, \Lambda \,|\, \Psi) \propto \sum_{\mathrm{Y}} f(\Psi \,|\, \mathrm{Y}, \tau, \mathrm{D}, \Lambda) \times p(\mathrm{Y} \,|\, \tau, \Lambda) \times p(\mathrm{D}, \tau, \Lambda). \tag{9.3.4}$$

Moreover, Gibbs sampling kernels from Redelings and Suchard (2005) provide an MCMC algorithm that makes this integration step relatively straightforward. Currently, the freely available software packages BALi-Phy and StatAlign allow for joint inference of alignment and phylogeny.

Jointly estimating the alignment and the phylogeny has fundamentally altered research in molecular phylogenetics (Wong, Suchard, and Huelsenbeck, 2008). Some hurdles remain before the approach becomes more widely used, specifically computational hurdles, but in the near future, it is likely that researchers will no longer treat Y as the observed data and unknowingly bias their inference. Redelings and Suchard (2009) provide a more comprehensive review of statistical phyloalignment.

9.3.2 Multilocus Data

The preceding section outlines a Bayesian framework to find the posterior distribution of τ and D from a single contiguous stretch of raw molecular information Ψ. Nowadays, however, researchers typically have information from multiple genes or even entire genomes due to the advances in genetic sequencing. Better put, *phylogenomic* or *multilocus* applications contain continuous stretches of information from M genes or loci, $\mathrm{Y}_1, \mathrm{Y}_2, \ldots, \mathrm{Y}_M$. For ease of presentation, we assume the alignment Y_m is known in this subsection.

Investigators generally agree that the data likelihoods $f(\mathrm{Y}_m \mid \tau_m, \mathrm{D}_m)$ remain adequate models for Y_m on each of the M loci (Rannala and Yang, 2008). More unexplored stands a hierarchical phylogenetic model to formalize the dependence within $p(\tau_1, \tau_2, \ldots, \tau_M)$ that addresses biologically motivated research questions while effectively borrowing strength across loci (Suchard et al., 2003a). Hierarchical phylogenetic models $p(\mathrm{D}_1, \ldots, \mathrm{D}_M)$ are relatively straightforward because a finite set of exchangeable parameters on each loci M govern D_m (Suchard et al., 2003a).

To borrow strength across loci, researchers initially suggested one of two competing approaches: the *supermatrix* or *concatenation* approach (Kluge, 1989) or the *supertree* or *independent* (Bininda-Emonds, 2005) approach. Under the tenants of the supermatrix approach, researchers ignore the partitioning in the data, and set $\tau_1 = \cdots = \tau_M = \tau$. Under the alternative supertree methodology, researchers adopt a completely independent prior, and pool no information about $\tau_1, \ldots, \tau_M$; instead, researchers rely on descriptive tools to form *a posteriori* consensus estimates of $\tau_1, \ldots, \tau_M$. As might be suspected, both of these approaches possess serious statistical and biological flaws (Rannala and Yang, 2008). In the supermatrix methodology, assumption of a single underlying τ ignores the presence of recombination, reassortment, gene flow, hybridization, and other such non-vertical processes. Under the other alternative supertree approach, pooling no information across loci leads to noisy results that are difficult to interpret (Rokas et al., 2003).

To bridge these two extremes, Suchard et al. (2003b), Suchard (2005), and Ané et al. (2007) stochastically cluster the tree topologies τ_m from each loci through $p(\tau_m \mid \Upsilon)$, where Υ characterizes this clustering process. Suchard et al. (2003b) adopt a multinomial process for $p(\tau_m \mid \Upsilon)$, where the form of Υ is hypothesis driven. Ané et al. (2007) parameterize $p(\tau_m \mid \Upsilon)$ under a flexible Dirichlet process. Lastly, Suchard (2005) models the dependence amongst the trees using a stochastic random walk process across tree space $\mathscr{T}$, with a particular emphasis on horizontal gene

transfer. All these approaches help bridge the gap between the supertree-supermatrix ideologies. Nevertheless, the three models for $p(\tau_m \mid \Upsilon)$ remain statistical abstractions without a direct biological mechanism as to why trees may differ.

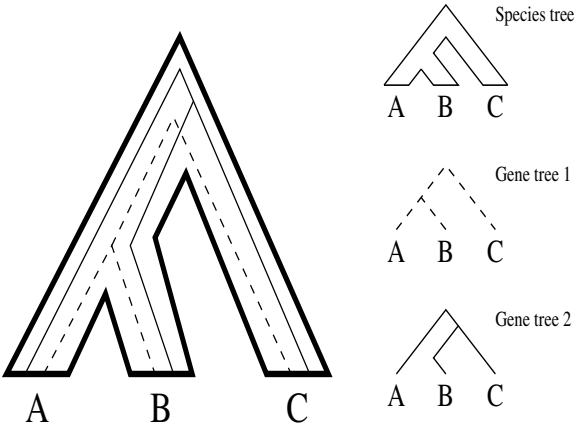


FIGURE 9.13. Incomplete lineage sorting.

How to address this last issue remains unresolved. The most thoroughly studied solution thus far involves incomplete lineage sorting. When multiple alleles of a gene or loci exist in a population, genetic drift and rapid multiple speciation events (Rannala and Yang, 2008) can result in gene histories differing from species histories (Rannala and Yang, 2003). Figure 9.13 displays this process whereby genetic drift causes gene tree 2 to conflict with the species tree. Fortunately, coalescent theory (Kingman, 1982) allows us to model incomplete lineage sorting in a single unified framework. The framework assumes an underlying species history τ_{species} that describes the history of the species under study. In Figure 9.13, τ_{species} corresponds to the large outer tree. The coalescent framework then embeds the drift of each τ_m inside the species tree $p(\tau_m \mid \tau_{\text{species}}, \theta)$, where θ characterizes the drift process. Rannala and Yang (2003) provide a formal derivation of $p(\tau_m \mid \tau_{\text{species}}, \theta)$ in a phylogenetic context. In the final step, the coalescent framework assumes conditional independence across the loci and hierarchically models the relationships amongst the M loci through

$$\prod_{m=1}^{M} p(\tau_m \mid \tau_{\text{species}}\theta) \times p(\tau_{\text{species}}) \times p(\theta),$$

where $p(\tau_{\text{species}})$ and $p(\theta)$ are prior distributions. Edwards, Liu, and Pearl (2007) parameterize $p(\tau_{\text{species}})$ according to a Yule process.

The idea of using a species tree τ_{species} and coalescent theory to model the variation amongst the M loci works extremely well for incomplete lineage sorting (Edwards, Liu, and Pearl, 2007). This framework, however, does have limitations. Most

glaring, this framework does not appropriately model horizontal gene transfer, recombination, reassortment, secondary contact, gene flow, hybridization, introgression, or other such processes. These later processes tend to dominate across many scales of evolution and cannot be ignored (Koonin, 2009).

Statistical models that formally account for these non-vertical processes remain in their infancy. Recently, Bloomquist, Dorman, and Suchard (2009) took initial steps to correct these deficiencies when dealing with data under the influence of recombination. Recognizing that a recombination event requires two parental sequences, Bloomquist, Dorman, and Suchard (2009) suggest that once a recombination event is inferred, an inferred recombinant sequence should be split into its two inferred parentals, effectively increasing the size of the dataset by one. As a benefit to this augmentation strategy, we can infer a single tree from the data and obtain information on the dates of recombination events. Unfortunately, data augmentation requires the investigator to make a distinction between a parental sequence and recombinant sequence before conducting an analysis.

Ancestral recombination graphs (ARG) and phylogenetic networks provide another solution (Hudson, 1983; Huson and Bryant, 2006). ARGs are graph theoretic structures $\mathscr{G}$ that provide a complete framework to model both vertical evolution and non-vertical evolution (Bloomquist and Suchard, 2009). Intuitively, ARGs act in a hierarchical fashion by drawing inference

$$\prod_{m=1}^{M} p(\tau_m \,|\, \mathscr{G}) \times p(\mathscr{G} \,|\, \xi), \tag{9.3.5}$$

across the M trees τ_m through $p(\tau_m \,|\, \mathscr{G})$. The coalescent with recombination is the most commonly used distribution for $p(\mathscr{G} \,|\, \xi)$ (Hudson, 1983), although other such distributions exist (Bloomquist and Suchard, 2009). As a case study, we focus on an example of recombination between bacterial symbionts living in communion with deep sea clams (Stewart, Young, and Cavanaugh, 2009). In the original analysis, Stewart, Young, and Cavanaugh (2009) adopt a supertree approach, infer trees τ_m on each individual locus m, and then analyze the discordance patterns between the trees τ_m to confirm recombination. Instead of this multi-stage approach, we combine a novel rate-variation model with the ARG model of Bloomquist and Suchard (2009) to infer the presence of recombination. This rate-variation model allows for more flexibility in τ_m than the model presented in Bloomquist and Suchard (2009). We use a MCMC sampler developed in the phylogenetic software package BEAST to make inference, and display the most probable ARG representing the history of these symbionts in Figure 9.14. In Figure 9.14, the data give a 16% posterior probability for this particular history, double the probability of any other scenario. In particular, the figure displays two recombination events. In the event closer to the root, the gene *acrB* has a distinct history from the four other genes. In the event closer to the current time, the gene *COI* has a distinct from the other four. As shown, the data support two recombination events in the evolutionary history of these organisms, which agrees with the results in Stewart, Young, and Cavanaugh (2009). Nev-

ertheless, this joint analysis based upon ARGs provides a more statistically sound approach for inference.

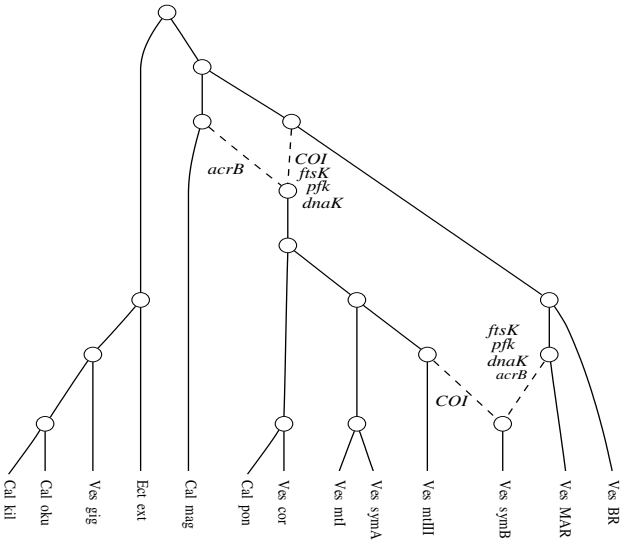


FIGURE 9.14. Ancestral recombination graph representing the most probable history.

9.3.3 Looking Ahead

Broadly speaking, Bayesian methods for molecular phylogenetics are in their third wave of growth. The first wave provided the initial spark and fostered a recognition of the value of the Bayesian framework (Holder and Lewis, 2003). The second wave furnished far more sophisticated models, especially the substitution model D (Suchard, Weiss, and Sinsheimer, 2001). Finally, the third and current wave draws upon the full power of the Bayesian paradigm to formally and rigorously integrate molecular evolution with other types of data. As examples, the Bayesian paradigm allows for the integration of temporal rate variation, the integration of indel processes (Lunter et al., 2005; Redelings and Suchard, 2005). and the integration of incomplete lineage sorting and non-vertical processes for multilocus datasets (Edwards, Liu, and Pearl, 2007).

This move towards joint models is likely to continue. Two broad areas of current interest include the integration of continuous traits such as gene expression and climatic data into the molecular evolution (Guo et al., 2007), and the joint modeling of geography and molecular evolution. In addition, investigators have begun to explore the relationships between molecular evolution, clinical medicine, and public health

(Nesse and Stearns, 2008). For example, Kitchen et al. (2009) adopt a hierarchical Bayesian framework to model the relationship between HIV evolution and antiretroviral therapy. On a final note, the deluge of data from next-generation sequencing technology opens up vast new areas of inquiry.

Acknowledgments: This work was supported by the National Institutes of Health [GM086887, AI07370] and the John Simon Guggenheim Memorial Foundation.

Chapter 10
Bayesian Data Mining and Machine Learning

Researchers in machine learning have developed methods for largely automated inference with large data sets. With increasingly more powerful computing resources and ever increasing needs for statistical inference for massive data sets, similar methods are also being developed by researchers in Bayesian analysis. The distinction between machine learning and Bayesian analysis is starting to blur. This chapter discusses several examples of such research.

10.1 Bayesian Model-based Principal Component Analysis

Bani K. Mallick, Shubhankar Ray, and Soma Dhavala

Principal component analysis (PCA) for dimension reduction finds use in many applications ranging from biology to image processing, wherever redundant and high-dimensional data are encountered. Recently, Tipping and Bishop (1999a) in the so-called "Probabilistic PCA" have taken a classical linear model approach to PCA. Minka (2000) later extends this into a Bayesian paradigm which enables automatic selection of the dimensionality from PCA and precludes the use of previously used heuristics. However, this model as briefed in Section 10.1.1, relies heavily on Laplace approximation which has questionable performance for highly parameterized models. In contrast, we perform *exact* inference to justify our estimates and the derived models follow either in conjunction or subsequently from this initial *Random Principal Components* (RPC) model. West (2003) developed PCA based regression but their the main intention was prediction and the dimension of the principal components was fixed. The assumptions set forth by the PCA model remain largely unsatisfied in practice and dimension reduction by PCA is rendered suboptimal. Highly non-linear structures are not uncommon and even though not of immediate use, are of interest in the study of manifolds. Hastie and Stuetzle (1989) account for non-linear structure through principal curves, which although of theo-

M.-H. Chen et al. (eds.), *Frontiers of Statistical Decision Making and Bayesian Analysis*, DOI 10.1007/978-1-4419-6944-6_10,

retical interest lack interpretation and scalability in higher dimensions. In contrast, our approach is conceptually closer to Tibshirani's (1992) work which tries to approximate the principal curve by a mixture smaller principal curvelets. We employ several RPC models distributed over a partitioned space to create a piecewise linear RPC model. As shown in Section 10.1.2, the common familiarity with Bayesian linear model theory makes this approach handy and inviting for more evolved models. For example, we allow the significant dimensions to vary between subspaces for better conformity with the local variations or inflections in the sample space. Recently, mixture modeling approaches have been used to develop flexible model based PCA (Hinton, Dayan, and Revow, 1997; Tipping and Bishop, 1999b). The locally adaptive partition model presented here is different than the mixture formulation and it is equivalent to local modeling of the PCA where the local parameters are estimated based on the partition or neighborhood.

PCA forms another common mode of exploratory analysis where a dimension reduction step precedes the clustering procedure. The idea of using principal components for clustering is appealing, in that the clustering algorithm deals only with the relevant information available in the data and remains largely unaffected by the unimportant variations. Several well-established clustering strategies have been employed namely, hierarchical clustering (Eisen et al., 1998), and self-organizing maps (SOMs) (Tamayo et al., 1999). All these methods, although purposeful, work with the knowledge of the number of clusters to identify. In practice however, such knowledge is unavailable and it is desirable to have a clustering technique which would additionally estimate and adopt to the number of clusters (to the limits, that the data would support). In a Bayesian context, clusters have been modeled using a multivariate normal mixture and the BIC approximation (Fraley and Raftery, 2002) is sought to simplify the posterior rule towards the estimation process.

In Section 10.1.3, we extend the RPC model to a general Bayesian framework for clustering analysis by combining the two-fold problem of dimension reduction and clustering rather than performing them separately. The novelty of the proposed Bayesian model is it draws information from both the clustering and the PCA model. In most of the existing models, a naive approach is used, where the estimates from the PCA model are simply plugged into the clustering model. Thus the PCA model is unaware of additional information in the clustering process as it completely overlooks the clustering problem. Clustering is accomplished through Bayesian partition models, whereby we assume the existence; and then look for separable partitions in the sample space each supporting a multivariate normal distribution. Throughout this paper, we emphasize on exact Bayesian posterior inference in estimation, and avoid reliance on BIC or Laplace approximation methods (Kass and Raftery, 1995) which have questionable performance for high-dimensional models.

The details of the reversible jump-based MCMC used to implement all the presented models are given in Section 10.1.4. Finally the experimental results for real and synthetic data sets are presented in Section 10.1.5.

10.1.1 Random Principal Components

In their recent study of PCA, Tipping and Bishop (1999a) produce classical estimates of various parameters accompanying the eigendecomposition. The model in its general form which can be variously approached for improvements and to mould PCA into other well-known statistical models. For example, Minka (2000) automates the process of selecting the effective number of principal components using a Bayesian methodology, an aspect that has been addressed in the past by use of various rules of thumb (Jolliffe, 1986).

Our aim is to select the PCs that effectively capture the significant attributes in data. Let $\mathscr{S}_d \subset \mathbb{R}^d$ be the sample space of observations $Y = (y_1, y_2, \ldots, y_n)$ and $y_i \in \mathscr{S}_d$. The dimension reduction in PCA is expressed in a linear model as

$$Y = \mu 1_{1\times n} + X\beta + \varepsilon,$$

where the overall mean μ is assumed to be distributed uniformly on $\mathbb{R}^d$, the unknown transformation X projects $\mathscr{S}_d$ onto a smaller dimensional subspace of principal components $\mathscr{S}_k \subset \mathbb{R}^k$ ($k \ll d$), the coefficient β, denotes the collection Y on projection, i.e., $\beta = (\beta_1, \beta_2, \ldots, \beta_n) \in \mathscr{S}_k$ and the error vector ε follows $\mathrm{N}(0, \sigma^2 I_d \otimes I_n)$.

The likelihood function is given by

$$f(Y|\mu, X, \sigma^2, k) \propto |U|^{-n/2} \exp\Big(-\frac{n}{2}\mathrm{tr}U^{-1}\Sigma\Big),$$

where $U = XX' + \sigma^2 I_d$ and Σ is the sample covariance. Tipping and Bishop (1999a) derive the maximum likelihood estimates of X in the form $\hat{X} = \Gamma(\Lambda - \sigma^2 I_k)^{1/2}\Theta$, where the pair (Γ, Λ) comprises the eigendecomposition of Σ after retaining k significant dimensions and Θ is an arbitrary rotation. Thereafter, the residual variance σ^2 is estimated as the sample mean of the remaining $d-k$ observed eigenvalues.

The model is highly parameterized and any attempts to fully describe the rich parameter class through prior distributions is computationally prohibitive. Conjugate prior distributions are convenient and can be motivated from the likelihood function based on the above mentioned estimates

$$\begin{aligned}
&f(Y|X, \sigma^2, k) \\
&\propto |XX' + \sigma^2 I_d|^{-(n-1)/2} \exp\Big[-\frac{1}{2}\mathrm{tr}\{(XX' + \sigma^2 I_d)^{-1}\Sigma\}\Big] \\
&= \Big\{\prod_{i=1}^{k} \lambda_i \sigma^{2(d-k)}\Big\}^{-(n-1)/2} \exp\Big\{-\frac{n\mathrm{tr}(\Sigma - \Gamma'\Sigma\Gamma)}{2\sigma^2} - \frac{n\mathrm{tr}(\Lambda^{-1}\Gamma'\Sigma\Gamma)}{2}\Big\}, \quad (10.1.1)
\end{aligned}$$

where the overall mean μ has been averaged out with a uniform prior over $\mathbb{R}^d$; and λ_i's are the eigenvalues in the model. A balance between the significant eigenvalues and the residual variance is apparent in (10.1.1). In the next subsection, we propose priors to explore the full uncertainty distribution of k. This full uncertainty distribu-

tion is also relevant to later developments in Sections 10.1.2 and 10.1.3, when we introduce localized RPC models.

10.1.1.1 Priors

The posterior estimation of the transformation X rests on the knowledge of the dimensionality k of the problem. Following Minka (2000), we set three distinct priors for the parameters $(X, \sigma^2) \equiv (\Gamma, \Lambda, \Theta, \sigma^2)$ with the aim of facilitating the estimation of k.

For any reduced dimension k ($< d$), the eigenvectors Γ form k frames in $\mathbb{R}^d$. It is instructive to express this jointly with the arbitrary rotation as a pair (Γ, Θ) to form k frames uniformly distributed in $\mathbb{R}^d$. We write $S(d-1,k)$ for the manifold spanned by the frame and $f(\Gamma,\Theta) \equiv f(\Gamma) = \mathrm{U}(S(d-1,k))$ for the uniform distribution on $S(d-1,k)$. Even though more accurate distributions have been derived for the k-frames (for example, see Bingham, 1974), they tend to burden posterior inference without contributing significantly to the model and we stay in our present course for ease of treatment and implementation.

By standard practice of prior specification in the linear model, the eigenvalues Λ are assumed to be i.i.d. inverse-gamma distributed for conjugacy,

$$f(\Lambda) \propto |\Lambda|^{-(\alpha+2)/2} \exp\left(-\frac{\alpha}{2}\mathrm{tr}\Lambda^{-1}\right) \equiv \mathrm{IG}(\alpha,\alpha), \tag{10.1.2}$$

where α is a shape hyperparameter. The residual variance σ^2 is elicited similarly as $f(\sigma^2) \sim \mathrm{IG}(\alpha(d-k),(\alpha+2)(d-k)/2-2)$ with hyperparameters tuned so that the joint prior $f(\Lambda,\sigma^2)$ is conjugate to the likelihood (10.1.1).

The dimension k is proposed to lie in the interval $[1,d]$ as $f(k) = \mathrm{Poisson}(\eta d)1_{[1,d]}$ for some $0 < \eta \leq 1$. For tighter estimates, one might restrict η to smaller values, while more conservative estimates would require larger values of η that are close to 1. The estimation can benefit from these adjustments when there is insufficient information in the data.

Some assumptions made here are not altogether realistic, for example, the eigenvectors and eigenvalues are specified independently. In practice however, dealing with dependent or correlated parameters in a high-dimensional model such as ours, can be computationally restrictive and it is compelling to have relaxed conditions for inferential ease.

10.1.1.2 Posterior Inference

The posterior distribution of the parameters conditional on the dimension k can be obtained by combining these priors with (10.1.1),

$$
\begin{aligned}
& f(\Gamma, \Lambda, \sigma^2 | Y, k) \\
& \propto |U|^{-m/2} \exp\Big[-\frac{1}{2} \mathrm{tr}\{U^{-1}(\Sigma + \alpha I)\}\Big] \\
& = \Big\{ \prod_i \lambda_i \sigma^{2(d-k)} \Big\}^{-m/2} \exp\Big\{ -\frac{n\mathrm{tr}(\Sigma - \Sigma^*)}{2\sigma^2} - \frac{\mathrm{tr}\Lambda^{-1}(n\Sigma^* + \alpha I)}{2} \Big\}, \qquad (10.1.3)
\end{aligned}
$$

where $m = (n + 1 + \alpha)/2$ and $\Sigma^* = \Gamma' \Sigma \Gamma$. Our interest lies in the posterior estimates of the parameters associated with the transformation (the number of principal components retained and the corresponding transformation). First, we estimate the significant number of principal components (PCs) k and conditional on it, the tuple $(\Gamma, \Lambda, \sigma^2)$ is estimated. Calculations for marginalized inference can be tedious and result in unstable forms that are difficult to contain in MCMC. On the other hand, inference from the full posterior is quite tractable contrary to its involved look. Alternately, the parameters $(\Gamma, \Lambda, \sigma^2)$ can be averaged out using Laplace approximation (Minka, 2000) which is known to introduce bias for highly parameterized models and leads to questionable estimates of k. We rely on the reversible jump sampler (RJS) (Green, 1995) in MCMC methods to work with the full posterior distribution (10.1.3) and to explore the parameter space of variable dimensions.

While the RPC model seems to fit quite well in fulfilling the objectives of traditional PCA, the linearity assumptions may not be satisfied by the data. Rather, in various applications of model-based PCA such as in machine learning, it is really a chance that the data would embody such assumptions and transformations are typically used to sufficiently attain them. Transformation to normality becomes increasingly difficult with increasing dimensions, where the data are sparse and tend to group in clusters. In such situations, it might help to treat the partitioned sample space with the hope that each partition would better match the model conditions. We take up this extension of the RPC model in the next section.

10.1.2 Piecewise RPC Models

The model heretofore (Tipping and Bishop, 1999b; Minka, 2000) attempts to describe the d-dimensional space by a small number k ($< d$) of basis vectors. The method performs optimally when the assumptions of the linear model framework are sufficiently satisfied. Otherwise, more evolved models that can capture the nonlinear structure will tend to be more effective. Many nonparametric approaches for classification exist such as learning vector quantization (Kohonen, 1988), clustering techniques (Duda, Hart, and Stork, 2000) etc.

A Bayesian scheme however has clear advantages over other methods, because it becomes possible to easily incorporate and enrich it by extra information (or specifications) into the model through priors. We employ Bayesian partitioning schemes to direct the tessellation of the sample space which is represented as a collection of ℓ disjoint (linear) subspaces $\mathscr{S}_{d(i)}$, each equipped with its own RPC model and containing n_i samples ($\sum n_i = n$).

The RPC model in Section 10.1.1, is highly parameterized for its descriptive nature and for the use of full posterior distribution for exact Bayesian inference. We look for representations of the tessellated sample space which do not heavily weigh upon the already involved model. The Voronoi tessellation (Aurenhammer, 1991) comprises such a frugal representation of the tessellated geometry and forms partitions which directly relate to the decision regions between normal clusters in decision theory. These properties make Voronoi tessellation attractive for a partitioning scheme that is delineating distinct regions (in the sense presumed by the model).

Let $G = (g_1, g_2, ..., g_\ell)$ be the collection of generators corresponding to the ℓ Voronoi partitions or cells, such that all enclosed sample points in a partition are closer to its generator than any other. What remains is to propose these generating points from a suitable statistical model such that the resulting Voronoi cells (V-cells) form envelopes enclosing homogeneous groupings in the sample space.

Considering a tessellation with ℓ partitions, we furnish each cell with a separate RPC model with individual number of PCs and the corresponding eigentransformation,

$$f(Y|g_i, X_i, \mu_i, \sigma_i^2, k_i) = \mathrm{N}(\mu_i, X_i X_i^T + \sigma_i^2 I_k),\ i = 1 \ldots \ell,$$

where X_i is a $d \times k_i$ transformation matrix for the region $\mathscr{S}_{d(i)}$ and σ_i^2 is the noise variance in each cell. The priors for each V-cell are independently specified as in Section 10.1.1:

- $f(\mu_i|g_i, \sigma_i^2, k_i) = \mathrm{N}(0, c_i\sigma_i^2)$, where c_i is a suitable scaling constant.
- $f(\sigma_i^2|g_i, k_i) = \mathrm{IG}(\alpha_1(d-k_i), (\alpha_2+2)(d-k_i)-2)$.
- $f(\Lambda_i|g_i, k_i) = \prod_{j=1}^{k_i} f(\lambda_{j(i)}) = \prod_{j=1}^{k_i} \mathrm{IG}(\alpha_1, \alpha_2)$, where α_1, α_2 are hyperparameters.
- $f((\Gamma, \Theta)_i|g_i, k_i) \propto 1/\text{area of the manifold spanned by } k_i \text{ orthoframes}$.
- $f(k_i|g_i) = \mathrm{Poisson}(\eta d)1_{[1,d]}$.

Without affecting the posterior inference significantly, it may be possible to directly plug some of the local parameters, for example, the model may be conditioned on local sample means, $\hat{\mu}_i = \sum_i y_i I(y_i \in \mathscr{S}_{d(i)})/n_i$ and sample variances, $\hat{\sigma}_i^2 = \sum_i (y_i - \hat{\mu}_i)^2 I(y_i \in \mathscr{S}_{d(i)})/n_i$ in $\mathscr{S}_{d(i)}$. Otherwise, the complete posterior distribution is a combined product of ℓ individual posterior distributions of the form (10.1.3).

The number and orientation of partitions or equivalently the generators G are randomly proposed by an external reversible jump sampler in search for the ideal partitioning supported by the complete posterior distribution and the non-linear structure in data. This requires additional prior elicitation for the number of partitions ℓ and the intra-partition dimensions k_i:

(a) **Tessellation priors** specify the size ℓ and the corresponding generators G. Following Green (1995), we can setup reversible MCMC moves between models of varying size. A combination of birth, death and update moves selected in a systematic or random manner can be used to find the optimal tiling. The size ℓ is specified by a Poisson prior distribution, truncated to some suitable size $\ell_{\max}(\ll n)$. Tessellations are assigned noninformative priors of the form $f(G) \propto 1/\ell$, to penalize for a large number of partitions.

(b) **Dimension priors** look at the compressibility of each subspace, that is the k_i's for the tessellation. It is convenient to use a Poisson prior (truncated at d) for the subspace dimensions k_i's.

Previous work on nonlinear PCA, notable among which are by Hastie and Stuetzle (1989) and Tibshirani (1992). have some serious limitations. The Principal curves developed by Hastie and Stuetzle although interesting may lack in its description of the nonlinear structure, as the concept is not extendible to multiple descriptive curves.

Later Tibshirani (1992) takes a different approach, where a mixture of smaller principal curvelets approximates the principal curves. This idea is related to the concept of approximation by functional bases which are well understood and are of interest due to their smoothing properties. While functional bases such as splines, wavelets, etc, are popular tools for approximation in small dimensions, they have limitations in higher dimensions.

Some improvement in the rates of approximation can be achieved by allowing dictionaries of functional bases. In this spirit, the above discussed mixing of sampled RPC models provides better approximation as the final approximation evolves as an average of various RPC models. Suppose $(G^{(1)}, G^{(2)}, \ldots, G^{(m)}, \ldots)$ is a collection of possible tessellations, where the m^{th} tessellation has a size $\ell^{(m)}$ and the polygons comprising it holds the parameters $\Pi^{(m)}$. The mixing provides estimates

$$\mathrm{E}(\Pi, G, \ell | Y) = \sum_{m>0} (\Pi, G, \ell) f(\Pi^{(m)} | Y, G^{(m)}, \ell^{(m)}) f(G^{(m)}, \ell^{(m)} | Y).$$

Gibbs sampling is used to draw the intra-partition parameters, $\{\Pi_i\}_{i=1}^{\ell}$, where $\Pi_i = (\mu_i, \sigma_i^2, \Lambda_i, \Gamma_i, k_i)$, and the tessellation, G alternately. For the m^{th} sampled tessellation $G^{(m)}$, the parameters in j^{th} partition $\Pi_j{}^{(m)}$, are sampled conditional on the remaining partitions, $\{\Pi_i^{(m)}\}_{-j}$. Later the generators $g_i^{(m+1)}$ defining the tessellation $G^{(m+1)}$ are sampled conditional on $\{\Pi_i{}^{(m)}\}_{i=1}^{\ell}$, one at a time by a reversible jump sampler (RJS) to admit changes in ℓ. Once the partition size $\hat{\ell}$ has been established from the sampled sizes $(\ell^{(1)}, \ell^{(2)}, \ldots, \ell^{(m)}, \ldots)$, another round of Gibbs sampling gives estimates for the tessellation $\hat{G} = \{\hat{g}_i\}_{i=1}^{\hat{\ell}}$ and the intra-partition parameters, which include the PCs that are oriented to capture the trend within each partition.

Sample statistics that assess the stationarity of the Markov chain can be useful in determining the number of MCMC samples. The estimates (of the tessellation size $\hat{\ell}$ or that of the tessellation $\hat{G}$) themselves can proceed from different statistics of the MCMC samples, for example, when interpretability is desired, the modal values come in handy. Details of the RJS based MCMC are left for Section 10.1.4, while we discuss another extension of the RPC model with applications in clustering high-dimensional data.

10.1.3 Principal Components Clustering

Clustering can be loosely defined as the process of grouping similar objects in the sample space. In the clustering literature, PCA is applied to reduce the dimensionality of the data set *a priori* and to extract the structure relevant to cluster separation. Partition models, used so far for isolation of linear subsets in data, can be modified to delineate regions in the space $\mathscr{S}_k$ of coefficients β. This projected space $\mathscr{S}_k$ is tessellated so that every coefficient in $\beta = (\beta_1, \beta_2, \ldots, \beta_n)'$ falls in exactly one partition.

In general, the tessellations for clustering in $\mathscr{S}_k$ would vary with k. A natural strategy is to search among tessellations in the whole range of spaces $\{\mathscr{S}_1, \ldots, \mathscr{S}_d\}$ that originate from our initial RPC model

$$Y = \mu 1_{1\times n} + X\beta + \varepsilon \tag{10.1.4}$$

for different values of k. The two-fold problem of dimension reduction and partitioning are closely connected, and it is tempting to unify the RPC and partition models. A unified model is rewarding as the uncertainties in two models are tied and can be mutually informative in the search for interesting clusters in $\mathscr{S}_k$('s). Additionally, as shown below, the data Y in this model-based approach, directly affects the partitioning algorithm through posterior quantities.

We allow the partition model to determine the relevant clusters in $\mathscr{S}_k$ provided by the anteceding RPC model in a hierarchical framework. Sticking with our notation for the piecewise RPC model, we define a partition as a collection of ℓ-disjoint subsets $\mathscr{S}_{k(j)} \subset \mathscr{S}_k$, $j = 1 \ldots \ell$. The cluster or data within each partition is assumed to follow a multivariate normal distribution and this is expressed collectively for all the partitions by writing a sub-linear model

$$\beta = \gamma Z + \zeta, \tag{10.1.5}$$

where $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_\ell)'$ is a collection of intra-partition means and Z links the response β_i in a particular partition to its mean, for example, if the i^{th} row of Z contains a 1 at the j^{th} column, it indicates $\beta_i \in \mathscr{S}_{k(j)}$. Thus the pair (γ, Z) fully describes the tessellation in the PCC model. The uncertainty in this sub-model is expressed by the error vector ζ containing elements $\zeta_i \sim \mathrm{N}(0, p_{ji}\sigma^2 I_k)$, where $p_{ji} = \sum p_j 1_{\beta_i \in \mathscr{S}_{k(j)}}$ determines the variance of β_i lying in $\mathscr{S}_{k(j)}$. In short, we write $\zeta \sim \mathrm{N}(0, \sigma^2 V)$, where $V = \mathrm{diag}(p_{j1} I_k, \ldots, p_{j\ell} I_k)$ and the elements of β lying in the j^{th} partition can be elicited by a prior

$$f(\beta_i | \gamma_j, \{p_j\}, \sigma^2) = \mathrm{N}(\gamma_j, p_{ji}\sigma^2).$$

The exposition benefits from a simplified model that follows by plugging (10.1.5) into (10.1.4) as

$$Y = \mu 1_{1\times n} + X\gamma Z + X\zeta + \varepsilon, \tag{10.1.6}$$

which produces the marginal-likelihood $f(Y|\mu,X,\sigma^2,\gamma,V,k) = \mathrm{N}(\mu 1_{1\times n} + X\gamma Z, U)$ where $\sigma^{-2}U = I_d \otimes I_n + X \otimes V \otimes X'$. A step further in the model hierarchy assumes a scaled prior for $\gamma \sim \mathrm{N}(0, qV\sigma^2)$, where q may be selected for the prior to be relatively diffuse on $\mathbb{R}^k$. The new marginal takes the usual form

$$f(Y|\mu,X,\sigma^2,k,\gamma,Z) \propto |U^*|^{-1/2} \exp\Big(-\frac{n}{2}\mathrm{tr}U^{*-1}\Sigma\Big),$$

where $U^* = U + q\sigma^2 X \otimes V \otimes X' = \sigma^2(I_d \otimes I_n + X \otimes V \otimes X' + X \otimes qV \otimes X')$.

To perform full Bayesian analysis, we again resort to the simplifications of Tipping and Bishop model. This is made possible by making the equal intra-partition variance assumption (i.e., $q_i = q_j$, $\forall i,j$), followed by a reparameterization that allows us to write the contributions of the transformation X in the model covariance U^* as

$$X \otimes (V + qV) \otimes X' \equiv \tilde{X}\tilde{X}'.$$

The estimates of $\tilde{X}$ arising from this simplification matches the estimates $\hat{X}$ discussed in Section 10.1.1. In other words, the modelling precision lost by making the equal variance assumption, is compensated amply by the sizeable reduction in model complexity. This new model relates to the equal spherical variance (EI) model of Banfield and Raftery (1993) who have shown it to perform better than heteroscedastic models in their BIC approximation clustering paradigm.

Once the marginalized likelihood has been rewritten by conditioning on the tessellation, the posterior inference follows directly from Section 10.1.1 with a few additions at a hierarchical level. We write

$$\begin{aligned} & f(Y|\tilde{X},\sigma^2,k,\gamma,Z) \propto |U^*|^{-1/2} \exp\Big(-\frac{n}{2}\mathrm{tr}U^{*-1}\Sigma\Big) \\ &= |\tilde{X}\tilde{X}' + \sigma^2 I_d|^{-n/2} \exp\Big[-\frac{1}{2}\mathrm{tr}\Big\{(\tilde{X}\tilde{X}' + \sigma^2 I_d)^{-1}\Sigma\Big\}\Big] \\ &= \Big\{\prod_i \lambda_i \sigma^{2(d-k)}\Big\}^{-n/2} \exp\Big\{-\frac{n\mathrm{tr}(\Sigma - \Sigma^*)}{2\sigma^2} - \frac{n\mathrm{tr}(\Lambda^{-1}\Sigma^*)}{2}\Big\}, \end{aligned} \tag{10.1.7}$$

where $U^* = \sigma^2 I_d \otimes I_n + \tilde{X}\tilde{X}'$, Σ retains the original notation for sample covariance and so does Σ^* which equals $\Gamma'\Sigma\Gamma$.

A model-based approach allows simultaneous evaluation of many underlying cluster models in the data at the cost of some extra-parameterization. In the preceding section, we were successful in expressing the variability jointly by writing the combined model (10.1.6) and consequently, expediting the estimation process of the model defining parameters. In the following discussion, we look at the estimation of the size ℓ, the partitions means $(\gamma_1, \gamma_2, \ldots, \gamma_\ell)$ for the partitions separating the unknown principal components β.

Each partition supports the model $\beta_i = \gamma_j + \varepsilon_i, \forall\, \beta_i \in \mathscr{S}_{k(j)}$, where γ_j is the model mean associated with the j^{th} cluster and $\varepsilon_i \sim \mathrm{N}(0, \sigma^2 p_{ji} I_k)$. The joint likelihood of all the β_i's lying in $\mathscr{S}_{k(j)}$ is

$$\prod_{\beta_i \in \mathscr{S}_{k(j)}} f(\beta_j|\gamma_j, p_{ji}, \sigma^2) = \prod_{\beta_i \in \mathscr{S}_{k(j)}} \mathrm{N}(\gamma_j, p_{ji}\sigma^2).$$

The joint prior for γ_j is conveniently expressed by a scaled conjugate normal prior, $f(\gamma_j|\sigma_j^2) = \mathrm{N}_{n_j}(0, qp_{ji}\sigma^2 I_k)$, which induces the posterior form

$$f(\gamma_j|p_{ji}, \sigma^2, \beta) = \mathrm{N}(n_j^{*-1}\sum_i \beta_i, n_j^{*-1}\sigma^2),$$

where $n_j^* = n_j + (qp_{ji})^{-1}$ and n_j is the number of β_i falling in $\mathscr{S}_{k(j)}$. We assume the knowledge of within-partition variance $p_j\sigma^2$ (see (10.1.5) for notation), in that we plug in the estimated variance within each partition in the model, without severely affecting the full Bayesian flavor of our model. The partition size is again elicited by $f(\ell) \propto \exp(-\upsilon\ell)$, where υ is a penalizing hyperparameter. The conditional distribution of β given γ is obtained by combining (10.1.4) and with the sub-model (10.1.5) which is now treated as a model generating the prior distribution.

$$f(\beta|X, \sigma^2, k; \gamma, Z, V, Y) = \mathrm{N}(\beta^*, \sigma^2 V^*),$$

where $V^* = (V^{-1} + X'X)^{-1} = (V^{-1} + I_k)^{-1}$ and $\beta^* = V^*(V^{-1}\gamma Z + X'Y)$.

Starting with a randomly generated tessellation (γ, Z), the parameters $(k, \Gamma, \Lambda, \sigma^2, \beta)$ are sampled from their respective distributions by Gibbs sampling and the tessellation is later updated conditional on the new states and this is repeated alternately. Each component of the Gibbs sampler is shown below, assuming the priors listed in Section 10.1.1.1 and a random initial partition Z, randomly oriented k orthoframe and eigenvalues contained in X,

RPC Parameters
1. $\sigma^2|\Gamma, \Pi, k, Y \sim \mathrm{IG}(n\mathrm{tr}(\Sigma - \Sigma^*) + \alpha(d-k), (n+\alpha+1)(d-k)/2 - 2)$,
2. $\Lambda|\Gamma, \Pi, k, Y \sim \mathrm{IG}(n\Sigma^* + \alpha, n + \alpha + 1)$,
3. $\Gamma|\sigma^2, \Lambda, \Pi, k, Y \sim$M-H proposal of orthoframes in $\mathbb{R}^d$,
4. $\beta|X, \sigma^2, \Pi, k, Y \sim \mathrm{N}(\beta^*, \sigma^2 V^*)$,

Clustering Parameters
5. $\gamma_j|\beta, \sigma^2, \Pi_{-\gamma_j}, k \sim \mathrm{N}(n_j^{*-1}\sum_i \beta_i, n_j^{*-1}\sigma^2)$, $\forall j$,
6. $k|X, \beta, \sigma^2, \Pi, Y \sim$ M-H proposal with full conditional, and
7. $Z, \ell|X, \beta, \sigma^2, \Pi_{-\{Z,l\}}, k, Y \sim$ RJS discussed in Section 10.1.4,

where $\Pi = \{\gamma, Z, \ell, q\}$ is the collection of parameters from our partition model. Since the number of partitions ℓ is unknown, we require another RJS step (conditional on the number of PCs k) to perform a model selection in the parameter space $\cup_{k=1}^{\ell}\{k\} \times \mathscr{S}_k$. The details of RJS implementation (Steps 6 and 7) are detailed in Section 10.1.4.

10.1.4 Reversible Jump Proposals

This section is devoted to the Reversible jump based MCMC used to setup trans-dimensional moves throughout our work. In the RJS terminology, we perform dimension matching by defining three distinct Markov chain moves, namely birth, death and update with probability b_k, d_k and $1-b_k-d_k$, respectively. The birth and death moves are usually used to signify an increase or decrease in the dimension of the space, while the update step simply updates the remaining parameters on a fixed dimension. We commence by elaborating the RJS to search the parameter space $\cup_k\{k, S(d-1,k), \mathbb{R}^{k+1}\}$ for the initial RPC model. These ideas are later extended to design the RJS based moves for the extensions namely, the piecewise RPC and PCC model.

10.1.4.1 Reversible Jump for RPC Model

The parameter space to be searched $\cup_k\{k, S(d-1,k), \mathbb{R}^{k+1}\}$ encompasses all the posterior models for different dimensions k. A birth step results in a unit increment in the dimension k accompanied by an increase in the number of eigenvectors and eigenvalues and necessary adjustments to the residual variance σ^2 are made to maintain a balance. The death step works oppositely to decrease k. Finally, the eigenvalues, residual variance and the eigenvectors are updated by fixing the dimension k in the update step. While most of the parameters are directly sampled from their respective inverse gamma posteriors, the eigenvectors require a Metropolis step with proposals that isometrically and randomly rotate the k-frames in $S(d-1,k)$ (the Steifel manifold of k-frames in dimension d).

Some discussion about the cross-dimensional transitions is in order. During birth, a new eigenvector is proposed uniformly in the orthogonal complement to the span of given k-frames in $S(d-1)$, the d-dimensional hypersphere. We write $S_\perp(d-k-1)$ to denote this orthogonal complement. The next step involves the proposal of the new residual variance σ^2_{k+1}, from σ^2_k and a uniformly distributed random variable,

$$\sigma^2_{k+1} \sim \tau\sigma^2_k, \text{ where } \tau \sim \mathrm{U}(0,1). \tag{10.1.8}$$

The proposal of the new eigenvalue λ_{k+1} can be inspired from the estimates discussed in Section 10.1.1, where the residual variance is the arithmetic mean of the $d-k$ insignificant eigenvalues. This relationship however does not lead to sound proposals of λ_{k+1}, which in most cases go unaccepted in a high-dimensional search. Better chain movement is achieved by assuming a geometric relationship of the form

$$\sigma^2_k = \Big(\prod_{i=k+1}^{d}\lambda_i\Big)^{1/(d-k)} \text{ and } \sigma^2_{k+1} = \Big\{\sigma_k^{2(d-k)}/\lambda_{k+1}\Big\}^{1/(d-k-1)}.$$

Thus, once we have the proposed σ^2_{k+1} (10.1.8), the new eigenvalue λ_{k+1} is obtained as $\sigma_k^{2(d-k)}/\sigma_{k+1}^{2(d-k-1)}$. The advantage of this altered set of calculations is apparent from the acceptance probability of the new proposal—the polynomial terms in the full posterior do not dominate the dimension matching transitions. Figure 10.1 depicts a birth step for a *data cloud* in $\mathbb{R}^3$.

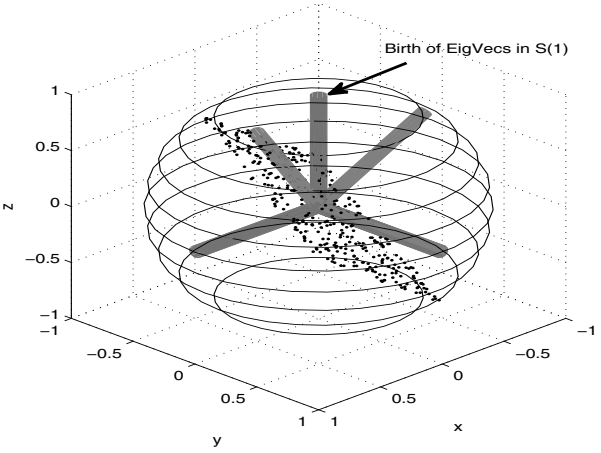


FIGURE 10.1. A birth step from $\{2, S(2,2), \mathbb{R}^3\} \rightarrow \{3, S(2,3), \mathbb{R}^4\}$.

In much the opposite fashion, the death step $(k+1 \rightarrow k)$ means that one of the $k+1$ eigenvalues and the corresponding eigenvector, is randomly dropped. Yeung et al. (2001) make the important observation that the first few PCs may not carry information about the separability of the clusters and we acknowledge this fact by not requiring the death step to remove only the last component. The new variance σ_k^2 is the product $(\sigma_{k+1}^{2(d-k-1)}\lambda_{k+1})^{1/(d-k)}$. The acceptance probability of a birth step from k to $k+1$ is $\min(R_b(k), 1)$, where

$$\begin{aligned} R_b &= (\text{posterior ratio}) \times (\text{proposalratio}) \times (\text{jacobian}) \\ &= \frac{f(\Gamma, \Lambda, \sigma^2_{k+1}, k+1|Y)}{f(\Gamma, \Lambda, \sigma^2_k, k|Y)} \frac{d_{k+1}/(k+1)}{b_k f(\sigma^2_{k+1}) \mathrm{U}(S_\perp(d-k-1))} \left| \frac{(d-k)}{(d-k-1)^2} \frac{\sigma^2_{k+1}}{\sigma^2_k \lambda_{k+1}} \right|, \end{aligned} \tag{10.1.9}$$

where $b_k = \min(1, p(k+1)/p(k))$ and $d_k = \min(1, p(k)/p(k+1))$. The acceptance probability of a death step from k to $k-1$ is $\min(R_b(k-1)^{-1}, 1)$. The steps detailed above directly provide the estimates $(\hat{\Gamma}, \hat{\Lambda}, \hat{\sigma}^2, \hat{k})$. Most importantly, the estimate $\hat{k}$, gives an idea of the number of PCs that would balance (with respect to posterior risk) the informative and unwanted sections of the data.

10.1.4.2 Extensions for the Piecewise RPC Model

In order to move between tessellations, we rely on the classical split-merge procedure to search through RJS. We again setup the birth, death, and update moves with the same interpretation as before except that in the present context it applies to the number of partitions. The partitions are assumed to be located around corresponding generators.

Since the individual generators g_j are independent by assumption, it is possible to sample them individually by a symmetric proposal distribution $q(g_j, g'_j)$, where g'_j is the newly proposed generator accepted with an acceptance probability shown in its general form below.

The split-merge technique is applied as follows. In a death step, two generators are randomly selected and merged (vectorically added), whereas in a birth step a randomly picked generator is split into two using a random variable u drawn from a suitable distribution. The birth and death steps can be summarized as follows

$$\begin{aligned} \text{Death} - (g_1, ..g_r, ..g_s, ..g_{\ell+1}) &\rightarrow (g'_1, \ldots, g'_{\ell-1}, (g_r + g_s)/2), \\ \text{Birth} - (g_1, ..g_r, ..g_\ell) &\rightarrow (g'_1 \ldots, g'_{\ell-1}, g_r - u, g_r + u). \end{aligned}$$

In general, it is less trivial to devise reversible proposals for the other parameters. For example, the mapping between the dimensionalities of the merged partition k'_ℓ and the two preexisting partitions (k_r, k_s) is not evident without some knowledge of the local trends in data. Similar mapping for other parameters which rely on k'_ℓ is thus undetermined. An alternative that is attractive from the computational side, is to propose new parameters by perturbing sampled estimates from the data within newly formed partitions. The birth acceptance ratio for these proposals is $\min\{1, R_b\}$, where

$$\begin{aligned} R_b &= (\text{posterior ratio}) \times (\text{proposal ratio}) \times (\text{Jacobian}) \\ &= \frac{\prod_{j=1}^{\ell+1} f(\Gamma'_j, \Lambda'_j, \sigma'^2_j, k'_j, G'_j | Y)}{\prod_{j=1}^{\ell} f(\Gamma_j, \Lambda_j, \sigma^2_j, k_j, G_j | Y)} \frac{d_{\ell+1}(\ell+1)}{b_\ell f(u)\ell(\ell+1)/2} \left| \frac{\partial(g'_\ell, g'_{\ell+1})}{\partial(g_r, u)} \right|. \end{aligned}$$

The Jacobian for the above mentioned split-merge procedure comes out to be 2. Some care in selecting the dispersion of u, as very small values lead to small jumps that are almost always accepted, whereas large values lead to excessively high rejection probability. All the partitions are unnecessarily involved in this sampling step, to account for the changes that might occur in the tessellation for a localized birth or death proposal. Things can be simplified by invoking properties of Voronoi tessellations, whereby a proposed generator locally affects only those partitions that encompass it in its Delaunay radius. In the update step, the parameters $(\Gamma_i, \Lambda_i, \sigma_i^2, k_i)$ for each partition updated separately following the RJS moves (10.1.9).

10.1.4.3 Reversible Jump for the PCC Model

As the reader would expect, the RJS moves for clustering the PCs are much the same as that used for piecewise RPC model, but some new constraints are observed on the Voronoi cells. We assume that the Voronoi cells are centroidal, that is the intra-partition means γ_i coincides with the generators g_i. Thus when a new partition mean γ_i is sampled, the encompassing Voronoi cell is implicitly defined to hold the data points y_i which are closest to γ_i than any other partition mean. The compound parameter space of interest now is $\cup_{k=1}^{d}\{k, S(d,k), \mathbb{R}^{k+1}, \cup_{\ell}\{\ell, \cup_{j=1}^{\ell}\mathscr{S}_{k(j)}\}\}$ unlike $\cup_{\ell}\{\ell, \cup_{j=1}^{\ell}\{\mathscr{S}_{d(j)}, \cup_{k_j=1}^{d}\{k_j, S(d,k_j), \mathbb{R}^{k_j+1}\}\}\}$ for the piecewise RPC model. The birth acceptance ratio for the RJS sampling new partitions conditional on other parameters can be written as $\min\{1, R_b\}$, where

$$\begin{aligned} R_b &= (\text{posterior ratio}) \times (\text{proposal ratio}) \times (\text{ Jacobian}) \\ &= \frac{\prod_{j=1}^{\ell+1} f(\gamma_j'|\beta, \sigma^2, \Pi_{-\gamma_j}, k) f(\ell+1)}{\prod_{j=1}^{\ell} f(\gamma_j|\beta, \sigma^2, \Pi_{-\gamma_j}, k) f(\ell)} \frac{d_{\ell+1}(\ell+1)}{b_\ell f(u)\ell(\ell+1)/2} \left| \frac{\partial(\gamma_\ell', \gamma_{\ell+1}')}{\partial(\gamma_r, u)} \right|. \end{aligned}$$

10.1.5 Experimental Results

We consider three real data sets and one synthetic data set to evaluate the three methodologies presented here in the paper. The data sets considered are:

SatImage data. The data, which is popular in the machine learning community, consists of the multi-spectral values of pixels in 3×3 neighborhoods in satellite images. This data was generated from Landsat Multi-Spectral Scanner image data. There are 6 classes in the data indicating the geography of the scene and each scene is characterized by 36 features (9 pixel intensities $\times$ 4 spectral bands). A total of 6435 observations are recorded.

Integrated spatial and spectral data for the interpreting a scene: its topography, land use, etc are important in remote sensing. It is of interest to classify distinct patterns based on a reduced set of features generated by PCA of the Satellite image data (Chang and Ghosh, 1998).

Gold Assay data. The data, which was used in their work on principal curves by Hastie and Stuetzle (1989), consists of 250 pairs of gold assays. Following their example, the data used here have been log-transformed for stabilizing the variance. The data compare results of two laboratories assaying gold content of outgoing cargo from a computer chip industry.

Hastie and Stuetzle use bootstrapping to check the adequacy of a linear model for characterizing the assay data. In a Bayesian context, we justify a nonlinear characterization with the Bayes factor which compares the posterior preference between the simple RPC model and the piecewise RPC models. These results are discussed below in Section 10.1.5.2.

Yeast cell data. The ability of the DNA microarray technology to produce expression data on a large number of genes in a parallel fashion has resulted in a wealth of data points. The main difficulty with microarray data analysis is that the sample size n is very small when compared to the dimension of the problem (the number of genes) p. The number of genes on a single array are usually in the thousands, so the number of regressors p easily exceeds the number of observations n. This is known as the "large p, small n" problem. In this situation, dimension reduction is needed to reduce the high-dimensional gene space. RPC and piecewise linear PCA approaches will be useful for this data reduction.

One commonly used approach in making conclusions from microarray data is to identify groups of genes with similar expression patterns across different experimental conditions through a cluster analysis (D'haeseleer, Liang, and Somogyi, 2000). The biologic significance of results of such analyses has been demonstrated in numerous studies (Eisen et al., 1998; Tamyo et al., 1999; Tavazoie et al., 1999). We will use hierarchical model based principal component clustering for the microarray data.

Cho et al. (1998) measure the fluctuation of roughly 6000 genes expressed over 2 cell cycles. There are 17 time points within the cycles of data which are chosen for analysis and they record the peaking of 384 genes at different periods within the cycles. The cycles were identified at 5 different phases and it is the purpose of the cluster analysis to identify these 5 cycles from the gene expression data. For the experiments, the data was standardized and pre-processed by using multivariate normality tests on the 5 cell cycles.

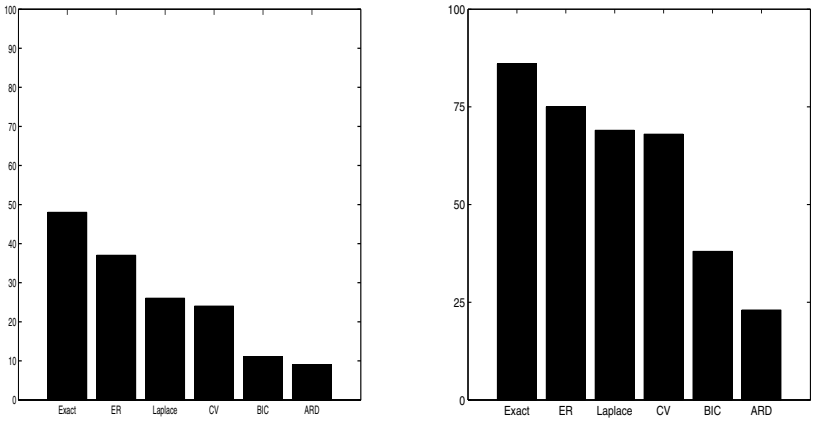
Synthetic normal mixture data. The data containing normally distributed clusters are generated for evaluation of to check the performance and model adequacy of the PCC model as real data sets can be noisy with indistinct classes. The results presented here are for a collection of 3000 observations sampled from a multivariate normal mixture model with four classes in thirty dimensions. Each class is generated by randomly sampling from a multivariate normal distribution with a near diagonal covariance matrix, which is common to all the classes.

10.1.5.1 Random Principal Components

Comparison with preexisting PCA algorithms. To test the performance relative to other PCA algorithms for model selection, we simulate data from a known model and compare the incidence of correct dimension selection by the various algorithms. The method is compared with five estimators, namely the Laplace's approximated model (Minka, 2000), the BIC approximation for the same method (Kass and Raftery, 1995; Minka, 2000), the Everson and Roberts's method (Everson and Roberts, 2000), Bishop's ARD algorithm (Bishop, 1998) and Cross-Validation.

The comparisons are based on synthetic data sets generated from known multivariate normal models exemplifying various instances of PCA analysis. We consider two models with dimensions, namely $d = 10$ and 50. For the former model, two sample sizes of $n = 20, 100$ are considered, while for the latter, we take $n = 120, 300$.

In Figure 10.2, the estimators are compared for these four combinations over 500 runs, which have been posed to evaluate the effect that various d/n ratios have on the performance. The figure plots the percentage of times the estimator hits upon the true model dimension. Our estimator performs better than the rest in all these situations, with the most notable improvement in large d or d/n valued cases.



(a) (10, 10) (b) (100,300)

FIGURE 10.2. Comparison of correct dimension selection of various estimators for various (n,d). The chances (percentage) of the estimators achieving the true model dimension is plotted for six estimators (Exact, ER, Laplace, CV, BIC, and ARD).

Experiments with the SatImage and Yeast Cell data. The MCMC for the RPC model was run for 10000 iterations for each dataset. The results in Table 10.1 give an idea of the posterior preference of the significant PCs for the two data sets measured after sufficient burn times:

TABLE 10.1. Relative frequency of PCs selected by RPC.

Dataset \ PCs	5	6	7	8	9	10	11
SatImage	0.01	0.05	0.11	0.22	0.28	0.09	0.02
Yeast Cell	0.04	0.22	0.31	0.25	0.05	0.02	0.01

For the SatImage data, the model with 9 PCs was the most favorable followed by the 8 PC model. Modest reduction (0.22-0.25) in dimensionality in the SatImage data, suggests the relative redundancy of information across the four spectral bands and locations in the image. In contrast, the compressibility $(\equiv \hat{k}/d)$ for the yeast cell data (0.41-0.47) is low, which might indicate insufficient adherence with the model

assumptions. In most cases, and for the examples considered, the data consists of several correlated measurements of an event in time or space. Low compressibility in such situations may result from the heterogeneity (or existence of clusters) in the population. It is only pertinent therefore to explore the yeast cell data in the Piecewise RPC paradigm and see if the compression helps from these refinements.

10.1.5.2 Piecewise RPC

We analyze the Gold data set, which has been typically associated with nonlinear dimension reduction models. Hastie and Stuetzle (1989) used this data to fit their principal curve, which is a nonlinear extension of the idea of principal components. A principal curve is lacking in description and the concept of secondary principal curves is seemingly vague and lacks a concrete definition. Direct extensions in this spirit are Principal surfaces (Hastie and Stuetzle, 1989) which owing to the problems in modeling remain largely unexplored.

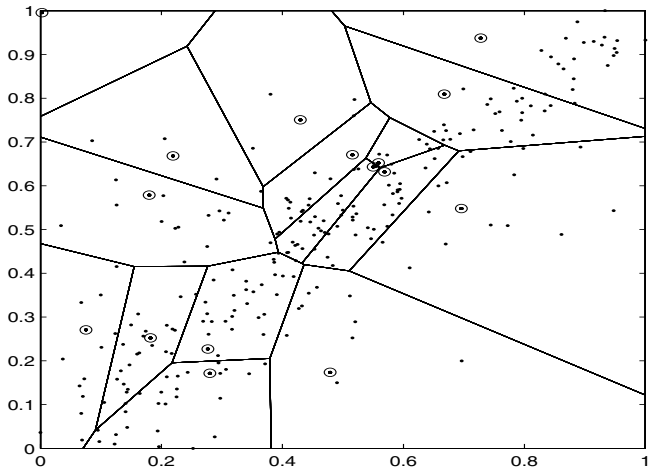


FIGURE 10.3. Nonlinear RPC for Gold data, showing 16 Voronoi polyhedra. The circles placed at the generators within each cell, are aligned to the PCs.

In this regard, the piecewise RPCs although not entirely comparable with principal curves, seem to give a piecewise approximation to principal surfaces as shown in Figure 10.3. More so, the each surface is free to have its own dimensionality. Although, not considered in the present discussion, the transitions between surfaces can also be smoothened by spline penalizers, in a way similar to Hastie and Stuetzle (1989) and Tibshirani (1992). The results are tabulated in Table 10.2.

TABLE 10.2. Piecewise RPC simulations for the Gold data.

Gold \ Size	15	16	17	18	19	20	21
Rel Freq	0.034	0.080	0.305	0.273	0.165	0.126	0.011
Bayes Factor	1.535	1.610	3.150	2.785	2.530	2.215	1.282

The piecewise RPC model is implemented for the higher dimensional yeast cell data, wherein each partition is allowed to have its own dimensionality. A straightforward comparison is possible, between the piecewise RPC model and the simple RPC model via Bayes Factors to assess the model choice. These results are presented in the results below in Table 10.3 along with the partition size and the average dimensionality (since each cluster differs in its dimensions) of the overall partition. The results strongly support the presence of heterogeneity and nonlinear structure in yeast cell responses.

TABLE 10.3. Piecewise RPC simulations for the yeast cell data.

Yeast \ Size	3	4	5	6	7	8	9
Rel Freq	0.090	0.125	0.229	0.265	0.165	0.080	0.010
Bayes Factor$\times 10^2$	1.282	1.570	4.520	6.630	2.890	2.230	1.090
Avg. k	6.750	6.230	5.663	5.275	4.772	4.100	3.607

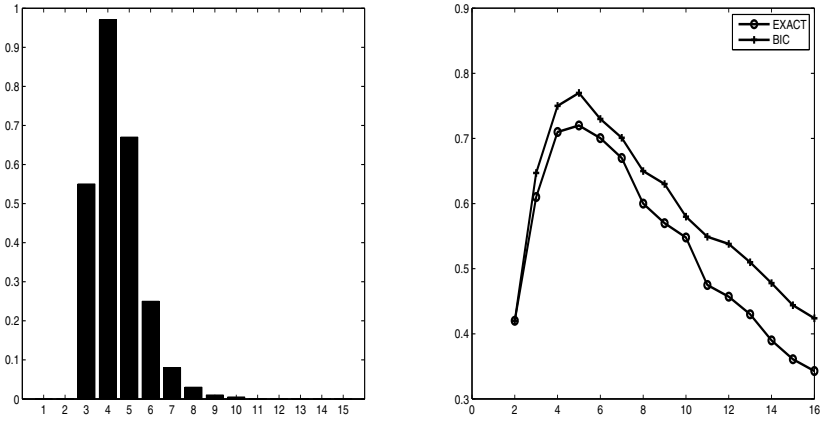
Additionally, we calculate the *average compressibility* (taken to be the average of the significant number of PCs within each partition) for each cluster configuration. Better dimension reduction is evident from these figures, which supports the stance of the preceding subsection about the improved model adequacy from partitioning.

The results from the gold data and the yeast cell data, also reflects the relative ease with which our piecewise RPC model is extended to model high-dimensional nonlinear data, a flexibility that has been missing in preexisting nonlinear PCA models.

10.1.5.3 Principal Components Clustering

We demonstrate the usefulness of exact Bayesian inference in a model-based approach to clustering. The mixture model based clustering procedure of Banfield and Raftery (1993) and Yeung et al. (2001) with BIC approximation and the heuristic CAST algorithm (Ben-Dor and Yakhini, 1999) form the benchmark for our cluster quality assessments. Liu et al. (2002) also used PCA based clustering and adopted Bayesian approach to the choice of the number of PCs.

Simulations for PCC exhibit a different aspect of principal component selection, where it is not necessary that the most significant PCs have a contribution towards classification. Indeed, RPC when combined with the clustering model resembles a variable selection procedure. The Rand index is used as a measure of agreement



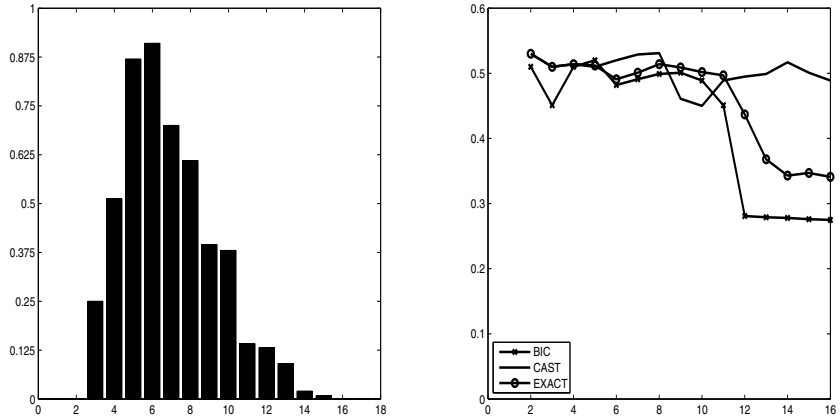
(a) (b)

FIGURE 10.4. (a) Histogram of the number of partitions in the synthetic normal mixture data and (b) Comparison of ARIs for the synthetic normal mixture data.

between a known partition to that provided by an external clustering criterion. The Rand index is a combinatorial measure of the number of pairs of objects falling in same or different groups in the two partitions and takes a value in the range [0,1]. The adjusted rand index (ARI) (Hubert and Arabie, 1985), which has been used in the simulations to evaluate the quality of clusters, additionally enforces that the expected Rand index for random partitions be 0. A high ARI indicates a fair agreement between two partitions.

Experiments were also performed for the simulated data set, in which case the initial 4-10 PCs where found useful for clustering. The relative frequencies of the number of clusters ℓ and the ARIs for the synthetic data set are shown in Figure 10.4. The PCC model uniformly dominates the BIC approximated clustering model and there is a clear preference for four clusters.

The PCC model was implemented for the Yeast cell data which showed a strong preference for the initial 5-12 PCs. The relative frequencies of the number of clusters in MCMC simulations are shown in Figure 10.5 (a). There is a strong evidence of 5 partition models in the yeast cell data, followed by the close competitors of 4 and 6 partitions. Figure 10.5 (b) presents the ARI when all the PCs from 5-12 are retained and comparisons show improvement over the BIC approximated spherical variance (EI) model used by Yeung et al. (2001).



(a) (b)

FIGURE 10.5. (a) Histogram of the number of partitions in the Yeast cell data and (b) Comparison of ARIs for the Yeast Cell data.

10.2 Priors on the Variance in Sparse Bayesian Learning: the demi-Bayesian Lasso

Suhrid Balakrishnan and David Madigan

Sparse Bayesian Learning (SBL) using automatic relevance determination as typified by the Relevance Vector machine (Tipping, 2001), has proved to be a very effective and accurate method for a wide variety of regression and classification problems. The SBL paradigm performs parameter learning via type-II maximum likelihood where a marginal data likelihood maximization provides the parameter estimates. Two related tracks, the Lasso (Tibshirani, 1996) and the Bayesian Lasso (Park and Casella, 2008), approach the estimation task in rather different ways. The Lasso considers regression and classification in the loss plus ℓ_1-regularization framework. The resulting optimization problem can also be viewed in the Bayesian setting as a maximum-*a-posteriori* (MAP) solution to a regression problem with parameters having individual Laplace (or double exponential) priors. The Bayesian Lasso instead makes use of the equivalence of a hierarchical Gaussian-Exponential prior to the Laplace prior, and conducts fully Bayesian inference (via Markov chain Monte Carlo or MCMC sampling algorithms) for parameter inference.

A number of recent papers have explored connections between these three approaches and our work is in that vein. For example Wipf and Nagarajan (2008) clearly delineate the connection between SBL's type-II maximum likelihood and MAP estimation, by showing that SBL's type-II maximum likelihood is equivalent

to MAP estimation where the prior on the parameters is "non-factorial" (in other words, the prior depends on the input basis functions, and cannot be decomposed into independent terms involving each parameter). A natural question that arises is whether type-II maximum likelihood is an effective way to train the Bayesian Lasso model as well. This would have two advantages over the Bayesian Lasso. First, parameter estimates would be sparse, and second, the parameter estimates would be obtained by optimization and not by computationally more demanding MCMC.

10.2.1 Background and Notation

We consider SBL, the Lasso and the Bayesian Lasso in the context of the classical Gaussian linear regression modeling. Specifically, given a regressor matrix/feature dictionary Φ, an observation/response vector $\mathbf{y}$ and i.i.d. Gaussian noise/errors ε, we consider linear models of the form

$$\mathbf{y} = \Phi\beta + \varepsilon. \tag{10.2.1}$$

These assumptions lead to a likelihood of the form:

$$p(\mathbf{y}|\beta, \sigma^2, \Phi) = 2\pi\sigma^{2^{-N/2}} \exp\Big\{ - \frac{\|\mathbf{y} - \Phi\beta\|}{2\sigma^2} \Big\},$$

where the dataset, $\mathscr{D}$ comprises N responses $\mathbf{y} = (y_1, \ldots, y_N)^{\mathrm{T}}$ and the N x p design matrix $\Phi = [\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_N)]^{\mathrm{T}}$. The Gaussian noise distribution is mean zero and variance σ^2, $p(\varepsilon) = \mathscr{N}(\varepsilon|0, \sigma^2 I)$ and the parameter vector β is p dimensional. We assume that the intercept parameter, if any, is estimated outside the estimation schemes discussed here (for example, by centering the response). Loosely speaking, the Lasso is the least "Bayesian" of three approaches while the Bayesian Lasso is the most Bayesian. SBL along with the "demi-Bayesian" approach we describe below are somewhere in between.

10.2.1.1 The Lasso

The Lasso formulation estimates β by solving the following convex optimization problem:

$$\min_{\beta} (\mathbf{y} - \Phi\beta)^{\mathrm{T}} (\mathbf{y} - \Phi\beta) + \rho \|\beta\|_1$$

(ρ is a non-negative scalar regularization parameter). The Lasso optimization problem has a MAP-Bayesian interpretation as follows (Tibshirani, 1996). Assign each component β_j of β an independent Laplacian or double-exponential prior distribution with mean 0:

$$p(\beta_j|\rho_j) = \frac{\rho_j}{2} \exp(-\rho_j|\beta_j|),\ \rho_j > 0, j = 1, \ldots, p$$

with $p(\beta) = \prod_j p(\beta_j)$ and all $\rho_j = \rho$. A prior of this form places high probability mass near zero and along individual component axes thereby promoting sparsity (see Figure 10.6). Figure 10.6 highlights the higher probability mass the Laplace assigns along the axes and at zero as well as its heavier tails. It also has heavier tails than a Gaussian distribution leading to some theoretical difficulties with regard to variable selection[1].

Now, in this setting, the Lasso optimization problem results in β estimates that correspond to the posterior mode estimates ($\text{argmax}_\beta\, p(\beta|\mathscr{D}, \rho)$). Predictions are then made using this point posterior mode. By contrast, fully Bayesian inference would typically integrate over the entire posterior distribution rather than conditioning on a specific value. In fact, while the posterior mode is an optimal point estimate under zero-one loss, there is no particular reason to expect such a loss function to be reasonable in any particular application. Nonetheless, the Lasso has provided excellent predictive performance in many applications (Genkin, Lewis, and Madigan, 2007).

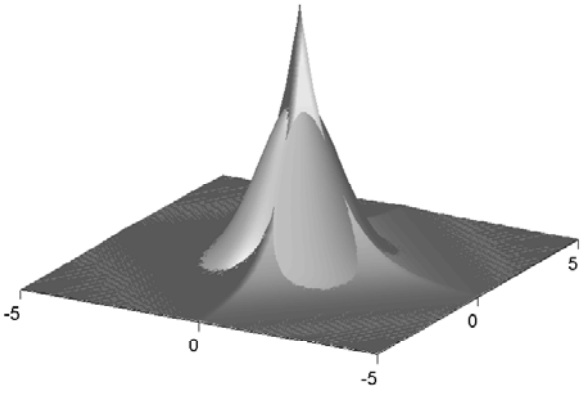


FIGURE 10.6. A superposition of a standard (zero mean, unit variance) two dimensional Gaussian distribution, and a Laplace distribution ($\rho = 1$).

10.2.1.2 Sparse Bayesian Learning

An alternative sparse linear modeling approach was proposed by Tipping (2001) in his work on the relevance vector machine (referred to as SBL here). In this line of work, a zero-mean Gaussian prior is assumed for each of the regression parameters:

[1] It is now well-known that the Lasso does not possess an "Oracle Property," typically failing to set enough components of β to zero. See, for example, Zou (2006).

$$p(\beta|\gamma) = \prod_{j=1}^{p} \mathcal{N}(\beta_j|0,\gamma_j),\ \gamma_j > 0, j = 1,\ldots,p, \tag{10.2.2}$$

where crucially, each unknown weight has a separate non-negative hyperparameter γ_j controlling it's variance (with γ being the p vector of these hyperparameters). In the learning procedure, sparsity is achieved if certain γ_j are set to zero. A further hierarchical specification of the hyperparameters (for both the variance of the weights and the noise) completes the prior specification, with $p(\gamma) = \prod_j Gamma(\gamma_j|a,b)$ and $p(\sigma^2) = Gamma(\sigma^2|c,d)$. In the relevance vector machine and further works however, these priors are specified as flat and hence improper priors (a,b,c,d=0), an important point of difference with what we propose.

Learning in the SBL paradigm involves exact posterior inference for the predictions, where the hyperparameters are chosen to maximize the marginal data likelihood. The literature refers to this procedure as type-II maximum likelihood or evidence maximization (Berger, 1985; Mackay, 1992). Equivalently, SBL minimizes:

$$-\log \int p(\mathbf{y}|\beta)p(\beta|\gamma)d\beta = \log|\Sigma_y| + \mathbf{y}^{\mathrm{T}}\Sigma_y^{-1}\mathbf{y}, \tag{10.2.3}$$

where $\Sigma_y = \sigma^2 I + \Phi\Gamma\Phi^{\mathrm{T}}$ and $\Gamma = \mathrm{diag}[\gamma]$ (see Tipping, 2001 or Wipf and Nagarajan, 2008). This optimization leads to some γ_*, which then leads to the posterior distribution of the weights $p(\beta|\mathscr{D},\gamma_*,\sigma^2) = \mathcal{N}(\beta|\mu,\Sigma)$. Here, $\mu = \Gamma_*\Phi^{\mathrm{T}}\Sigma_{y*}^{-1}\mathbf{y}$ and $\Sigma = \Gamma - \Gamma\Phi^{\mathrm{T}}\Sigma_{y*}^{-1}\Phi\Gamma$ [2]. The expression for the posterior mean μ, further emphasizes how if $\gamma_{*,j} = 0$, the corresponding β_j is also zero and removed from the model. Finally, the predictive density for a new/test point $\phi(\mathbf{x}_t)$ integrates over the posterior density of β leading to a closed-form Gaussian expression:

$$\begin{aligned} p(y_t|\mathscr{D},\gamma_*,\sigma^2,\phi(\mathbf{x}_t)) &= \mathcal{N}(y_t|m_{y_t},\sigma_t). \\ m_{y_t} &= \mu^{\mathrm{T}}\phi(\mathbf{x}_t), \\ \sigma_t^2 &= \sigma^2 + \phi(\mathbf{x}_t)^{\mathrm{T}}\Sigma\phi(\mathbf{x}_t). \end{aligned} \tag{10.2.4}$$

We note that SBL can be shown to be equivalent to Gaussian process regression under particular restrictions; see, for example, Tipping (2001).

The objective function in the SBL optimization problem in (10.2.3) is multimodal, non-convex, and has fixed points at sparse solutions. Various algorithms have been proposed in the literature for obtaining local minima (Mackay, 1992; Tipping, 2001; Tipping and Faul, 2003; Wipf and Nagarajan, 2008).

10.2.1.3 The Bayesian Lasso

The Bayesian Lasso (Park and Casella, 2008) starts with the data model of (10.2.1) and the same Gaussian prior for the weights as in SBL (10.2.2). The hierarchical

[2] The expressions are modeled on the Wipf and Nagarajan (2008) paper, and are equivalent to the ones in the relevance vector machine paper where the notation is slightly different.

prior model differs slightly from that of SBL insofar as the variance parameters are assumed to be drawn from an exponential distribution with rate hyperparameter p-vector λ, instead of a gamma distribution,i.e.:

$$p(\gamma|\lambda) = \prod_{j=1}^{p} \frac{\lambda_j}{2} \exp\Big\{ -\frac{\lambda_j \gamma_j}{2} \Big\},\ \lambda_j > 0, j = 1, \ldots, p.$$

The reason why this relates to the Lasso and sparse learning, is because this particular form of hierarchical prior results in a Laplace prior on β after marginalizing out $\gamma\,(p(\beta) = \int p(\beta|\gamma)p(\gamma|\lambda)d\gamma)$. This result derives from the representation of the Laplace distribution as a scaled mixture of Gaussians with an exponential mixing density (Park and Casella, 2008):

$$\frac{\sqrt{a}}{2} \exp(-\sqrt{a}|z|) = \int_0^\infty \frac{1}{\sqrt{2\pi s}} \exp\{-z^2/(2s)\} \frac{a}{2} \exp(-as/2) ds,\ a > 0.$$

Inference in the Bayesian Lasso is carried out in a fully Bayesian manner via posterior simulation. Exploiting closed form marginal distribution calculations, Park and Casella (2008) outline a Gibbs sampler that can be used to draw samples from the posterior distribution $p(\beta|\mathscr{D})$ (they also propose various techniques to estimate/set/sample from the hyperparameter distribution). While this represents a satisfying Bayesian solution, MCMC sampling poses a significant obstacle in terms of the size of the applications this technique can reasonably be expected to handle. In addition, the Bayesian Lasso does not yield a sparse solution unless ad-hoc rules are used to threshold components of β that are small *a posteriori*. Other minor sampling related drawbacks include difficulty in assessing convergence of the MCMC sampler, and tuning of the sampling algorithm itself.

10.2.2 The demi-Bayesian Lasso

With the above background in place we turn to our proposals. To circumvent the computational complexities associated with the MCMC sampling required for the Bayesian Lasso, we propose fitting the Bayesian Lasso model through a type-II maximum likelihood procedure (i.e., by maximizing the marginal data likelihood). Conceptually, this inherits the benefits of the SBL framework and alleviates the corresponding sampling associated problems. We now find hyperparameters via optimization and not sampling (thus greatly expanding the dimensionality of models that can be learnt efficiently), the resultant posterior distribution is analytically tractable (Gaussian), and sparse models for prediction are obtained without thresholding the posterior distribution. Of course, the flip side is that first, this proposal, like SBL, is less than fully Bayesian, and second, also like SBL, it results in a non-convex optimization problem.

Specifically, we propose to learn the Bayesian Lasso linear model $\mathbf{y} = \Phi\beta + \varepsilon$, with $p(\varepsilon) = \mathcal{N}(\varepsilon|0, \sigma^2 I)$ (we assume σ^2 given in this work, and pick its value from among a set of candidates based on predictive accuracy estimates such as cross validation/validation error). Further, $p(\beta|\gamma) = \mathcal{N}(\beta|0, \Gamma)$ (recall that $\Gamma = \text{diag}[\gamma]$) and we place an exponential prior on the variance components,

$$p(\gamma|\lambda) = \prod_{j=1}^{p} \frac{\lambda_j}{2} \exp - \frac{\lambda_j \gamma_j}{2}.$$

However, as in SBL, we choose to estimate the non-negative hyperparameters γ by type-II maximum likelihood. In other words, we maximize the marginal data likelihood in order to learn the hyperparameters:

$$\begin{aligned} p(\gamma|\mathcal{D}, \lambda) &\propto p(\mathbf{y}|\gamma) p(\gamma|\lambda) \\ &= \left(\int p(\mathbf{y}|\beta) p(\beta|\gamma) d\beta \right) p(\gamma|\lambda). \end{aligned}$$

Taking the negative logarithm, using the result from (10.2.3), and removing quantities irrelevant to the optimization problem results in the following objective function to be minimized:

$$\mathcal{L}(\gamma) = \log |\Sigma_y| + \mathbf{y}^{\mathrm{T}} \Sigma_y^{-1} \mathbf{y} + \lambda \sum_{j=1}^{p} \gamma_j \tag{10.2.5}$$

Note that for parsimony and convenience in further estimation, we set all the $\lambda_j = 2\lambda$, which we assume to be given (again picked from candidates using cross validation). Also note that the key difference compared to SBL is the presence of the proper variance prior, which results in the extra term in (10.2.5) as compared to (10.2.3), and provides extra shrinkage. After obtaining (local) maximum values for the hyperparameters γ_* (the next section outlines algorithms for this purpose), we then make posterior predictions also according to the SBL machinery, via the expressions for $p(y_t|\mathcal{D}, \gamma_*, \sigma^2, \phi(\mathbf{x}_t))$ and the related expressions for the mean and variance, (10.2.4). We call this approach the demi-Bayesian Lasso (dBL).

It is worth mentioning that the above formulation can be obtained by considering the original SBL formulation with a particular form of the Gamma prior on the variance components γ_j. This links the Bayesian Lasso model to the SBL model and provides the motivation for our proper prior on the variances.

10.2.2.1 Algorithms

The key learning task with the model is finding optimal prior variance, γ values. This then allows us to compute the posterior distribution over the weights and compute the posterior predictive distribution (10.2.4). Due to the similarity with the SBL objective function, many of the SBL algorithms apply with minor modifications. Here we discuss two variants. The first is a modification of the EM algorithm that was proposed in Tipping (2001). Starting with some γ, we iteratively apply the E

step:

$$\Sigma = \Gamma - \Gamma \Phi^{\mathrm{T}} \Sigma_{y*}^{-1} \Phi \Gamma,$$

with $\mu = \Gamma_* \Phi^{\mathrm{T}} \Sigma_{y*}^{-1} \mathbf{y}$ and the M step:

$$\gamma_j = \frac{2(\mu_j^2 + \Sigma_{jj})}{1 + \sqrt{1 + 4\lambda(\mu_j^2 + \Sigma_{jj})}},$$

for all $j = 1, \ldots, p$, until convergence. We will refer to this algorithm as EM dBL.

The second variant modifies a recent algorithm by Wipf and Nagarajan (2008) that possesses several nice properties, such as a global convergence analysis (to a local minimum) and sparsity along the solution path. We state the algorithm next, which we will call ℓ_1 dBL, followed by a brief deviation. We refer the reader to Wipf and Nagarajan (2008) for further details.

The ℓ_1 dBL Algorithm

Data: $\mathscr{D}, \lambda, \gamma$.
Result: Sparse β, γ, at each iteration.
Initialize $\beta = \mathbf{0}$, $\mathbf{z} = [1, \ldots, 1]^{\mathrm{T}}$.
while *Convergence criteria not met* **do**

$\beta_* = \operatorname{argmin}_\beta \|\mathbf{y} - \Phi\beta\|_2^2 + 2\sigma^2 \sum_j (z_j + \lambda)^{1/2} |\beta_j|$,
$\gamma_j = (z_j + \lambda)^{-1/2} |\beta_{*,j}|$,
$\mathbf{z}_* = \nabla_\gamma \log |\Sigma_y| = \operatorname{diag}[\Phi^{\mathrm{T}} \Sigma_y^{-1} \Phi]$,
$\beta = \beta_*$,
$\mathbf{z} = \mathbf{z}_*$.

end
$\beta = E[\beta | \mathbf{y}, \gamma_*] = \Gamma_* \Phi^{\mathrm{T}} \Sigma_{y*}^{-1} \mathbf{y}$.

The algorithm outlined above is guaranteed to converge monotonically to a local minimum or saddle point of (10.2.5). This follows trivially from Theorem 1 and analysis in Wipf and Nagarajan (2008). The algorithm notably uses iterated reweighted ℓ_1 regression (step 1 in the while loop) to estimate the weights β, also known as an adaptive Lasso problem (Zou, 2006). The ℓ_1 penalty results in sparse β, which correspondingly results in sparse estimates of variance components γ—we will refer to this algorithm as ℓ_1 dBL. The auxiliary variables $\mathbf{z}$ (a p-vector) arise from the upper bound of the log-determinant term (see below). The choice of an Exponential prior results in very small computational difference between the SBL algorithm in Wipf and Nagarajan (2008) and the one presented here. In particular, replacing $z_j + \lambda$ with z_j is the only difference. Similarly, the prior results in a small difference in the M-step in the corresponding update in Tipping (2001) algorithm, where it is: $\gamma_j = \mu_j^2 + \Sigma_{jj}$. As expected, the proper prior results in additional regularization of the variance parameters towards zero. We expect that this additional regularization will come with a bias-variance trade-off, the additional flexibility created by the single extra parameter λ potentially allowing us to generalize better.

Deriving the ℓ_1 dBL algorithm. Here we briefly outline the algorithm derivation. The log-determinant term in $\mathscr{L}(\gamma)$ (10.2.5) is concave in γ, and so can be expressed via

$$\log|\Sigma_y| = \min_{\mathbf{z}}\{\mathbf{z}^{\mathrm{T}}\gamma - g^*(\mathbf{z})\}.$$

In that expression, $g^*(\mathbf{z})$ is the concave conjugate of $\log|\Sigma_y|$, $g^*(\mathbf{z}) = \min_{\gamma}\mathbf{z}^{\mathrm{T}}\gamma - \log|\Sigma_y|$. This then leads to the upper bounding cost function:

$$\mathscr{L}(\gamma,\mathbf{z}) = \mathbf{z}^{\mathrm{T}}\gamma - g^*(\mathbf{z}) + \mathbf{y}^{\mathrm{T}}\Sigma_y^{-1}\mathbf{y} + \lambda\sum_{j=1}^{p}\gamma_j \geq \mathscr{L}.$$

Following Wipf and Nagarajan (2008), the optimal $\mathbf{z}$ occurs when

$$\mathbf{z}_* = \nabla_{\gamma}\log|\Sigma_y| = \mathrm{diag}[\Phi^{\mathrm{T}}\Sigma_y^{-1}\Phi].$$

Re-expressing the term

$$\mathbf{y}^{\mathrm{T}}\Sigma_y^{-1}\mathbf{y} = \min_{\beta}\frac{1}{\sigma^2}\|\mathbf{y}-\Phi\beta\|_2^2 + \sum_j\frac{\beta_j^2}{\gamma_j},$$

we get an upper bounding term

$$\mathscr{L}_{\mathbf{z}}(\gamma,\beta) = \frac{1}{\sigma^2}\|\mathbf{y}-\Phi\beta\|_2^2 + \sum_{j=1}^{p}\left((z_j+\lambda)\gamma_j + \frac{\beta_j^2}{\gamma_j}\right) \geq \mathscr{L}$$

which is jointly convex in β and γ, which can be globally minimized solving for γ and then β (Wipf and Nagarajan, 2008). Now, for any β, $\gamma_j = (z_j+\lambda)^{-1/2}|\beta_j|$ minimizes $\mathscr{L}_{\mathbf{z}}(\gamma,\beta)$. This then results in the ℓ_1 dBL algorithm which is an iterative application of the steps of finding the optimal γ (minimizing the upper bounding cost), and then finding the optimal $\mathbf{z}$ (which then leads to recomputing the optimal upper bounding cost).

10.2.2.2 An EN Heuristic

While the use of of iterated re-weighted ℓ_1 regularized regression results in sparsity which is desirable, it also inherits some of the drawbacks of ℓ_1 regression. In particular, an issue of concern is the instability of ℓ_1 regression solutions with respect to highly correlated regressors (Zou and Hastie, 2005). Essentially, with highly correlated regressors/basis functions, the weights β computed based on the ℓ_1 solution are unstable—small differences in the dataset can result in the selection of very different subsets of a set of correlated regressors[3]. Zou and Hastie's (2005) "elastic net" seeks to address this issue. The elastic net imposes an $\alpha\ell_1 + (1-\alpha)\ell_2$

[3] For two perfectly correlated relevant regressors, one amongst them is chosen to have a non-zero weight either at random or due to the particulars of the the algorithm implementation.

penalty, $0 \leq \alpha < 1$, on the weights. This has the attractive property of imposing a simple additional convex loss and encourages a "grouping effect" which helps keep weights on correlated regressors similar (Theorem 1 in Zou and Hastie, 2005). Zou and Hastie (2005) show good results when applying this mixed penalty.

We attempt to capture the same effect in the dBL. This is done by solving an elastic net problem in the ℓ_1 dBL algorithm instead of the re-weighted ℓ_1 regression problem. Unfortunately, the heuristic doesn't correspond to an intuitive prior on the variance components and further is strongly tied to the iterated re-weighted ℓ_1 regression algorithm (an equivalent is hard to define for the EM style algorithms). Nonetheless, we explore this heuristic in the experiments that follow—we will refer to this as dBL+EN below.

10.2.3 Experiments and Results

We now turn to evaluation of the dBL via experimental studies. We consider both simulation studies and three real data examples from the literature and evaluate the strengths and weaknesses of the proposal.

10.2.3.1 Simulation Studies

Our simulation study models are based on the studies in (Zou and Hastie, 2005). (Examples 10.2 and 10.4 correspond exactly, Examples 10.1 and 10.3 are minor modifications of examples in their work). The aim is to highlight the differences between the techniques in terms of predictive performance, but also in terms of variable selection accuracy. We present five simulation study examples, each of which consist of a training set, a validation set and a test set (all independent). Models are fit using the training data only, and parameters/hyperparameters selected from appropriate grids on reasonable values using the validation set. For the EN heuristic, in all experiments we set the ℓ_1/ℓ_2 blending parameter $\alpha = 0.7$. Borrowing notation from Zou and Hastie (2005), we use $x/y/z$ to denote x training observations (size of the training data), y validation and z independent test samples. The four examples attempt to gauge the performance of the methods in various scenarios:

Example 10.1. We simulate 200 data sets consisting of 20/20/200 observations with 8 predictors. The data generating mechanism is a linear model with $\mathbf{y} = \Phi\beta + \kappa\varepsilon$, where $p(\varepsilon) = \mathcal{N}(\varepsilon|0, I)$ and $\kappa = 3$. We set $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)'$. The pairwise correlation between Φ_i and Φ_j is set as $\mathrm{Cov}(i, j) = \rho^{|i-j|}$. In this example, the covariance matrix is an identity matrix, $\mathrm{Cov}(i, j) = 0$ for all $i \neq j$ and $\mathrm{Cov}(i, i) = 1$. Finally, Φ is drawn from a multivariate Gaussian with zero mean and the above covariance matrix.

Example 10.2. This example is entirely analogous to Example 10.1 except with non-identity covariance (introducing mild correlation between the regressors). Here, $\rho = 0.5$.

Example 10.3. This example the same as Examples 10.1 and 10.2, except with higher correlation between the regressors. Here, $\rho = 0.85$.

Example 10.4. This is an example where the data generating mechanism is a linear model. We simulate 200 data sets with 100/100/400 observations and 40 predictors. This time $\beta = (0,\ldots,0,2,\ldots,2,0,\ldots,0,2,\ldots,2)'$, with alternating blocks of 10 indices of zeros and 2s. Here, $\kappa = 15$ and $\mathrm{Cov}(i,j) = 0.5$ for all $i \neq j$, and $\mathrm{Cov}(i,i) = 1$.

Example 10.5. This is an example where the data generating mechanism is *not* a linear model. Here we will not be able to gauge variable selection accuracy, but only predictive performance. In this case we include some interaction terms and powers of the regressors in computing the response. We simulate 40/40/400 observations following polynomial model (for a single observation): $y = 1.5\phi_1^2 + 2\phi_1\phi_2 - \phi_5\phi_1 + \phi_5^3 + 2\phi_7 + 3\varepsilon$ where ε is a zero mean, unit variance Gaussian error. The learning algorithms only get access to Φ and the responses.

For Examples 10.1 through 10.4, we compute the following quantities: i) the mean squared error (MSE), computed on test data, ii) mean "parametric" error (MPE), that is, the mean of the quantity $(\beta - \beta_{true})'\Sigma(\beta - \beta_{true})$, where Σ is the covariance of Φ. This attempts to quantify closeness to the parameters that actually generated the data. iii) Quantities related to structural errors: mean C ($\bar{C}$) and mean IC ($\bar{IC}$). C is defined as the number of true weights that were zero which are correctly estimated as zero by the model (thus higher values are better). Similarly, IC is defined as the number of non-zero true weights incorrectly estimated as zero by the model (and thus lower IC values are preferred). Models that are excessively sparse would tend to have high C values (good) and high IC values (not good). A model that is completely non-sparse would have the lowest possible C value (bad) but the lowest IC values (good) as well. For Example 10.5, since the data generating mechanism is outside the model hypothesis class we only report the test mean squared error.

We evaluate the optimization based approaches, namely the Lasso (Lasso in the results), the original SBL algorithm (Tipping, 2001, SBL), the Wipf and Nagarajan (2008) SBL algorithm (ℓ_1 SBL), the dBL model with parameters found using the EM algorithm (EM dBL) and the ℓ_1 variation (the ℓ_1 dBL algorithm) and finally, the ℓ_1 based proposal with the EN heuristic (dBL + EN).

Table 10.4 and Figure 10.7 show the results. In all cases (modest to large) improvements are made over the flat-prior variants and over the Lasso both in terms of prediction accuracy as well as structural accuracy. In the tables we show standard errors of the estimates, and in Figure 10.7, we show boxplots of the squared error showing the median, lower and upper quartiles, whiskers and outliers. We next turn to some real data examples.

10.2.3.2 Prostate Cancer Data

The data in this example comes from a prostate cancer study done by Stamey (1989). Eight clinical measurements serve as the regressors, which are, in order: log(cancer

TABLE 10.4. Simulation study results.

	Lasso	SBL	ℓ_1 SBL	EM dBL	ℓ_1 dBL	dBL + EN
Example 10.1						
MSE	14.40 (0.28)	14.39 (0.31)	14.66 (0.34)	14.23 (0.29)	**14.04** (0.30)	14.11 (0.28)
MPE	3.99 (0.21)	4.02 (0.24)	4.25 (0.27)	3.83 (0.22)	**3.64** (0.23)	3.70 (0.21)
$\bar{C}$	2.23 (0.12)	2.59 (0.12)	3.51 (0.10)	2.21 (0.11)	**3.61** (0.10)	3.29 (0.09)
$I\bar{C}$	0.24 (0.04)	0.26 (0.04)	0.38 (0.05)	**0.22** (0.04)	0.30 (0.04)	0.26 (0.04)
Example 10.2						
MSE	14.63 (0.36)	14.84 (0.37)	15.12 (0.42)	14.44 (0.36)	14.42 (0.36)	**14.22** (0.37)
MPE	3.91 (0.22)	4.12 (0.23)	4.44 (0.30)	3.72 (0.21)	3.72 (0.21)	**3.53** (0.22)
$\bar{C}$	2.24 (0.11)	2.77 (0.11)	3.56 (0.11)	2.23 (0.10)	3.58 (0.10)	**3.25** (0.10)
$I\bar{C}$	**0.22** (0.03)	0.39 (0.04)	0.48 (0.05)	**0.22** (0.03)	0.36 (0.04)	0.23 (0.03)
Example 10.3						
MSE	14.20 (0.32)	14.42 (0.31)	15.20 (0.42)	13.83 (0.30)	13.99 (0.30)	**13.44** (0.28)
MPE	3.33 (0.17)	3.56 (0.17)	4.21 (0.29)	2.96 (0.15)	3.09 (0.15)	**2.53** (0.13)
$\bar{C}$	2.42 (0.09)	2.85 (0.10)	3.23 (0.09)	2.52 (0.09)	**3.48** (0.09)	2.77 (0.08)
$I\bar{C}$	0.65 (0.05)	0.78 (0.05)	0.91 (0.05)	0.58 (0.05)	0.92 (0.05)	**0.38** (0.04)
Example 10.4						
MSE	316.92 (2.41)	311.37 (2.32)	327.13 (2.78)	283.72 (1.95)	286.50 (2.03)	**261.14** (1.67)
MPE	83.74 (1.64)	77.37 (1.39)	93.55 (2.02)	49.28 (0.90)	52.19 (1.03)	**26.22** (0.49)
$\bar{C}$	9.72 (0.34)	14.61 (0.24)	8.39 (0.16)	11.99 (0.19)	**14.66** (0.18)	8.03 (0.21)
$I\bar{C}$	5.92 (0.19)	9.87 (0.19)	5.79 (0.13)	6.58 (0.14)	8.77 (0.15)	**2.52** (0.12)
Example 10.5						
MSE	30.78 (0.40)	30.56 (0.42)	31.64 (0.47)	30.32 (0.40)	**30.07** (0.40)	30.37 (0.40)

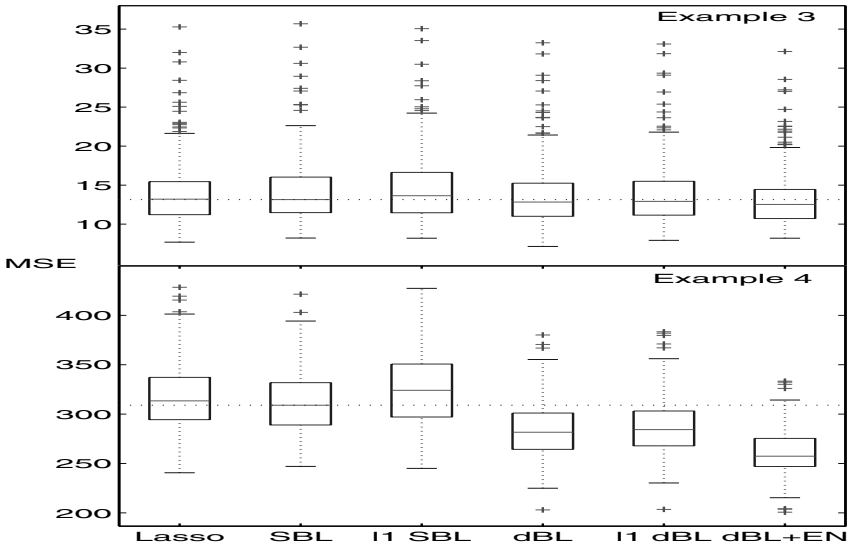


FIGURE 10.7. Boxplots for simulation studies 3 and 4, where Examples 3 and 4 correspond to Examples 10.3 and 10.4. The horizontal dashed line is a visual guide and marks the location of the minimum median from amongst the prior art, namely the Lasso, SBL and ℓ_1 SBL.

volume) *lcavol*, log(prostate weight) *lweight*, *age*, log(amount of benign prostatic hyperplasia) *lbph*, seminal vesicle invasion *svi*, log(capsular penetration) *lcp*, Gleason score *gleason* and percentage Gleason score 4 or 5 *pgg*45. The predictive quantity of interest is the log(prostate specific antigen) *lpsa*.

Following Zou and Hastie (2005), we divide the data into two parts, a training set with roughly two thirds the number of observations, 64 observations and a test set with 33 observations. Hyperparameters were selected from a grid of values via 10-fold cross validation using only the training data[4]. The methods are compared via the prediction mean-squared error on the test data.

TABLE 10.5. Prostate data results.

	Lasso	**SBL**	**ℓ_1 SBL**	**EM dBL**	**ℓ_1 dBL**	**dBL + EN**
MSE	0.4505	0.5539	0.5765	0.5367	**0.3781**	0.5216
Vars	all	all but 6	all but 6	all but 6	(1,3,4,5)	(1,5,7)
σ^2 (λ)	0.5	0.2	0.01	0.25 (1)	0.005 (500)	0.25 (1)

Our results (Table 10.5) show improved performance of the proposals over the Lasso[5] and SBL, with the ℓ_1 dBL providing the best performance. There is broad consensus on the selected variables, with *lcp* being rejected by all models in our experiments.

10.2.3.3 Diabetes Data

The data in this study come from Efron et al. (2004). The response is a quantitative measure of diabetes progression in 398 patients one year after baseline. The predictors include age, sex, body mass index, average blood pressure, and six blood serum measurements, for a total of 10 regressors. As Efron et al. (2004) point out, linear models are especially useful in this diagnostic application, because in addition to predictive accuracy for future patients, the models would ideally provide disease progression guidance by being interpretable. We standardized the regressors to have zero mean 0 and unit variance.

We partition the data into a 266 patient training sample and a 132 patient test sample. Hyperparameters were selected from a grid of values via 10-fold cross validation using only the training data. We show test mean squared error, variables selected and parameters (and hyperparameters used).

Our results agree with many reported findings on this dataset, and in our experiments, the dBL + EN variant proved predictively best by a slight margin (Table 10.6). In terms of variable selection, the least important regressors appear to be 1,

[4] For all the real data examples, we select the hyperparameters following the slight modification to k-fold CV suggested in Chapter 7 of Hastie, Tibshirani, and Friedman (2001), namely we pick the largest amount of regularization that is within 1 standard error of the minimum CV error.

[5] Note that in the table, the σ^2 is a proxy label for the regularization parameter for the Lasso.

TABLE 10.6. Diabetes data results.

	Lasso	**SBL**	**ℓ_1 SBL**	**EM dBL**	**ℓ_1 dBL**	**dBL + EN**
MSE	3031.2	3045.1	3032.4	3034.2	3031.2	**3029.3**
Vars	all but 1,6,8	all but 1,7	all	all but 1,8	all but 1,6,8	all
σ^2 (λ)	1	500	500	500 (0.001)	100 (0.1)	500 (0.001)

6 and 8, which is also evident from the findings in Park and Casella (2008). Note that in our experiments, the SBL model seems to deselect regressor 7, which is an anomaly.

10.2.3.4 Biscuit NIR Data

In this application, we examine the biscuit dough data from Brown, Fearn, and Vannucci (1999). The response we look at is fat content of the dough (centered), and the regressors are spectral characteristics of the dough, measured using near infrared (NIR) spectroscopy (standardized). The spectral characteristics are described using a grid of wavelengths, in particular reflectance measured at every 4nm from the range of wavelengths: 1202–2400 nm. The data is split into 39 training samples and 31 test samples, and we standardize the regressors.

Hyperparameters were selected from a grid of values via 5-fold cross validation using only the training data. The methods are compared via the prediction mean-squared error on the test data.

TABLE 10.7. Biscuit NIR data results.

	Lasso	**SBL**	**ℓ_1 SBL**	**EM dBL**	**ℓ_1 dBL**	**dBL + EN**
MSE	0.0565	0.0551	0.0696	0.0543	**0.0450**	0.1001
|Non-zero Vars|	18	6	269	11	54	43
σ^2 (λ)	1.25	0.25	0.05	0.2 (0.1)	0.15 (0.2)	1 (1)

Our results (Table 10.7 and Figure 10.8 are consistent with previous studies that use this data (West, 2003) and we find ℓ_1 dBL gives the best performance. In Figure 10.8, due to the coarse resolution of the plot, only high magnitude weights can be discerned. Note the similarity between the high magnitude weights of the Lasso and SBL solutions (the EM SBL high magnitude weights are very similar and hence omitted). In particular, the non-zero β found by ℓ_1 dBL around 1710 nm are significant because fat is known to have a characteristic absorbance in this range. Also note that for this example, the dBL + EN heuristic appears to perform worse than the others.

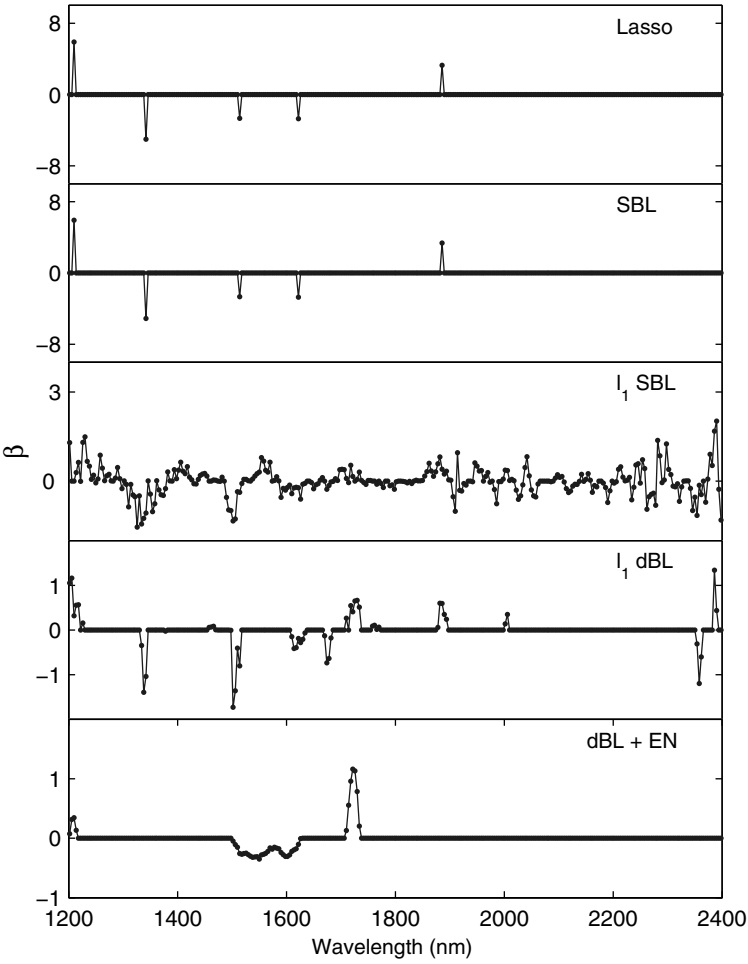


FIGURE 10.8. Biscuit data β values. Shown from top to bottom are the final parameter weights for the Lasso, SBL, ℓ_1 SBL, ℓ_1 dBL and dBL + EN.

10.2.4 Discussion

In this section we examine the use of proper priors in sparse Bayesian learning and showed some promising experimental results. We show that with a single additional hyperparameter (set through cross-validation), the model is augmented substantially enough to make better predictions. Further, the choice of an exponential distribution as a prior connects SBL to the recently proposed Bayesian Lasso, with our proposal amounting to an attractive alternative way of estimating Bayesian Lasso model hyperparameters by maximizing marginal likelihood rather than Monte Carlo simulation. We also explore the use of an EN-heuristic that, in our experiments, leads to

better performance in the presence of correlated regressors. In future work we would like to extend the proposals to classification problems. We would also like to examine the efficient SBL algorithm of Tipping and Faul (2003) to see if an analogous procedure can be applied in this case as well. Finally, other forms of prior distribution on the variance are the topic of our further exploration, including additionally sparsifying priors like the Laplace distribution etc.

10.3 Hierarchical Bayesian Mixed-Membership Models and Latent Pattern Discovery

Edoardo M. Airoldi, Stephen E. Fienberg, Cyrille J. Joutard, and Tanzy M. Love

Although hierarchical models have dominated the Bayesian literature since the early 1970s, several variations on *Hierarchical Bayesian Mixed-Membership Models* (HBMMMs) have recently gained popularity thanks to their ability to deal with minimal information and noisy labels in a systematic fashion. These models allow each subject of study, e.g., documents or individuals, to belong to more than one class, group, or cluster (Erosheva, Fienberg, and Lafferty, 2004; Erosheva and Fienberg, 2005; Airoldi et al., 2006, 2008; Airoldi, 2007).

We can specify HBMMMs in terms of a hierarchy of probabilistic assumptions (i.e., a directed acyclic graph) that involves: (i) observations, x, (ii) latent variables, λ, and (iii) parameters (constants) for the patterns associated with the groups or clusters, θ. The likelihood of the data is then $\ell(x|\theta) = \int_\lambda \ell(x, \lambda|\theta) D_\alpha(d\lambda)$, where $D_\alpha(d\lambda)$ is a prior distribution over the latent variables. During pattern discovery, i.e., posterior inference, we condition on the values of the observed data and maximize the likelihood with respect to a set of parameters θ that describe the group patterns.

The focus in pattern discovery with HBMMMs is not on the variable amount of information about the labels for the objects; rather, it is on the hierarchy of probabilistic assumptions that the analyst believes provide the structure underlying the data, which ultimately lead to the likelihood function. Whatever the amount of information about the class labels, full, partial, minimal, or none, we simply treat the information as observations about the attributes and we condition upon it. The missing information about the labels or weights on the classes or groups is recovered during pattern discovery (i.e., via posterior inference) as is the information about other non-observable patterns. In this sense, HBMMMs are essentially *soft-clustering* models in that the *mixed-membership* error model for the labels associates each observation with a vector of memberships that sum to one.

Because of their flexibility, instances of HBMMMs have gained popularity in a variety of applications, e.g., population genetics (Pritchard, Stephens, and Donnelly, 2000; Rosenberg et al., 2002), scientific publications (Blei, Ng, and Jordan,

2003; Erosheva, Fienberg, and Lafferty, 2004; Griffiths and Steyvers, 2004), words and images (Barnard et al., 2003), disability analysis (Erosheva, 2002a,b, 2003; Erosheva, Fienberg, and Joutard, 2007), fraud detection (Neville et al., 2005), biological sequences and networks (Airoldi et al., 2008). HBMMMs are closely related to popular unsupervised data mining methods such as probabilistic principal component analysis (Tipping and Bishop, 1999a), parametric independent component analysis, mixtures of Gaussians, factor analysis, and hidden Markov models (Rabiner, 1989).

A fundamental issue of HBMMMs is that of *model choice*, involving the choice of the number of latent categories, groups, or clusters. Positing an explicit model for the category labels requires a choice regarding the number of existing categories in the population, i.e., the *choice* of the crucial model dimension. A parametric model for the labels would assume the existence of a predetermined number, K, of categories, whereas a nonparametric error model would let the number of categories grow with the data. We explore the issue of model choice in the context of HBMMMs, both theoretically and computationally, by investigating the nexus between strategies for model choice, estimation strategies, and data integration in the context of data extracted from scientific publications and measures of disability for Americans aged 65+, cf. (Erosheva, Fienberg, and Joutard, 2007; Joutard et al., 2007; Airoldi et al., 2009).

Overview of the section. In this section, we (i) describe HBMMMs a class of models that respond to the challenges introduced by modern applications, and we characterize HBMMMs in terms of their essential probabilistic elements; (ii) identify the issue of *model choice* as a fundamental task to be solved in each applied data mining analysis that uses HBMMMs; (iii) survey several of the existing strategies for model choice; (iv) develop new model specifications, as well as use old ones, and we employ different strategies of model choice to find "good" models to describe problems involving text analysis and survey data; (v) study what happens as we deviate from statistically sound strategies in order to cut down the computational burden, in a controlled experimental setting.

PNAS biological sciences collection. Our data consists of abstracts and references for a collection of articles from the *Proceedings of the National Academy of Sciences* for 1997–2001. Erosheva, Fienberg, and Lafferty (2004) and Griffiths and Steyvers (2004) report on their estimates about the number of latent topics, and find evidence that supports a small number of topics (e.g., as few as 8 but perhaps a few dozen) *or* as many as 300 latent topics, respectively. There are a number of differences between the two analyses: the collections of papers were only partially overlapping (both in time coverage and in subject matter), the authors structured their dictionary of words differently, one model could be thought of as a special case of the other but the fitting and inference approaches had some distinct and non-overlapping features. The most remarkable and surprising difference comes in the estimates for the number of latent topics: Erosheva, Fienberg, and Lafferty (2004) focus on values such as 8 and 10, but admit that a careful study would likely produce somewhat higher values, while Griffiths and Steyvers (2004) present analyses they claim support on the order of 300 topics! Should we want or believe that there are only a dozen or so

topics capturing the breadth of papers in PNAS or is the number of topics so large that almost every paper can have its own topic? A touchstone comes from the journal itself, which states that it classifies publications in biological sciences according to 19 topics. When submitting manuscripts to PNAS, authors select a major and a minor category from a predefined list of 19 biological science topics, and possibly those from the physical and/or social sciences.

Here, we develop an alternative set of analyses using the version of the PNAS data on biological science papers analyzed in Erosheva, Fienberg, and Lafferty (2004). We employ both parametric and non-parametric strategies for model choice, and we make use of both text and references of the papers in the collection. This case study gives us a basis to discuss and assess the merit of the various model choice strategies.

Disability survey data. In the second example, we work with data extracted from the National Long-Term Care Survey (NLTCS) by Erosheva (2002a) to illustrate the important points of our analysis. The NLTCS is a longitudinal survey of the U.S. population aged 65 years and older with waves conducted in 1982, 1984, 1989, 1994, 1999, and 2004. It is designed to assess chronic disability among the US elderly population, especially those who show limitations in performing some activities that are considered normal for everyday living. These activities are divided into *activities of daily living* (ADLs) and *instrumental activities of daily living* (IADLs). ADLs are basic activities of hygiene and healthcare: eating, getting in/out of bed, moving inside the house, dressing, bathing and toileting. IADLs are basic activities necessary to reside in the community: doing light and heavy housework and laundry, cooking, grocery shopping, moving outside the house, traveling, managing money, taking medicine, and telephoning. The data extract we work with consists of combined data from the first four survey waves (1982, 1984, 1989, 1994) with $21,574$ individuals and 16 variables (6 ADLs and 10 IADLs). For each activity, individuals are either disabled or healthy on that activity. We then deal with a 2^{16} contingency table. Of the $2^{16} = 65,536$ possible combinations of response patterns, only $3,152$ occured in the NLTCS sample.

Here we complement the earlier analyses in Erosheva (2002a), Erosheva and Fienberg (2005), and Erosheva, Fienberg, and Joutard (2007) and employ both parametric and non-parametric strategies for model choice. We focus on increasing the number of latent profiles to see if larger choices of K result in better descriptions of the data and to find the value of K which best fits the data.

From the case studies we learn that: (i) Independently of the goal of the analysis, e.g., predictive versus descriptive, similar probabilistic specifications of the models often support similar "optimal" choices of K, i.e., the number of latent groups and patterns; (ii) Established practices aimed at reducing the computational burden while searching for the best model lead to biased estimates of the optimal choices for K, i.e., the number of latent groups and patterns.

Arriving at a "good" model is a central goal of empirical analyses. These models are often useful in a predictive sense. Thus our analyses in this section are relevant as input to those managing general scientific journals as they re-examine current indexing schemes or consider the possible alternative of an automated indexing sys-

tem, and to those interested in the implications of disability trends among the US elderly population as the rapid increase in this segment of the population raises issue of medical care and the provision of social security benefits.

10.3.1 Characterizing HBMM Models

There are a number of earlier instances of mixed-membership models that have appeared in the scientific literature, e.g., see the review in Erosheva and Fienberg (2005). A general formulation due to Erosheva (2002a), and also described in Erosheva, Fienberg, and Lafferty (2004), characterizes the models of mixed-membership in terms of assumptions at four levels. In the presentation below, we denote subjects with $n \in [1,N]$ and observable response variables with $j \in [1,J]$.

A1–Population Level. Assume that there are K classes or sub-populations in the population of interest and J distinct characteristics measured on each subject. We denote by $f(x_{nj}|\theta_{jk})$ the probability distribution of jth response variable in the kth sub-population for the nth subject, where θ_{jk} is a vector of relevant parameters, $j \in [1,J]$ and $k \in [1,K]$. Within a subpopulation, the observed responses are assumed to be independent across subjects *and* characteristics.

A2–Subject Level. The components of the membership vector $\lambda_n = (\lambda_{n[1]}, \ldots, \lambda_{n[K]})'$ represent the mixed-membership of the nth subject to the various sub-populations. Conditional on the mixed-membership scores, the response variables x_{nj} are independent of one another, and independent across subjects.

A3–Latent Variable Level. Assume that the vectors λ_n, i.e., the mixed-membership scores of the nth subject, are realizations of a latent variable with distribution D_α, parameterized by vector α.

A4–Sampling Scheme Level. Assume that the R replications of the J distinct response variables corresponding to the nth subject are independent of one another. The probability of observing $\{x^r_{n1}, \ldots, x^r_{nJ}\}_{r=1}^R$, given the parameters, is then

$$P(\{x^r_{n1}, \ldots, x^r_{nJ}\}_{r=1}^R | \alpha, \theta) = \int \left(\prod_{j=1}^{J} \prod_{r=1}^{R} \sum_{k=1}^{K} \lambda_{n[k]} f(x^r_{nj}|\theta_{jk}) \right) D_\alpha(d\lambda).$$

The number of observed response variables is not necessarily the same across subjects, i.e., $J = J_n$. Likewise, the number of replications is not necessarily the same across subjects and response variables, i.e., $R = R_{nj}$.

10.3.2 Strategies for Model Choice

Although pathological cases can be built, where slightly different model specifications lead to quite different analyses, in real situations we expect models with similar probabilistic specifications to suggest an optimal number of groups, K, in the same ballpark.

In our application to scientific publications and survey data we explore the issue of model choice by means of different criteria, two of which are popular in the data mining community; namely, cross-validation and a Dirichlet process prior (Antoniak, 1974; Hastie, Tibshirani, and Friedman, 2001).

Cross-validation is a popular method to estimate the generalization error of a prediction rule (Hastie, Tibshirani, and Friedman, 2001), and its advantages and flaws have been addressed by many in that context, e.g., see Ng (1997). More recently, cross-validation has been adopted to inform the choice about the number of groups and associated patterns in HBMMMs (Barnard et al., 2003; Wang, Mohanty, and McCallum, 2005). Guidelines for the proper use of cross-validation in choosing the optimal number of groups K, however, has not been systematically explored. One of the goals of our case studies is that of assessing to what extent cross-validation can be trusted to estimate the underlying number of topics or disability profiles. In particular, given the non-negligible influence of hyper-parameter estimates in the evaluation of the held-out likelihood, i.e., the likelihood on the testing set, we discover that it is important not to bias the analysis towards unprincipled estimates of such parameters, or with arbitrary ad-hoc choices that are not justifiable using preliminary evidence, i.e., either in the form of prior knowledge, or outcome of the analysis of training documents. Expert prior information was sought, but those consulted expressed views on only a relatively narrow component of the data that did not inform the hyper-parameters. In this situation, estimates obtained following good statistical properties, e.g., empirical Bayes or maximum likelihood estimates, should be preferred to others (Carlin and Louis, 2005).

Positing a Dirichlet process prior on the number of latent topics is equivalent to assuming that the number of latent topics grows with the log of the number of documents or individuals (Ferguson, 1973; Antoniak, 1974). This is an elegant model selection strategy in that the selection problem becomes part of the model itself, although in practical situations it is not always possible to justify. A nonparametric alternative to this strategy uses the Dirichlet Process prior as an infinite dimensional prior with a specific parametric form as a way to mix over choices of K, e.g., see McAuliffe, Blei, and Jordan (2006). This prior appears reasonable for static analyses of scientific publications that appear in a specific journal.

The statistical and data mining literatures contain many criteria and approaches to deal with the issue of model choice, e.g., reversible jump MCMC techniques, Bayes factors and other marginal likelihood methods, cross-validation, and penalized likelihood criteria such as the Bayesian Information Criterion (BIC), the Akaike information criterion (AIC), the deviance information criterion (DIC), and minimum description length (MDL). For further details and discussion see Joutard et al. (2007).

10.3.3 Case Study: PNAS 1997–2001

In this section we introduce model specifications to analyze the collection of papers published in PNAS, which were submitted by the respective authors to the section on biological sciences. Earlier related analyses appear in Erosheva, Fienberg, and Lafferty (2004) and Griffiths and Steyvers (2004). After choosing an optimal value for the number of topics, K^*, and its associated words and references usage patterns, we also examine the extent to which they correlate with the actual topic categories specified by the authors.

We organize our models into finite and infinite mixtures, according to the dimensionality of the prior distribution, D_α, posited at the latent variable level. We characterize an article, or document, by the words in its abstract and the references in its bibliography. Introducing some notation, we observe a collection of N documents. The nth document is represented as (x_{1n}, x_{2n}). We assume that words and references come from finite discrete sets (vocabularies) of sizes V_1 and V_2, respectively. For simplicity, we assume that the vocabulary sets are common to all articles, independent of the publication time, although this assumption can be relaxed (Joutard et al., 2007). We assume that the distribution of words and references in an article is driven by an article's membership in each of K basis categories, $\lambda = (\lambda_1, \ldots, \lambda_K)$, and we denote the probabilities of the V_1 words and the V_2 references in the kth pattern by θ_{k1} and θ_{k2}, for $k = 1, 2, \ldots, K$. These vectors of probabilities define the multinomial distributions over the two vocabularies of words and references for each basis semantic pattern. Below, whenever the analysis refers to a single document, we omit the document index n.

10.3.3.1 Finite Mixture Model

For an article with R_1 words in the abstract and R_2 references in the bibliography, the generative sampling process is as follows:

1. Sample $\lambda \sim \text{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_K)$, where $\alpha_k = \alpha$, for all k.
2. Sample $x_1 \sim \text{Multinomial}(p_\lambda, R_1)$, where $p_\lambda = \sum_{k=1}^{K} \lambda_k \theta_{k1}$.
3. Sample $x_2 \sim \text{Multinomial}(q_\lambda, R_2)$, where $q_\lambda = \sum_{k=1}^{K} \lambda_k \theta_{k2}$.

The conditional probability of words and references in a article is then

$$P\{(x_1, x_2)|\theta, \alpha\} = \int \prod_{j=1}^{2} \prod_{v=1}^{V_j} \left(\sum_{k=1}^{K} \lambda_k \theta_{kj[v]} \right)^{x_{j[v]}} dD_\alpha(\lambda).$$

The hyper-parameters of this model are the symmetric Dirichlet parameter α, and the multinomial parameters for words, θ_{k1}, and references, θ_{k2}, for each of the latent topics[6] $k = 1, \ldots, K$. That is, through corresponding pairs of θ vectors (θ_{k1} and

[6] In this application, we refer to the sub-populations of assumption A1 in Section 10.3.1 as "topics." Despite the suggestive semantics, topics are pairs of latent distributions over the vocabulary and the set of known citations, from a statistical perspective.

θ_{k2}) we define a parametric representation of each of the K sub-populations (see assumption A1 in Section 10.3.1), which we refer to as topics in this application. Technically, they are pairs of latent distributions over the vocabulary and the set of known citations. In other words, element v of θ_{k1} encodes the probability of occurrence of the v-th word in the vocabulary (containing V_1 distinct words) when the k-th topic is active, with the constraint that $\sum_v \theta_{k1[v]} = 1$ for each k. Similarly, element v of θ_{k2} encodes the probability of occurrence of the v-th reference in the set of known citations (V_2 of them) when the k-th topic is active. In this finite mixture model, we assume that the number of latent topics is unknown but fixed at K. Our goal is to find the optimal number of topics, K^*, which gives the best description of the collection of scientific articles.

10.3.3.2 Infinite Mixture Model

In the infinite mixture case we posit a simpler and more traditional type of clustering model, by assuming that each article is generated by one single topic. However, in this case we do not need to fix the unknown number of topics, K, prior to the analysis. This full membership model can be thought of as a special case of the mixed membership model where, for each article, all but one of the membership scores are restricted to be zero. As opposed to traditional finite mixture models that are formulated conditional on the number of latent categories, however, this model variant allows the joint estimation of the characteristics of the latent categories, θ, and of the number of latent categories, K. That is, prior to the analysis, the number of sub-populations (see assumption A1 in Section 10.3.1) is unknown and possibly infinite.

We assume an infinite number of categories and implement this assumption through a Dirichlet process prior for λ, D_α, introduced and discussed in Ferguson (1973) and Neal (2000). The distribution D_α models the prior probabilities of latent pattern assignment for the collection of documents. In particular, for the nth article, given the set of assignments for the remaining articles, λ_{-n}, this prior puts a probability mass on the kth pattern (out of K distinct patterns observed in λ_{-n}) which is proportional to the number of documents associated with it. The prior distribution also puts a probability mass on a new, $(K+1)$th latent semantic pattern, that is distinct from the patterns $(1,\ldots,K)$ observed in λ_{-n}. That is, D_α entails prior probabilities for each component of λ as follows:

$$p(\lambda_{n[k]} = 1|\lambda_{-n}) = \begin{cases} \frac{m(-n,k)}{N-1+\alpha} & \text{if } m(-n,k) > 0, \\ \frac{\alpha}{N-1+\alpha} & \text{if } k = K(-n)+1, \\ 0 & \text{otherwise,} \end{cases}$$

where λ_{-n} denotes the full-membership vectors for all but the nth document; $m(-n,k)$ is the number of documents that are associated with the kth latent pattern, other than the nth document, i.e., $m(-n,k) = \sum_{m=1}^{N} \mathbb{I}(\lambda_{m[k]} = 1, m \neq n)$; and

$K(-n)$ is the number of distinct latent patterns that are associated with at least one document other than the nth document.

The generative sampling process for the infinite mixture model is as follows:

1. Sample $\lambda \sim \text{DirichletProcess}(\alpha)$.
2. For each of the N articles
   2.1. Sample $x_{1n} \sim \text{Multinomial}(\theta_{c1}, R_1)$ where $\lambda_{n[c]} = 1$.
   2.2. Sample $x_{2n} \sim \text{Multinomial}(\theta_{c2}, R_2)$ where $\lambda_{n[c]} = 1$.

The hyper-parameters of this model are the scaling parameter of the Dirichlet process prior, α, and the multinomial parameters for words, θ_{k1}, and references, θ_{k2}, for each of the latent topics $k = 1, \ldots, K$. In this model, we assume that the number of latent topics, K, is unknown and possibly infinite, through the prior for λ, D_α, and we examine the posterior distribution of λ.

10.3.3.3 Empirical Results

We fit six models for latent topics in the PNAS dataset: using words alone or with references, finite or infinite mixture models, and (for finite mixture) fitted or fixed Dirichlet parameter α. We used variational methods for the finite mixtures, and MCMC methods for the infinite mixture. For further details see Airoldi (2007) and Joutard et al. (2007).

In Figure 10.9, we give the cross-validated log-likelihood obtained for the four finite mixture models (at $K = 5, 10, \cdots, 50, 100, 200, 300$). The plots of the log likelihood in Figure 10.9 suggest we choose a number of topics between 20 and 40 whether words or words and references are used. Values of K that maximize the held-out loglikelihood are somewhat greater when the database is expanded with references compared to when the database contains only words. Thus, adding references allows for finer refinement of topics.

The infinite model generates a posterior distribution for the number of topics, K, given the data. Figure 10.10 shows the posterior distribution ranges from 17 to 28 topics. The maximum a posteriori estimate of K is smaller for the model with words and references compared to the model with words only. Further, the posterior range of K is smaller for the model with words and references. Thus adding references to the models reduces the posterior uncertainty about K.

Overall, the values of K in the region of 15-40 are supported by all our models. A number in that range would be a plausible choice for the number of latent basis categories in PNAS biological sciences research reports, 1997-2001. By choosing $K = 20$ topics, we can meaningfully interpret all of the word and reference usage patterns. We then fit the data with a 20 topics model for the finite mixture model using words and references and focused on the interpretation of the 20 topics.

To summarize the distribution of latent aspects over distributions, we provide a graphical representation of the distribution of latent topics for each of the PNAS submission classification in Figure 10.11. When the references are included, the relationship of estimated latent categories with designated PNAS classifications be-

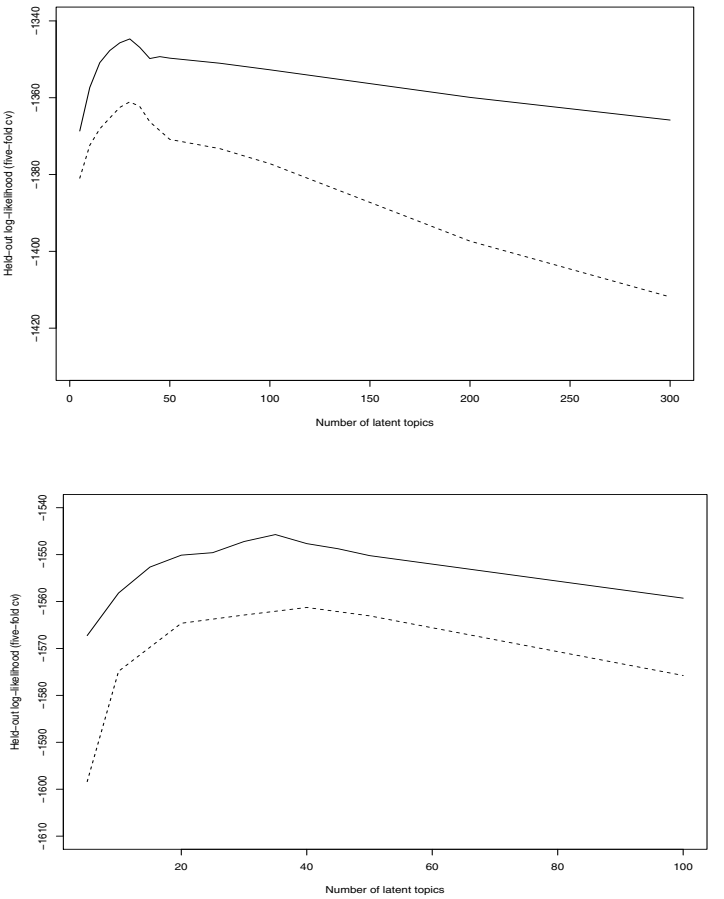


FIGURE 10.9. Top panel: Average held-out log-likelihood corresponding to four mixed-membership models we fit to the PNAS Biological Sciences articles, 1997-2001, using words from article abstracts (bottom panel) and words and references (top panel). Solid lines correspond to models fitted by estimating the hyperparameter α; dashes lines correspond to models fitted by setting the hyperparameter equal to $\alpha = 50/K$.

comes more composite for both estimation methods. Models where α is fixed are less sparse than the corresponding models with α fit to the data. For 20 latent topics, we fix $\alpha = 50/20 = 2.5 > 1$—each latent topic is expected to be present in each document and a priori we expect equal membership in each topic. By contrast the fitted values of α less than one lead to models that expect articles to have high membership in a small number of topics. See Joutard et al. (2007) for further consequences of these assumptions. The PNAS topics tend to correspond to fewer latent topics when we estimate α and to low to moderate numbers topics when we fix α.

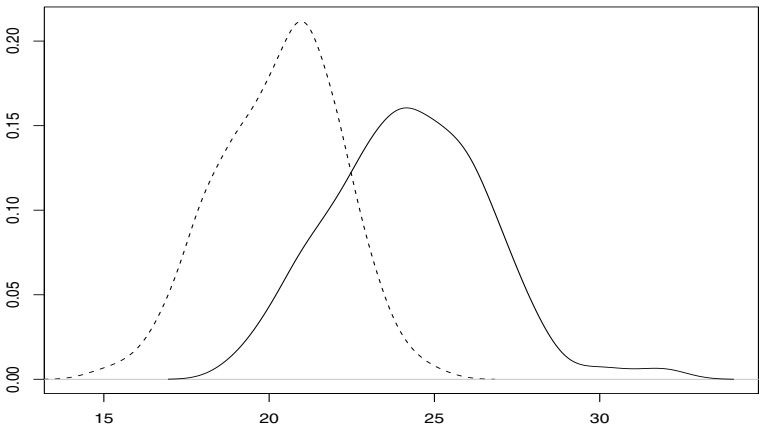


FIGURE 10.10. Posterior distribution of the number of mixture components K for infinite models for the PNAS Biological Sciences articles, 1997-2001, using words from article abstracts (solid line) and words and references (dashed line).

Further, by examining Figure 10.11, we note that nearly all of the PNAS classifications are represented by several word and reference usage patterns in all of the models. This highlights the distinction between the PNAS submission categories and the discovered latent topics. The assigned PNAS categories follow the structure of the historical development of Biological Sciences and the divisions/departmental structures of many medical schools and universities. These latent topics, however, are structured around the current biological research interests.

We consider the best model of words and references, with $K^* = 20$, and we offer the following interpretation of *all* of the topics to demonstrate what a reasonable model fit should look like in Table 10.8.

10.3.4 Case Study: Disability Profiles

All our models are special cases of HBMMMs presented in Section 10.3.1. Below, we organize them into finite and infinite mixture models, as before, according to the dimensionality of the prior distribution, D_α, posited at the latent variable level—assumption A3.

We characterize an individual by a set of responses, x_{jn} for $j = 1,\ldots,J$, which were measured through a questionnaire. In our analysis we selected $J = 16$ binary responses that encode answers to questions about the ability to perform six *activities of daily living* (ADL) and ten *instrumental activities of daily living* (IADL). The

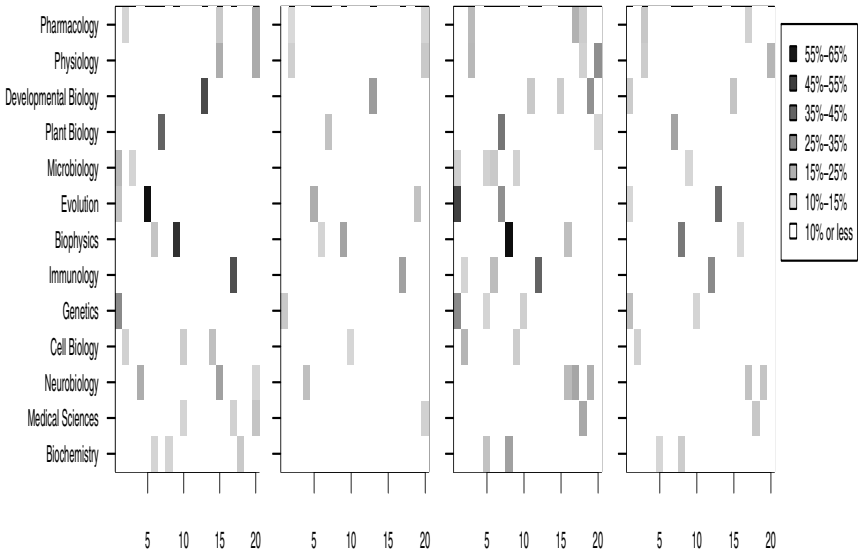


FIGURE 10.11. Estimated average mixed membership of articles in 20 estimated topics by PNAS submission classifications. In each panel, we plot average membership values for each submission category (ordered on the Y axis) in the topics (ordered on the X axis). Panels 1 and 2 represent models with only words while panels 3 and 4 use words and references. Panels 1 and 3 represent models with α estimated from the data while panels 2 and 4 use a fixed value of α.

TABLE 10.8. A post-analysis interpretation for the model with $K^* = 20$ basis catagories.

Topic	Interpretation
1	population genetics
2	enzymes by protein kinases
3	problems of hormone levels
4 & 5	nuclear activity production of cdna and mrna & catalysts for dna copying
6 & 12	HIV and immune response & T-cell response to HIV infection
7	plant evolution and phylogenetic relationships
8 & 11	protein structure and folding & protein promotion by transcription binding factors
9	procedural explanations
10	genetic mutation
14	cancer markers
13 & 18	mutant mice and tumor suppression & tumor treatment for mice and humans
15	bone marrow stem cells
16	functional and visual responses to changes in the brain
17	neurons and neurotransmitters
19	nervous system development
20	electrical excitability of cell membranes

jth response, x_{jn}, is recorded as zero if the nth individual does not have problems performing the jth activity (he is considered healthy, to that extent, for the purpose of the survey), whereas it is recorded as one if the nth individual has problems performing the jth activity (he is considered disabled, to that extent, for the purpose of the survey).

10.3.4.1 Finite Mixture Model

To carry out the analysis of the NLTCS data in the finite mixture setting we use the GoM model described in Erosheva, Fienberg, and Joutard (2007) and Joutard et al. (2007), which posits the following generative process for all N individuals in the survey.

1. Sample $\theta_{jk} \sim \text{Beta}\,(\sigma_1, \sigma_2)$ for each j and k.
2. For each of the N seniors
   2.1. Sample $\lambda_n \sim \text{Dirichlet}\,(\alpha_{[1]}, \ldots, \alpha_{[K]})$.
   2.2. Sample $x_{jn} \sim \text{Bernoulli}\,(p_{j\lambda})$ for each j, where $p_{j\lambda} = \sum_{k=1}^{K} \lambda_k \theta_{jk}$.

We sample the elements of θ from a symmetric Beta distribution with fixed hyper-parameter $\sigma_1 = \sigma_2 = 1$. Note that the distribution on λ is not the symmetric distribution we used in the previous case study, in the finite setting. In this model, θ is a matrix that encodes the probability of being disabled with respect to each one of the 16 activities for seniors who display disability characteristics specific to each of the K latent profiles. That is, θ_{jk} is the probability of being disabled with respect to the j-th activity for a person who "belongs" completely to the k-th latent profile. Note that in this model there are no constraints on the sum of the total probability of having being disabled given any specific profile. For example, $\sum_{j=1}^{J} \theta_{jk}$ is not necessarily one as in the model of Section 10.3.3. The hyper-parameters of this model are α and σ. In Joutard et al. (2007), we develop a variational approximation to perform posterior inference on such hyper-parameters, and on the latent variables λ_n.

In our analyses, we also consider a fully Bayesian version of the GoM model, following Erosheva (2002a), which posits the following generative process for all N individuals in the survey.

1. Sample $\xi \sim D_\alpha$.
2. Sample $\alpha_0 \sim \text{Gamma}\,(\tau_1, \tau_2)$.
3. Sample $\theta_{jk} \sim \text{Beta}(\sigma_1, \sigma_2)$ for each j and k.
4. For each of the N seniors
   4.1. Sample $\lambda_n \sim \text{Dirichlet}(\alpha_0 \xi_{[1]}, \ldots, \alpha_0 \xi_{[K]})$.
   4.2. Sample $x_{jn} \sim \text{Bernoulli}(p_{j\lambda})$ for each j, where $p_{j\lambda} = \sum_{k=1}^{K} \lambda_k \theta_{jk}$.

In this fully Bayesian setting we fix the hyper-parameter for convenience. According to our model specifications D_α is a symmetric Dirichlet distribution with fixed hyper-parameter $\alpha_1 = \cdots = \alpha_K = 1$. The kth component of ξ, $\xi_{[k]}$, represents the proportion of the seniors in the survey who express traits of the kth latent disability

profile. Further, we fix a diffuse Gamma distribution, $\tau_1 = 2$ and $\tau_2 = 10$, to control for the tails of the Dirichlet distribution of the mixed membership vectors, λ_n.

In both of the finite mixture models we presented in this section, we assume that the number of latent profiles is unknown but fixed at K. Our goal is to find the number of latent disability profiles, K^*, which gives the best description of the population of seniors.

10.3.4.2 Infinite Mixture Model

In the infinite setting we do not fix the number of sub-populations K. As in the previous case study, we restrict subjects (elderly Americans) to complete membership in one group (profile) and the mixed membership vectors $\lambda_{1:N}$ reduce to single membership vectors. The generative sampling process for the infinite mixture model is as follows:

1. Sample $\lambda \sim \text{DirichletProcess}(\alpha)$.
2. Sample $\theta_{jk} \sim \text{Beta}\ (\sigma_1, \sigma_2)$ for each j and k.
3. Sample $x_{jn} \sim \text{Bernoulli}(\theta_{jc})$ where $\lambda_{n[c]} = 1$ for each j and n.

Here D_α is the Dirichlet process prior described in Section 10.3.3.2. As in the finite models, we specify a symmetric Beta distribution for the disability probabilities, θ, however, here we fix $\sigma_1 = \sigma_2 = 10$ to make moderate disability probabilities more likely *a priori* than extreme probabilities. Further, we fix the hyper-parameter of the Dirichlet process prior at $\alpha = 1$, which encodes "indifference" toward additional groups.

10.3.4.3 Empirical Results

We fit three models for disability propensity profiles to the NLTCS: the finite mixture with random Dirichlet parameter α, the finite mixture with fixed but unknown α using variational and MCMC methods, and the infinite mixture model using MCMC methods. See Joutard et al. (2007), Erosheva, Fienberg, and Joutard (2007), and Airoldi (2007) for details about inference and a variety of different approaches to the choice of K, including a method based on residuals for the most frequent response patterns, and information criteria such as DIC and BIC. These analyses yield results consistent with those for cross-validation using variational approximation methods, shown in Figure 10.12, which suggest a choice of 8 or 9 profiles.

The infinite model generates the posterior distribution for the number of profiles, K, in Figure 10.12, which is concentrated on from 11 to 15 profiles. We expect that the infinite model requires more profiles because it involves "hard clustering."

Multiple criteria suggest that $K = 9$ is a reasonable choice for the NLTCS data. Figure 10.13 shows the latent profiles obtained for the 9 profiles GoM model using MCMC methods. The conditional response probabilities represented on the Y-axis are the posterior mean estimates of $\theta_{jk} = P(x_{jn} = 1 | \lambda_{n[k]} = 1)$, the probability of

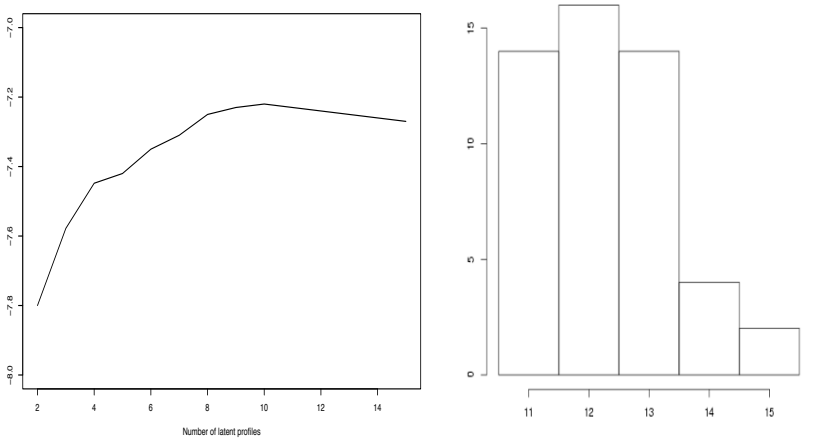


FIGURE 10.12. Left Panel: Log-likelihood (5 fold cv) for $K = 2, \ldots, 10, 15$ for the finite model. Right Panel: Posterior distribution of K for the infinite model.

being disabled on the activity j for a complete member of latent profile k. We can clearly distinguish two profiles for "healthy" individuals; these are the lower curves (the solid, black curve and the dashed, black curve). The upper curve (solid, grey curve) corresponds to seriously "disabled" individuals since most of the probabilities are greater than 0.8. One profile (long-dashed, grey curve) has the second highest values for the IADLs "managing money," "taking medicine," and "telephoning." This focuses on individuals with some cognitive impairment. The profile with the second highest probabilities for most of the ADLs/IADLs (dashed, grey curve) characterizes "semi-disabled" individuals. The profile with very high probabilities for all the activities involving mobility including the IADL "outside mobility" (dot-dashed, grey curve) characterizes mobility-impaired individuals. Another profile characterizes individuals who are relatively healthy but can't do "heavy housework" (long-dashed, black curve). The two remaining profiles (the dot-dashed, black curve and the dotted, black curve) corresponds to individuals who are "semi-healthy" since they show limitations in performing some physical activities.

We found similar interpretations with the estimates based on variational methods and MCMC methods despite some differences in the estimated values of the conditional disability propensity probabilities θ_{jk}.

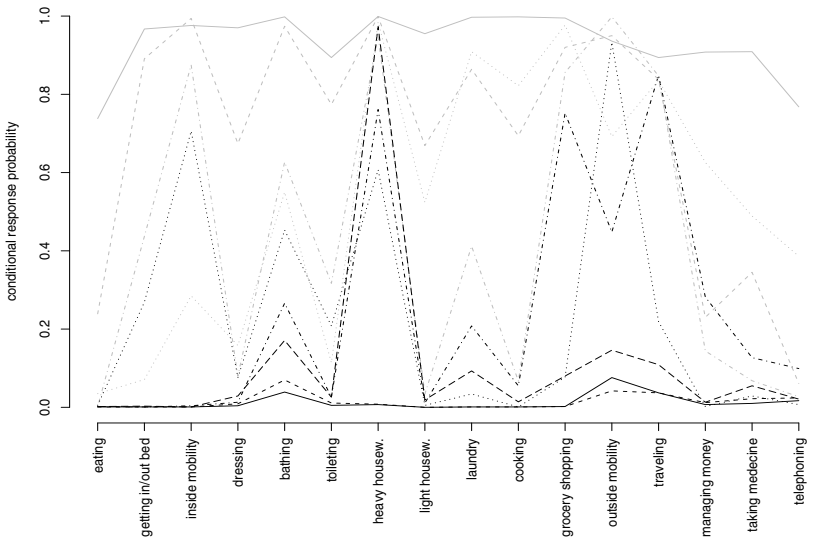


FIGURE 10.13. Latent profiles (θ_k's) for the GoM model with K=9.

10.3.5 Summary

In this section, we have studied the issue of model choice in the context of mixed-membership models. Often the number of latent classes or groups is of direct interest in applications, but it is always an important element in determining the fit and meaning of the model.

We used extensions of "latent Dirichlet allocation" (LDA) to analyze a corpus of PNAS biological sciences publications from 1997 to 2001. We included k-fold cross-validation and the Dirichlet process prior among our approaches for selecting the number of latent topics, focusing on six combinations of models and model choice strategies. We focused on $K = 20$ topics, a value that appears to be within the range of possibly optimal numbers of topics, and we saw that the resulting topics were easily interpretable and profile popular research subjects in biological sciences, in terms of the corresponding words and references usage patterns. Much higher choices for K lead to far more complex interpretations. For further details see Airoldi et al. (2009).

For the NLTCS data, we have developed parametric and nonparametric variations of the GoM model. We performed posterior inference using variational methods and MCMC. We have used different criteria to assess model fit and reached the conclusion that $K = 9$ latent profiles is an appropriate choice for the data set, cf., the related analyses reported in Erosheva, Fienberg, and Joutard (2007). This choice allows us to identify profiles such as the one for individuals who are able to perform

all activities except "doing heavy housework." Further, we were able to interpret all 9 of the profiles, although once we reach $K = 5$, the fit seems not to improve markedly.

Acknowledgments: This work was partially supported by National Institutes of Health Grant No. R01 AG023141-01, Office of Naval Research Contract No. N00014-02-1-0973, National Science Foundation Grant No. DMS-0240019, and Department of Defense Grant No. IIS0218466, all to Carnegie Mellon University, by National Institutes of Health Grant No. T32 ES007271 to University of Rochester, and by National Science Foundation Grant No. DMS-0907009 to Harvard University. We are especially grateful to Elena Erosheva for comments.

Chapter 11
Bayesian Inference in Political Science, Finance, and Marketing Research

Many current research challenges in Bayesian analysis arise in applications. A beauty of the Bayesian approach is that it facilitates principled inference in essentially any well-specified probability model or decision problem. In principle one could consider arbitrarily complicated priors, probability models and decision problems. However, not even the most creatively convoluted mind could dream up the complexities, wrinkles and complications that arise in actual applications. In this chapter we discuss typical examples of such challenges, ranging from prior constructions in political science applications, to model based data transformation for the display of multivariate marketing data, to challenging posterior simulation for state space models in finance and to expected utility maximization for portfolio selection.

11.1 Prior Distributions for Bayesian Data Analysis in Political Science

Andrew Gelman

Jim Berger has made important contributions in many areas of Bayesian statistics, most notably on the topics of statistical decision theory and prior distributions. It is the latter subject which I shall discuss here. I will focus on the applied work of my collaborators and myself, not out of any claim for its special importance but because these are the examples with which I am most familiar. A discussion of the role of the prior distribution in several applied examples will perhaps be more interesting than the alternative of surveying the gradual progress of Bayesian inference in political science (or any other specific applied field).

I will go through four examples that illustrate different sorts of prior distributions as well as my own progress—in parallel with the rest of the statistical research

M.-H. Chen et al. (eds.), *Frontiers of Statistical Decision Making and Bayesian Analysis*, DOI 10.1007/978-1-4419-6944-6_11,

community—in developing tools for including prior information in statistical analyses:

- In 1990, we fit a hierarchical model for election outcomes in congressional districts, using a mixture distribution with an informative prior distribution to model districts held by Democrats and Republicans.
- In 1994, we returned to this example, replacing the mixture model with a regression using incumbency as a predictor, with a flat prior distribution on the regression coefficient.
- In 1997, we used a hierarchical model with poststratification to estimate state-level public opinion from national polls. Formally, the model used noninformative prior distributions, but our poststratification actually used lots of external information from the Census.
- In 2008, we used a varying-intercept, varying slope model to explore the relation between income and voting in U.S. states. An attempt to extend this model to include additional predictors revealed the limitations of our default approach of marginal maximum likelihood.

11.1.1 Statistics in Political Science

Is there anything about the study of public opinion and politics (as compared to economics, psychology, sociology, or history, for example) that would show up in the statistical modeling, in particular in prior distributions? I don't think so.

Important statistical issues do arise in particular examples, however. For example, there have not been many national elections, but the fifty states are a natural setting for hierarchical modeling—the states are hardly exchangeable but it can be reasonable to model them with exchangeable errors after controlling for regional indicators and other state-level predictors[1]. Much work in political science goes into increasing the sample size, for example studying other countries (or, within the United States, by studying state and local elections) or replacing binary data with continuous variables. For example, students of the so-called "democratic peace" use continuous measures for democracy and peace, allowing quantitative researchers to examine more sophisticated hypotheses (see Garktze, 2007).

I now return to the statistical specifics of the examples listed above. As we shall see, our models do not show any linear or even monotonic development. Rather,

[1] I used to say that Alabama and Mississippi were exchangeable, along with North and South Dakota, until Brad Carlin—a resident of the neighboring state of Minnesota—explained to me the differences between these two sparsely populated northern states, thus also educated me in the general principle, emphasized by Bayesians from De Finetti to Berger and beyond, that exchangeability is a state of mind as much as it is a description of the physical and social world. To this day I remain blissfully ignorant of any important features distinguishing the two southern states mentioned above. Not so many years ago many would've considered New Hampshire and Vermont to be exchangeable as well, but the expanding Boston suburbs on one side and Ben & Jerry's on the other have made such a model untenable.

we have used more informative prior distributions where needed because of data limitations.

11.1.2 Mixture Models and Different Ways of Encoding Prior Information

Gelman and King (1990) present a model for estimating the so-called seats-votes curve: the expected percentage of seats won by a political party in a legislative election, as a function of the party's share of the national vote. For example, in 2008 the Democrats won 59% of the seats in the U.S. House of Representatives based on an average of 55% of the vote in House elections (after adjusting for uncontested seats; see Kastellec, Gelman, and Chandler, 2009). In 2006 they garnered 53% of the seats based on 52% of the vote. More generally, we can estimate a stochastic seats-votes relation-and thus compute its expectation, the seats-votes curve—by setting up a probability model for the vector of 435 congressional election outcomes.

For a Bayesian such as Jim Berger (or myself), inference under a probability model is conceptually straightforward (even though it might require computational effort and even research). The real challenge is setting up the model.

To use statistical notation, we have districts $i = 1, 2, \ldots, 435$, and in each there is y_i, the proportion of votes received in that district by the Democrats in the most recent election (as noted above, our model corrects for uncontested races, a detail which we ignore in our treatment here). We model

$$y_i \sim N(\theta_i, \sigma_y^2),$$

where θ_i represents the expected level of support for the Democrats in that district and year, with σ_y representing variation specific to that election. We estimated σ_y by looking at the residual variance predicting an election from the election six years ago, four years ago, and two years ago, and extrapolating this down to predict a hypothetical variance at lag zero.

With one data point y_i for each parameter θ_i, we certainly needed a prior distribution, and what we used was a mixture model with three components: two major modes roughly corresponding to Democratic and Republican-leaning districts, and a third component with a higher variance to capture districts that did not fit in either of the two main modes. This mixture of three normal distributions had eight hyperparameters, which we gave pretty strong prior distributions in order to separately identify the modes from a single election's worth of data.

Much has been written about the difficulty of estimating mixture models and the failure of maximum likelihood or noninformative Bayesian inference in this setting; here, we had to go even further because our mixture components had particular interpretations that we did not want to lose. To be specific, we assigned following informative prior distribution:

- Mixture component 1: mean had a $N(-0.4, 0.4^2)$ prior distribution, standard deviation had an inverse-$\chi^2(4, 0.4^2)$ prior distribution;
- Mixture component 2: mean had a $N(+0.4, 0.4^2)$ prior distribution, standard deviation had an inverse-$\chi^2(4, 0.4^2)$ prior distribution;
- Mixture component 3: mean had a $N(0, 3^2)$ prior distribution, standard deviation had an inverse-$\chi^2(4, 0.8^2)$ prior distribution; and
- The three mixture parameters had a Dirichlet$(19, 19, 4)$ prior distribution.

(The model was on the logit scale, which was why the priors for the modes were centered at -0.4, $+0.4$ rather than at 0.4, 0.6 as they would have been had the data been untransformed.)

Finally, having performed inference for the model using the Gibbs sampler (or, as we called it in those pre-1990 days, "the data augmentation method of Tanner and Wong (1987)"), we can simulate hypothetical replications of the election under different conditions and then map out a seats-votes curve by allowing different nationwide vote swings.

We followed up this study a few years later (Gelman and King, 1994) with a very similar model differing in only two particulars: First, we set up our model as a regression in which for each data point y_i there could be district-level predictors x_i, and as a predictor we took incumbency status: a variable that equaled 1 for Democratic congress members running for reelection, -1 for Republican incumbents, and 0 for "open seats"—those districts with no incumbents running. The model was essentially the same as before, except that the district-level variance represented unexplained variation after accounting for this (and any other) predictors.

The other way in our 1994 model differed from that published four years earlier was that we got rid of the mixture model and its associated informative prior distribution! It turned out that all the information captured therein—and more—was contained in the incumbency predictor. This illustrates the general point that what is important is the information, not whether it is in the form of a "prior distribution" or a "likelihood[2]."

11.1.3 Incorporating Extra Information Using Poststratification

In the wake of the successes of hierarchical Bayes for agricultural, social, and educational research (see, for example, Lindley and Smith (1972), and the accompanying references and discussion), survey researchers began using these methods for small-area estimation (Fay and Herriot, 1979).

Gelman and Little (1997) applied these models to the problem of estimating state-level opinions from national surveys, using hierarchical logistic regression to obtain estimates of the average survey response within population subgroups defined by sex, ethnicity (2 categories), age (4 categories), education (4 categories), and

[2] Contrary to what Bayesians sometimes say, however, neither a loss function nor any formal decision analytic framework was needed to set up the model and use it to perform useful inferences.

state (51, including the District of Columbia)—3264 cells in all, and thus certainly a case of small-area estimation—and then summing those estimates over the 64 cells within each state to estimate state-level averages.

Looked at in the traditional Bayesian way, the regression model was innocuous, with predictors including sex×ethnicity, age×education, and state indicators, fitted normal prior distributions for the 16 age×education, and a group-level regression with normal errors for the 51 state predictors. The unmodeled coefficients and hyperparameters were given noninformative uniform prior distributions, and it was easy enough to program a Metropolis algorithm that converged well and yielded simulation-based inference for all the regression parameters, simulations that we directly propagated to obtain inference for the 3264 population cells—a nice trick, given that the procedure performs well even when fit to samples of 1500 or less.

A key place where external information enters into this example, though, is in the next step, in which we construct inferences for the 51 states. The key step is poststratification: summing over the cells in proportion to their population sizes within each state. This step is not particularly Bayesian—given the computations already done, it's nothing more than the computation of 51 weighted averages for each of our posterior simulations—but it does use prior information, in this case the population counts from the Census. The poststratification framework allows us to include external information structurally, as it were, in a way more natural than would be the formal elicitation of a prior distribution.

This multilevel regression and poststratification approach has been useful in other studies of public opinion. For example, Lax and Phillips (2009a) estimate state-level opinion on several gay-rights issues and compare to state policies in this area. Lax and Phillips (2009b) demonstrate that this approach outperforms classical methods while using far smaller samples. The formal prior distribution is not important here, but what is crucial is the use of prior information in the form of state-level predictors (along with the external information from the Census, which is also implicitly used in survey weighting).

11.1.4 Prior Distributions for Varying-Intercept, Varying-Slope Multilevel Regressions

A striking feature of the American political map in the twenty-first century is that the Democratic Party does best in the richer states of the northeast and west coast, while the Republicans' strength is in the poorer states in the south and middle of the country—even while the parties retain their traditional economic bases, with Democrats and Republicans continuing to win the votes of poorer and richer voters, respectively. Gelman et al. (2008b) use Bayesian multilevel modeling to explore this juxtaposition, using both individual-level and state-level incomes to predict vote choice in a logistic regression model that includes unexplained variation at both levels. The coefficients for state and individual incomes go in opposite directions,

corresponding to rich Democratic states with rich Republican voters within each state.

Our central model included varying intercepts and slopes—that is, the relation between income and voting was allowed to be different in each state—and we blithely fit it using noninformative uniform prior distributions for the hyperparameters, which for this model included the unmodeled regression coefficients, the group-level standard deviation parameters, and the correlation between the errors in the state-level intercepts and slopes. All worked well, and we had the agreeable choice of fitting the full Bayesian model in Bugs (Spiegelhalter et al., 1994, 2002) or running a quick approximate fit using a program in R that computed marginal maximum likelihood estimates (Bates, 2005).

But we ran into trouble when we tried to extend the model by adding religious attendance as a predictor (Gelman et al., 2008a, Chapter 6), thus requiring four varying coefficients per state (income, religious attendance, their interaction, and a constant term). A group-level covariance of dimension 4×4 was just too much for a noninformative prior distribution to handle. Bugs simply choked—the program ran extremely slowly and failed to move well through the posterior distribution—and the marginal maximum likelihood estimate moved straight to the boundary of parameter space, yielding an estimated covariance matrix that was not positive definite. These problems arose even with sample sizes in the tens of thousands; apparently, the hyperparameters of even moderately-dimensional hierarchical regression models are not well identified from data.

In our particular example of modeling vote choice given income and religious attendance, we managed to work around the problem by accepting this flawed estimate—our focus here was on the four coefficients for each state rather than on the hyperparameters themselves—but we are convinced that a good general solution to this problem requires an informative prior distribution for the group-level covariance matrix, possibly using the scaled-inverse-Wishart family (O'Malley and Zaslavsky, 2005), whose redundant parameterization allows the user to supply different prior precisions for scale and correlation parameters.

11.1.5 Summary

In conclusion, prior information is often what makes Bayesian inference work. I won't say it's always necessary–noninformative machine learning methods seem to work pretty well in classification problems with huge sample sizes and simple questions–but in the political science examples of which I'm aware, information needs to come in, whether as regression predictors or regularization (that is, prior distributions) on parameters. An important challenge for Jim Berger and his successors in the theory of Bayesian statistics is to study the mapping from prior to posterior in indirect-data settings such as hierarchical models, and thus to figure out which aspects of the prior distribution we need to be particularly careful to specify well. Such theory may indirectly inform our understanding of public opinion,

elections, and international relations, by enabling us to study social and political phenomena with ever more realistic (and thus complicated and parameter-laden) models.

Acknowledgments: We thank the National Science Foundation, National Institutes of Health, Institute of Educational Sciences, and the Columbia University Applied Statistics Center for partial support of this work.

11.2 Bayesian Computation in Finance

Satadru Hore, Michael Johannes, Hedibert Lopes, Robert E. McCulloch, and Nicholas G. Polson

Modern-day finance uses arbitrage and equilibrium arguments to derive asset prices as a function of state variables and parameters of the underlying dynamics of the economy. Many applications require extracting information from asset returns and derivative prices such as options or to understand macro-finance models such as consumption-based asset pricing models. To do this the researcher needs to combine information from different sources, asset returns on the one hand and derivative prices on the other. A natural approach to provide inference is Bayesian (Berger, 1985; Bernardo and Smith, 1994; Gamerman and Lopes, 2007).

Our computational challenges arise from the inherent nonlinearities that arise in the pricing equation, in particular through the dependence on parameters. Duffie (1996) and Johannes and Polson (2009) show that empirical asset pricing problems can be viewed as a nonlinear state space models. These so-called affine models provide a natural framework for addressing the problem as well. Whilst affine pricing models in continuous time go a long way to describe the evolution of derivative prices, empirically extracting the latent state variables and parameters that drive prices has up until now received less attention due to computational challenges. In this paper, we address these challenges by using simulation-based methods, such as Markov chain Monte Carlo (MCMC), Forward filtering backward sampling (FFBS) and particle filter (PF). Hence we solve the inverse problem of filtering state variables and estimating parameters given empirical realizations on returns and derivative prices.

The statistical tools that we describe include MCMC methods, with particular emphasis on the FFBS algorithm of Carter and Kohn (1994) and Frühwirth-Schnatter (1994). For sequential methods we describe PF algorithms, with particular emphasis on the sequential importance sampling with resampling (SISR) filter of Gordon, Salmond, and Smith (1993) and the particle learning (PL) algorithm of Lopes et al. (2010). This current research shows how to also estimate parameters such as agents' preferences from empirical data. In many cases the agents will be given the underlying parameters and the problem becomes one of filtering the hidden states as conditioning information arrives.

The rest of the section is outlined as follows. Section 11.2.1 describes asset pricing problems and Bayesian inference for these problems. Section 11.2.2 describes sequential Bayesian computation. Section 11.2.3 provides an illustration to equilibrium-based stochastic volatility models. Finally, Section 11.2.4 concludes.

11.2.1 Empirical Bayesian Asset Pricing

In order to solve the inference problem, namely calculating the joint posterior $p(\theta, X_t|Y^t)$ where $Y^t = (Y_1, \ldots, Y_t)$ is a set of discretely observed returns, we have to pick a suitable time-discretization of the continuous-time model and then perform Bayesian inference on the ensuing state space model. We use an Euler discretization.

In order to make our discretization an accurate representation of the continuous time model, we may discretize at a higher frequency than that of the observed data. In this case, we simulate additional state variables between the observations via the Euler discretization scheme. This introduces the concept of missing data. These missing data are drawn via a Gibbs step.

One novel feature of Bayesian methods is that they allow data in the form of observations of derivative prices to aid in the estimation problem. For example, suppose that we observe a call price $C(X_t, \theta)$ or a variance swap price. This information can be combined with the current posterior distribution on states and price, namely $p(X_T, \theta|Y^T)$, to obtain sharper parameter estimates.

We now provide the relevant Bayesian calculations. The conditional likelihood can be written as

$$p(R_{t+1}, P_{t+1}|X_{t+1}, \theta) = p(R_{t+1}|X_{t+1})p(P_{t+1}|X_{t+1}, \theta),$$
$$p(P_{t+1}|X_{t+1}, \theta) \sim N(F(X_t, \theta), \sigma_D^2),$$

where σ_D is a pricing error of say $(1,5)\%$ and $F(X_t, \theta)$ describes the pricing formula obtained from equilibrium arguments. In the discretized system, log-returns are given by $R_{t+1} = Y_{t+1} - Y_t = \ln(S_{t+1}/S_t)$ and $p(R_{t+1}|X_{t+1})$ describes the evolution of asset returns.

The goal of empirical asset pricing is to learn about the risk neutral and objective parameters $(\theta^{\mathrm{Q}}, \theta^{\mathrm{P}})$, respectively. Moreover, one can also recover filtered estimates of the state variables X_{t+1}, namely volatility, V_{t+1}, jump times, J_{t+1} and jump sizes, and the model specification from the observed asset returns and derivative prices. In general, we have a joint system with data $Y_{t+1} = (R_{t+1}, P_{t+1})$ corresponding to returns and derivative prices evolving according to the system

$$\begin{pmatrix} P_{t+1} \\ R_{t+1} \end{pmatrix} = \begin{pmatrix} A(\theta^{\mathrm{P}}, \theta^{\mathrm{Q}})X_{t+1} + B(\theta^{\mathrm{P}}, \theta^{\mathrm{Q}}) \\ \mu^{\mathrm{P}} \end{pmatrix} + \begin{pmatrix} \sigma_D \varepsilon_{t+1}^D \\ \sqrt{V_{t+1}}\varepsilon_{t+1}^R \end{pmatrix}.$$

Here σ_D is a pricing error and there exist pricing formulas for $A(\theta^{\mathrm{P}}, \theta^{\mathrm{Q}})$ and $B(\theta^{\mathrm{P}}, \theta^{\mathrm{Q}})$. See Chernov and Ghysels (2000) and Polson and Stroud (2003) for fur-

ther discussion. Traditional pure inversion methods infer parameters by first taking the derivative price as given and then trying to match state and parameter values to the given price, P_t, by inversion, namely $(\hat{X}_t, \hat{\theta}) = C^{-1}(P_t)$. However, taking the time series of implied states $\hat{X}_t$ can also lead to parameter estimates that are inconsistent.

Posterior distribution. The Bayesian posterior distribution now uses both returns and derivative pricing information to estimate $(\theta^{\mathrm{P}}, X_t)$ and implicitly determine θ^{Q}. This is given by

$$p(\theta^{\mathrm{P}}, \theta^{\mathrm{Q}}, X_t | Y^t) \propto p(R^t | \theta^{\mathrm{P}}, X_t) p(P^t | \theta^{\mathrm{P}}, \theta^{\mathrm{Q}}, X_t) p(X_t | \theta^{\mathrm{P}}) p(\theta^{\mathrm{P}}, \theta^{\mathrm{Q}}),$$

where $Y_t = (R_t, P_t)$ contains returns and prices from the equilibrium model.

Hence we can sequentially filter $p(\theta^{\mathrm{P}}, \theta^{\mathrm{Q}}, X_{t+1} | Y^{t+1})$. For very small pricing errors the derivative price information will give very precise estimates of X_{t+1} which in turn will precisely estimate the parameters of the system. However, too restrictive will lead to noisy co-variance estimates and the derivative prices will be in conflict with the physical model evolution which may be a sign of model misspecification.

11.2.2 Bayesian Inference via SMC

Here we review the particle methods that are developed for state filtering and sequential parameter learning with and without derivative price information. Let X_t denote a latent state variable, θ underlying parameters and observed data Y_t at time t. This might just be returns R_t or a combination of returns and derivative prices (R_t, D_t). Then, the models considered in the paper are instances of the following general state-space model

$$\begin{aligned} (Y_t | X_{t-1}, \theta) &\sim p(Y_t | X_{t-1}, \theta), \\ (X_t | X_{t-1}, \theta) &\sim p(X_t | X_{t-1}, \theta) \end{aligned}$$

for $t = 1, \ldots, T$ and $X_0 \sim p(X_0 | \theta)$. The parameter θ is kept fixed (and omitted) for the moment. There are three filtering and learning posterior distributions:

1. *Filtering:* computation of or sampling from $p(X_t | Y^t)$ on-line for $t = 1, \ldots, T$;
2. *Filtering and learning:* computation of and/or sample from the posterior $p(\theta, X_t | Y^t)$ from which we can calculate marginals $p(X_t | Y^t)$ and $p(\theta | Y^t)$ for $t = 1, \ldots, T$; and
3. *Smoothing:* computation of or sampling from the full joint distribution $p(\theta, X^T | Y^T)$.

Predictive distributions are also straightforward to compute as forward functionals of the process taking into account this posterior uncertainty about parameters. The optimal Bayesian nonlinear filters (under squared error loss) are $\hat{X}_t = \mathbb{E}(X_t | Y^t)$

and $\hat{\theta} = \mathbb{E}(\theta|Y^t)$. Given these filtered posterior estimates of the states and the parameters we can then provide estimates of the following dynamics: i) *physical and risk-neutral:* $\mathbb{P}$ and $\mathbb{Q}$ dynamics; ii) *market price of risk:* given a specification of market prices of risk we can calculate the posterior $p(\lambda|Y^T)$ given a panel of derivative prices (calls, variance swaps).

Particle filters. Unfortunately, closed-form solution for the filtering and smoothing distributions is only available for simple cases. For the general state-space model, the filtered distributions are temporally connected via the following recursive propagation/update equations:

$$p(X_{t+1}|Y^t) = \int p(X_{t+1}|X_t)p(X_t|Y^t)dX_t, \tag{11.2.1}$$

$$p(X_{t+1}|Y^{t+1}) \propto p(Y_{t+1})p(X_{t+1}|Y^t), \tag{11.2.2}$$

i.e., the propagation rule emulates the prior distribution of X_{t+1} (a high dimensional integral) to be combined with the likelihood $p(X_{t+1}|Y^{t+1})$ via Bayes' theorem (also a function of a high-dimensional integral).

We use particle filter algorithms to sequentially update the particle set $\{X_t^{(1)}, \ldots, X_t^{(N)}\}$ to the particle set $\{X_{t+1}^{(1)}, \ldots, X_{t+1}^{(N)}\}$ once Y_{t+1} become available. Particle filters provide a natural alternative to MCMC methods which are computationally intensive for the sequential inference problem (see Johannes, Polson, and Stroud, 2009). The posterior distribution $p(X_t|Y^t)$ for the filtering and learning distribution on states and parameters is approximated by

$$p^N(X_t|Y^t) = \frac{1}{N}\sum_{i=1}^{N} \delta_{X_t^{(i)}}(X_t)$$

based on the particle set $\{X_t^{(i)}, i = 1, \ldots, N\}$, where $\delta_z(\cdot)$ denotes the delta-Dirac mass located in z. We need to show how to efficiently update particles. More precisely, after observing Y_{t+1}, we need to efficiently produce a new particle set $\{X_{t+1}^{(i)}, i = 1, \ldots, N\}$ that approximates $p(X_{t+1}|Y^{t+1})$. We review the seminal sample-importance resample or *bootstrap filter* of Gordon, Salmond, and Smith (1993) and the particle learning schemes of Lopes et al. (2010). A recent and thorough literature review of existing methods and recent advances in sequential Monte Carlo is provided by Cappé, Godsill, and Moulines (2007).

11.2.2.1 The Sample-Importance Resample (SIR) Filter

The classical bootstrap filter is also well known as the sequential importance sample with resampling (SISR) filter, and can be described in two steps mimicking the above propagation/update rule ((11.2.1) and (11.2.2)).

Let the particle set $\{X_t^{(i)}, i = 1, \ldots, N\}$ approximate $p(X_t|Y^t)$.

1. *Propagate:* Draw $\tilde{X}_{t+1}^{(i)} \sim p(X_{t+1}|X_t^{(i)})$ and compute associated (unnormalized) weights $w_{t+1}^{(i)} \propto p(Y_{t+1}|X_{t+1}^{(i)})$, for $i = 1,\ldots,N$;
2. *Resample:* $X_{t+1}^{(i)} = \tilde{X}_{t+1}^{(k^i)}$, where $k^i \sim \text{Multinomial}(w_{t+1}^{(1)},\ldots,w_{t+1}^{(N)})$, for $i = 1,\ldots,N$.

Now, the particle set $\{X_{t+1}^{(i)}, i = 1,\ldots,N\}$ approximate $p(X_{t+1}|Y^{t+1})$. In words, the propagation step generates particles that approximate the prior distribution at time $t+1$, i.e. $p(X_{t+1}|Y^t)$ (11.2.1). Then, a simple SIR argument (reweighing prior draws with their likelihoods) transforms prior particles into posterior particles that approximate $p(X_{t+1}|Y^{t+1})$ (11.2.2).

Despite (or due to) its attractive simplicity and generality, the SISR algorithm suffers from *particle degeneracy* and is bounded to break down after a few hundred observations, even in the simplest scenarios, such as the local level model. Additionally, SIR filters tend to become even more unstable when sequential parameter learning is dealt with. We propose below a particle filter that overcomes these obstacles in a large class of dynamic models.

11.2.2.2 Particle Learning

Bayes' rule links these to the next filtering distribution through Kalman-type updating. This takes the form of a smoothing and a prediction step that reverses the standard order of the propagation/update rule of (11.2.1) and (11.2.2). More precisely,

$$\begin{aligned} p(X_t|Y^{t+1},\theta) &\propto p(Y_{t+1}|X_t,\theta)p(X_t|Y^t,\theta), \\ p(X_{t+1}|Y^{t+1},\theta) &= \int p(X_{t+1}|X_t,\theta)p(X_t|Y^{t+1},\theta)dX_t, \\ p(\theta|X^{t+1},Y^{t+1},\theta) &= p(\theta|Z_{t+1}), \end{aligned}$$

where $Z_{t+1} = \mathscr{Z}(Z_t, X_{t+1}, Y_{t+1})$ is a vector of conditional sufficient statistics for θ. This leads us to the following particle simulation algorithm. As before, let the particle set $\{(X_t, Z_t, \theta)^{(i)}, i = 1,\ldots,N\}$ approximate $p(X_t, \theta, Z_t|Y^t)$.

1. *Resample:* $(\tilde{X}_t, \tilde{Z}_t, \tilde{\theta})^{(i)} = (X_t, Z_t, \theta)^{(k^i)}$, where $k^i \sim \text{Multinomial}(w_t^{(1)},\ldots,w_t^{(N)})$ and (unnormalized) weights $w_t^{(j)} \propto p(Y_{t+1}|X_t^{(j)}, \theta^{(j)})$, for $j = 1,\ldots,N$;
2. *Propagate:* Draw $X_{t+1}^{(i)}$ from $p(X_{t+1}|\tilde{X}_t^{(i)}, \tilde{\theta}^{(i)}, Y_{t+1})$, for $i = 1,\ldots,N$;
3. *Update sufficient statistics:* $Z_{t+1}^{(i)} = \mathscr{Z}(\tilde{Z}_t^{(i)}, X_{t+1}^{(i)}, Y_{t+1})$;
4. *Parameter learning:* $\theta|Z_{t+1}^{(i)} \sim p(\theta|Z_{t+1}^{(i)})$.

Central to PL algorithms is the possibility of directly sampling from the joint posterior distribution of state (augmented or not) and parameter conditional sufficient statistics. There are a number of advantages to using particle learning: (i) optimal filtering distributions in sequential parameter learning cases (Pitt and Shephard, 1999); (ii) parameter sufficient statistics for Gaussian and conditionally Gaussian

state-space models (Storvik, 2002); (iii) straightforward particle smoothing (Godsill, Doucet, and West, 2004); (iv) state sufficient statistics reduces Monte Carlo error (Chen and Liu, 2000); and (v) alternative to standard MCMC methods in state-space models (Carvalho et al., 2009). Lopes et al. (2010) introduces PL as a framework for (sequential) posterior inference for a large class of dynamic and static models.

11.2.2.3 The 2007-2008 Credit Crisis: Extracting Volatility and Jumps

Lopes and Polson (2010) used particle filtering methods to estimate volatility and examine volatility dynamics for three financial time series (S&P500, NDX100 and XLF) during the early part of the credit crisis. Standard and Poor's SP500 stock index and the Nasdaq NDX 100 index are well known. The XLF index is an equity index for the prices for US financial firms. They compared pure stochastic volatility models to stochastic volatility models with jumps. More specifically, the stochastic volatility jump model includes the possibility of jumps to asset prices and one possible model is

$$\begin{aligned} \frac{dS_t}{S_t} &= \mu + \sqrt{V_t} dB_t^{\mathbb{P}} + d\left(\sum_{s=N_t}^{N_{t+1}} Z_s\right), \\ d\log V_t &= \kappa_v(\theta_v - \log V_t) + \sigma_v dB_t^V, \end{aligned}$$

where the additional term in the equity price evolution describes the jump process and is absent in the pure stochastic volatility model. The parameter μ is an expected rate of return and the parameters governing the volatility evolution are κ_v, θ_v and σ_v. The Brownian motions $(B_t^{\mathbb{P}}, B_t^V)$ are possibly correlated giving rise to a leverage effect. The probabilistic evolution $\mathbb{P}$ describes what is known as the physical dynamics as opposed to the risk-neutral dynamics $\mathbb{Q}$ which is used for pricing. Sequential model choice shows how the evidence in support of the stochastic volatility jump model accumulates over time as market turbulence increases (Figure 11.1).

11.2.3 Bayesian Inference via MCMC

In this section we outline a simple and robust approach to the estimation of nonlinear state-space models derived from discretized asset pricing models. We illustrate the approach on asset pricing derived from the equilibrium conditions of a fully specified economic model. Our approach is to discretize the states and then apply the FFBS sampling algorithm (forward-filtering, backward-sampling, Frühwirth-Schnatter, 1994; Carter and Kohn, 1994) as an alternative to the general MCMC scheme of Carlin, Polson, and Stoffer (1992).

FIGURE 11.1. Sequential (log) Bayes factor for the SV with jumps model versus the pure SV model for the year 2007.

Our exposition and examples will assume a one-dimensional state space model. In higher dimensional cases, we can update the components of the state one at a time in a Gibbs sampler. The parameter θ is kept fixed during the FFBS step and is usually sampled jointly. Hore, Lopes, and McCulloch (2009), for instance, used a mixture of various Metropolis-Hastings chains to sample from $p(\theta|X^T, Y^T)$.

The rest of this section is largely based on Hore, Lopes, and McCulloch (2009). They derive option prices in a general equilibrium setting under recursive preferences and time-varying growth rates. The key distinction from the option prices from the previous section is that the market prices of risks are determined endogenously from the solution of the agent's utility maximization problem under recursive preferences and stochastic growth rates. This highlights the strength of our estimation procedure. We can apply our Bayesian methodology to estimate deep parameters of a dynamic general equilibrium model and obtain full inference on the underlying states (and parameters as well) given the data on prices (or return) and/or quantity dynamics that are implied by the economic system. It takes a valuable step forward in giving us inference on economic parameters that guide us to understand dynamic rational expectations models that form the building block of structural asset pricing problems.

11.2.3.1 Forward Filtering, Backward Sampling

A general, nonlinear state-space model is specified by the distributions

$$p(X_0),\ \ p(X_t \,|\, X_{t-1}),\ \ p(Y_t \,|\, X_t, X_{t-1}).$$

The first two distributions determine the distribution of the latent states $\{X_t\}$ and the third distribution gives the distribution of the observable $\{Y_t\}$ given the current and previous states. Often the observation equation is given as $p(Y_t \,|\, X_t)$ so that the current observation only depends on the current state. We will need the generalization given above for our financial application.

Given the three distributions above, the joint distribution of (X_0, X^T, Y^T), where $Z^t = (Z_1, \ldots, Z_t)$, is given by

$$p(X_0, X^T, Y^T) = p(X_0) \prod_{t=1}^{T} p(X_t \mid X_{t-1}) \prod_{t=1}^{T} p(Y_t \mid X_t, X_{t-1}). \tag{11.2.3}$$

We now review the basic steps involved in FFBS. After we have finished our review we will discuss the attractive simplicity of the state discretization strategy. FFBS consists of first forward filtering (FF) and then backward sampling (BS). The forward filtering step recursively updates

$$p(X_{t-1}, X_t \,|\, Y^t) \quad \Rightarrow \quad p(X_t, X_{t+1} \,|\, Y^{t+1}).$$

The backward sampling draws from the joint distribution of the states given the data using

$$p(X_1, X_2, \ldots, X_T \,|\, Y^T) = p(X_T, X_{T-1} \,|\, Y^T) \prod_{t=T-2}^{1} p(X_t \,|\, X_{t+1}, X_{t+2}, \ldots, X_T, Y^T).$$

That is, we first draw the last two states given all the data and then work backwards in time drawing each state conditional on all the subsequent ones.

Forward Filtering

We do the forward filtering (FF) iteration in two steps: an evolution step and an update step.

FF Evolution. In the first step, we extend our state knowledge at time t to include the future state using the state equation:

$$\begin{aligned} p(X_{t-1}, X_t, X_{t+1} \mid Y^t) &= p(X_{t-1}, X_t \,|\, Y^t)\ p(X_{t+1} \mid X_t, X_{t-1}, Y^t) \\ &= p(X_{t-1}, X_t \,|\, Y^t)\ p(X_{t+1} \mid X_t). \end{aligned}$$

The first term on the right hand side is what we assume we know from the previous iteration, i.e., the time t posterior distribution of (X_{t-1}, X_t), while the second term is the state equation of the state-space model. We then margin out X_{t-1} to obtain $p(X_t, X_{t+1} \mid Y^t)$:

$$p(X_{t-1}, X_t, X_{t+1} \mid Y^t) \Rightarrow p(X_t, X_{t+1} \mid Y^t).$$

Alternatively, we can first margin out X_{t-1} from $p(X_{t-1}, X_t \mid Y^t)$ and then use the state equation:

$$p(X_t, X_{t+1} \mid Y^t) = p(X_t \mid Y^t)\, p(X_{t+1} \mid X_t).$$

This first step may be viewed as computing our prior knowledge of (X_t, X_{t+1}) given Y^t, our observations up to time t.

FF Update. In the second step, we implement the Bayes' theorem to update our distribution of (X_t, X_{t+1}) to incorporate the additional information in Y_{t+1}. Keeping in mind that $Y^{t+1} = (Y^t, Y_{t+1})$,

$$\begin{aligned} p(X_{t+1}, X_t \mid Y^{t+1}) &\propto p(X_{t+1}, X_t, Y_{t+1} \mid Y^t) \\ &= p(X_t, X_{t+1} \mid Y^t)\, p(Y_{t+1} \mid X_t, X_{t+1}, Y^t) \\ &= p(X_t, X_{t+1} \mid Y^t)\, p(Y_{t+1} \mid X_t, X_{t+1}). \end{aligned}$$

The first term on the right hand side is available from the above FF evolution step, while the second term is the observation equation of the state-space model. The FF step is really just the Bayes theorem repeated over time.

Backward Sampling

The backward sampling step depends on the observation that

$$\begin{aligned} p(X_t \mid X_{t+1}, X_{t+2}, \ldots, X_T, Y^T) &= p(X_t \mid Y^t, X_{t+1}, Y_{t+1}) \\ &= p(X_t \mid X_{t+1}, Y^{t+1}). \end{aligned} \tag{11.2.4}$$

Let $\tilde{Z}^t = (Z_t, Z_{t+1}, \ldots, Z_T)$. So superscript t means everything up to and including t and superscript combined with a $\sim$ on top means from t on including t. Using this notation (11.2.4) becomes

$$p(X_t \mid X_{t+1}, \tilde{X}^{t+2}, Y^t, Y_{t+1}, \tilde{Y}^{t+2}) = p(X_t \mid X_{t+1}, Y^t, Y_{t+1}).$$

Stated in terms of conditional independence, this is equivalent to say that

$$X_t \perp (\tilde{X}^{t+2}, \tilde{Y}^{t+2}) \mid X_{t+1}, Y^t, Y_{t+1},$$

where $\perp$ indicates independence. If the state space model is written as a DAG (directed acyclic graph) and then converted to an undirected graph, this conditional independence statement is obvious. We can also see the property given by (11.2.4) directly from the joint distribution of (X^T, Y^T) by first noting that

$$p(X^{t-1}, X_t, Y^t \mid X_{t+1}, Y_{t+1}, \tilde{X}^{t+2}, \tilde{Y}^{t+2}) = p(X^{t-1}, X_t, Y^t \mid X_{t+1}, Y_{t+1}). \tag{11.2.5}$$

Since the seven quantities in both sides of the above comprise all of (X^T, Y^T), this relation may be easily seen simply by looking at the full joint given in (11.2.3) and remembering that the conditional is proportional to the joint. That is, we can rewrite (11.2.3) as

$$p(X_0, X^T, Y^T) = p(X_0)\, p(X^{t-1} \mid X_0)\, p(X_t \mid X_{t-1})\, p(X_{t+1} \mid X_t)\, p(\tilde{X}^{t+2} \mid X_{t+1}) \times p(Y^t \mid X^t)\, p(Y_{t+1} \mid X_{t+1}, X_t)\, p(\tilde{Y}^{t+2} \mid X_{t+1}, \tilde{X}^{t+2})$$

and then (11.2.5) becomes apparent. Equation (11.2.4) then is obtained by first further conditioning on Y^t and then margining out X^{t-1}.

Discretized FFBS

Now that the essentials of FFBS have been laid out, we can review the necessary computations and see that they are easily performed given a state discretization. Let each X take on values in the grid $(x_1, x_2, \ldots, x_M)$. Then $p(X_{t-1}, X_t \mid Y^t)$ may be represented by an $M \times M$ matrix with rows corresponding to X_{t-1} and columns corresponding to X_t. We then obtain the marginal for X_t by summing the rows. We then create the joint distribution of (X_t, X_{t+1}) by constructing the matrix whose (i, j) element is $p(X_t = x_i \mid Y^t)\, p(X_{t+1} = x_j \mid X_t = x_i)$. To update to conditioning on Y^{t+1}, we multiply each element of this matrix by $p(Y_{t+1} \mid X_t = x_i, X_{t+1} = x_j)$ and then renormalize so that the sum over all elements of the matrix is equal to one. For the BS step we can easily compute $p(X_t = x_i \mid X_{t+1} = x_j, Y^{t+1})$ from the matrices representing the joints $p(X_t, X_{t+1} \mid Y^{t+1})$ which we must store while doing the FF step.

11.2.3.2 Equilibrium Put Option Pricing

In this section we sketch for the reader the continuous time general equilibrium model in Hore, Lopes, and McCulloch (2009). The model simultaneously determines consumption dynamics and option prices given preferences and capital accumulation dynamics. The above application of FFBS is used to estimate the discretized version of the model.

At a very high level, the full state space dynamics of equilibrium quantities and the underlying latent variable is given by

$$\begin{aligned} \frac{dC_t}{C_t} &= \mu_C(X_t)\, dt + a_{11}(X_t)\, dB_k + a_{12}(X_t)\, dB_x, \\ p_{ti} &= f(X_t, S_i, \tau_{ti}, R_t), \\ dX_t &= \delta(\bar{X} - X_t)\, dt + \sigma_x dB_x, \end{aligned}$$

where the Brownian motion terms B_k and B_x are correlated. There is a single state variable X_t (here t denotes continuous time) which is the expected return on production technology. C_t denotes the path of consumption in the economy under equilibrium and p_{ti} is the option price consistent with equilibrium consumption dynamic. The ith option has strike price S_i, time till expiration τ_{it}, and R_t is the equilibrium wealth upon which the option is written.

We orthogonalize the Brownian motion terms and express the shocks driven by independent Brownian motions. We can then write the model as

$$\begin{aligned}
\frac{dC_t}{C_t} &= \mu_C(X_t)\,dt + a_{11}(X_t)\,dB_k + a_{12}(X_t)\,dB_u,\\
p_{ti} &= f(X_t, S_i, \tau_{ti}, R_t) \qquad i = 1,\ldots,I,\\
dX_t &= \delta(\bar{X} - X_t)\,dt + a_{21}(X_t)\,dB_k + a_{22}(X_t)\,dB_u,
\end{aligned}$$

where B_k and B_u are independent Brownian motions and p_{it} is the price of the ith put option having strike price S_i and expiration time τ_{it}, for $i = 1,\ldots,I$, while R_t is the price of the underlying asset upon which the option is written. The parameters of the model are δ and $\bar{X}$. The functions μ_C, a_{ij}, and f are complex. They also depend on other model parameters, but this dependence is suppressed in order to highlight the state-space nature of the model. If we let θ denote these suppressed parameters then θ includes, for example, utility parameters that capture the risk aversion, elasticity of inter-temporal substitution, and time discount factor. The functions that describe the consumption process are derived by maximizing the expected utility of the consumption time path subject to constraints driven by the evolution of the expected return on production. The option is written on equilibrium wealth R_t determined in equilibrium by capital growth and current consumption level.

To form a discrete time version of the model over time interval Δt, we discretize the consumption and state equations and then add independent error to the option pricing equations. Let g_{t+1} be the consumption growth from t to $t+1$. Our discretized model is

$$\begin{aligned}
g_{t+1} &= \mu_g(X_t) + a_{11}(X_t)\sqrt{\Delta t}\,Z_{t1} + a_{12}(X_t)\sqrt{\Delta t}\,Z_{t2},\\
p_{ti} &= f(X_t, S_i, \tau_{ti}, R_t) + \sigma\,\varepsilon_{ti} \qquad i = 1,\ldots,I,\\
X_{t+1} &= \alpha + \rho X_t + a_{21}(X_t)\sqrt{\Delta t}\,Z_{t1} + a_{22}(X_t)\sqrt{\Delta t}\,Z_{t2},
\end{aligned}$$

where $\mu_g(X_t) = \mu_C(X_t)\Delta_t$, $\rho = 1 - \delta\bar{X}$, $\alpha = (1-\rho)\Delta_t$ and all Z_{ti} and ε_{ti} independent and identically distributed standard normal shocks. To put this model in our general form we let $W_t = a_{11}(X_t)\sqrt{\Delta t}\,Z_{t1} + a_{12}(X_t)\sqrt{\Delta t}\,Z_{t2}$ and $V_t = a_{21}(X_t)\sqrt{\Delta t}\,Z_{t1} + a_{22}(X_t)\sqrt{\Delta t}\,Z_{t2}$. We then draw V_t from its marginal distribution and W_t from its conditional distribution given V_t, or equivalently, its conditional distribution given X_{t+1} and X_t. This gives rise to the following nonlinear state space model

$$\begin{aligned}
g_{t+1} &= \mu_g(X_t) + W_t,\\
p_{ti} &= f(X_t, S_i, \tau_{ti}, R_t) + \sigma\,\varepsilon_{ti}, \;\; i = 1,\ldots,I,\\
X_{t+1} &= \alpha + \rho X_t + V_t,
\end{aligned}$$

where $W_t \sim p(W_t|X_{t+1}, X_t)$. The state equation could not be simpler, it is just an AR(1). The observation equations relate the observed relative put prices and consumption growth to the current and previous state as in our general prescription.

Inference for this complex model is now conceptually straightforward. At the top level, we have Gibbs sampler that alternates between $p(\theta|X^T)$ and $p(X^T|\theta)$. Drawing the states given θ using the discretized FFBS is quite simple. The only drawback is that since the function f is very expensive to compute it is necessary to

precompute all possible $p(y_t|X_t, X_{t-1})$ where y_t varies over all the observed values and both X_{t-1} and X_t vary over the grid values. The more difficult step is to draw from $p(\theta|X^T)$. Hore, Lopes, and McCulloch (2009), for instance, used a mixture of various Metropolis-Hastings chains.

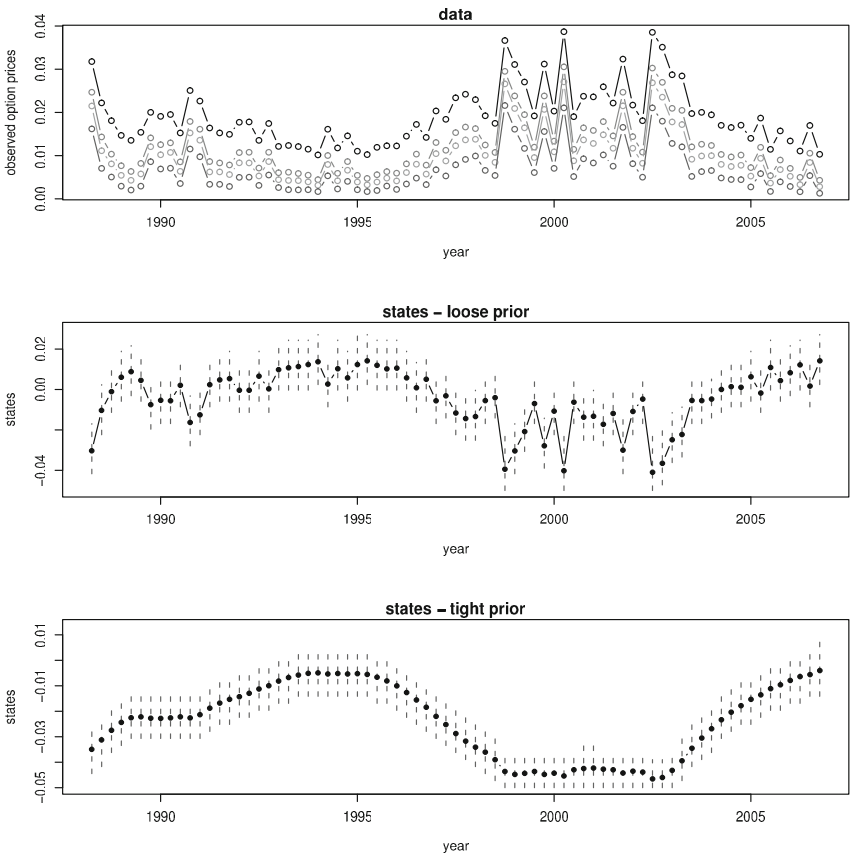


FIGURE 11.2. The top panel shows the time-series plot of the four option prices. The next two panels show the posterior distribution of the time-series of the underlying states corresponding to a loose and tight prior setting. The dashed line represents the 95% posterior band around each state.

Figure 11.2 shows the option data and state inference. Time is discretized to be months. Data on four different options corresponding to different strikes were used. The top panel plot the time series of relative option prices. The four options differ in their strike prices. The other two panels show the inference for the states X_t with t now denoting the month. The marginal distribution of each state is indicated by the vertical dashed line. Such inference is straightforwardly obtained from our FFBS draws. The difference between the two plots is the prior on the smoothness of the state evolution.

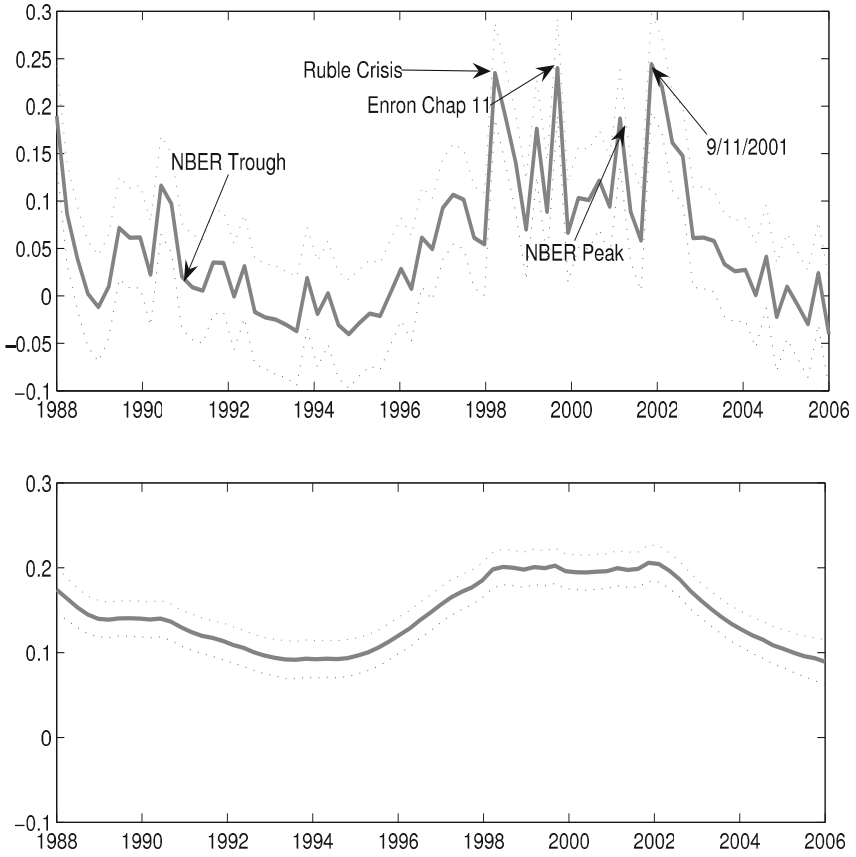


FIGURE 11.3. Time-Series of long-run risk estimated from option prices. Using the posterior distribution of parameters and the state, we compute the time-series of long-run risk. The median time-series estimate is presented in the above plot with the dotted line representing the 95% posterior interval at each point. The top time-series plot corresponds to loose prior, whereas the bottom time-series plot corresponds to the tight prior.

Inference of complex functions of our states and parameters are easily obtained given the MCMC draws. A quantity of economic interest is the risk-premia. The risk-premia in Figure 11.3 is clearly counter-cyclical. Higher growth rates imply lower risk-premia and vice versa. In bad times, the high precautionary savings motivation is consistent with high risk-premia. The agent feels more risk-averse at a time when the probability of a wealth loss is high. Clearly, the agent's demand for insurance is high and he is willing to pay more for put options in these states. Likewise, in good states the precautionary savings motive dissipates and the agent's risk-premia is low. In these states, the agent is not fearful of a wealth loss and his willingness to pay for put options to insure his wealth is low. This explains the counter-cyclical risk-premia pattern that we filter out of the time-series of put op-

tion prices. For further discussion of model specification and derivatives pricing and returns see Broadie, Chernov, and Johannes (2007, 2009).

Note that the difficulty in drawing θ also makes sequential estimation via particle filtering a practically infeasible alternative. While particle filtering might work for the draw of $X^T|\theta$ sequential, i.e., a pure state learning case, it is not a straightforward matter to get joint inference for X^T and θ, i.e., a state and parameter learning case. Liu and West (2001), for instance, who sequentially approximate $p(\lambda|y^t)$ by a multivariate normal density, is an unreasonable alternative here given the highly nonlinear nature of θ. Similarly, it would be hard to implement particle filter and particle learning algorithms (see Section 11.2.2), since the posterior distribution of X^T and θ can not be represented by a small dimensional set of conditional sufficient statistics. The discretized FFBS scheme seems to be the best available alternative.

11.2.4 Conclusion

Bayesian approaches are natural in the analysis of financial models. It has long been recognized that Bayesian thinking is relevant to fundamental questions about risk and uncertainty. Many modern financial moderns have a sequential structure expressed in terms of dynamics of latent variables. In these models, the basic Bayesian advantage in coherent assessment of uncertainty is coupled with powerful computational methods.

In this section we have reviewed two approaches to the Bayesian analysis of sequential models and attempted to illustrate ways to apply them to models derived from financial and economic theory. Particle filtering methods allow us to quickly and dynamically update our inferences about state and, in some cases, parameters. The suggested use of FFBS is much slower and more suited when complete joint inference is needed for a states over a fixed time period and underlying model parameters. The advantage of this approach is its simplicity and wide applicability without asking the user to make difficult choices about algorithm details.

11.3 Simulation-based-Estimation in Portfolio Selection

Eric Jacquier and Nicholas G. Polson

In this section we provide a simulation-based approach to optimal portfolio selection. The basic principles of portfolio selection have been known for a long time (de Finetti, 1940; Markowitz, 1959). The modern-day challenge is to apply the theory of flexible return distributions with varying degrees of conditioning information in large-scale problems. Modeling the returns distribution has a long history. Samuelson (1969) consider the i.i.d. case and showed that investors' allocation should be

horizon invariant. Merton (1972) describes the dynamic portfolio allocation with time varying conditioning information and Barberis (2000) provides a Bayesian perspective on the problem. From a statistical perspective, the big issue is accounting for estimation risk (a.k.a. parameter uncertainty) and how this affects the optimal portfolio rule, see Brandt (2009) for a recent survey.

In this section, we demonstrate how simulation-based approaches can be used to select optimal portfolios. The Bayesian approach provides a natural perspective on the problem and entails a decision-theoretic formulation (Berger, 1985) with different levels of computational tractability depending on the nature of the investor's return distribution beliefs and the horizon of the optimal allocation problem. Bayesian methods incorporate estimation risk and flexible return distributions ranging from the independent identically distributed returns, to predictability driven by exogenous latent state variables, stochastic volatility, or even multiple models. Our approach, therefore, will be flexible enough to handle complex and realistic returns distributions together with differing levels of conditioning information.

The investors' objective is to maximize expected utility. In its simplest form, we need to calculate $\max_{\omega} \mathbb{E}_t(U(\omega,R))$ where the expectation $\mathbb{E}_t$ is taken with respect to our current conditioning set Z_t of the investors' beliefs up to time t. The decision variable ω is a vector of asset weights and R a vector of future returns. There is often no analytical solution to the problem. A conceptually simple Monte Carlo approach for finding the optimal decision is as follows: first simulate a set of returns $R^{(i)} \sim p(R|Z_t)$ for $i = 1,\ldots,N$. Then, given these draws, we estimate the expected utility for a decision ω with an ergodic average of the form

$$\mathbb{E}_t\left(U(\omega,R)\right) = \frac{1}{N}\sum_{i=1}^{N} U(\omega,R^{(i)}),$$

and optimize this MC average of utility over the decision ω. This can be problematic, however, when the utility places weights on the tails of the future return distribution and will lead to a poor estimate. Later we provide an alternative MCMC approach that can simultaneously perform the averaging (over $\mathbb{E}_t$) incorporating parameter uncertainty, and the optimizing (over ω) to find the optimal weights. Other statistical issues that arise in formulating the future distribution of returns include analyzing assets of different history lengths, see, for example, Stambaugh (1999), Polson and Tew (2000) and Gramacy and Pantaleo (2009).

A more challenging problem arises when the investor wishes to solve a multiperiod problem. Again the investor will have a set of conditioning variables Z_t at his disposal that will typically include exogenous predictors such as dividend yield, term premium and current volatility state. To proceed, consider the value function is defined by

$$U_T(W_t,Z_t) = \max_{\omega_s;t\le s\le T} \mathbb{E}_t\left[U(W_T(\omega))|Z_t\right],$$

where W_t is current wealth. The evolution of terminal wealth now depends on the sequential portfolio allocation $\underline{\omega} = \{\omega_s; t \le s \le T\}$ with horizon T. The solution

clearly differs from a sequence of myopic portfolio rules — the difference being the hedging demand.

The rest of the section is outlined as follows. Section 11.3.1 describes the optimal portfolio selection problem as an expected utility optimization problem. We consider a number of cases including the possibility of Bayesian learning for the investor. Section 11.3.2 provides a simulation-based MCMC approach to simultaneously account for estimation risk and to find the optimal rule. Finally, Section 11.3.3 concludes.

11.3.1 Basic Asset Allocation

Since the foundational work of Samuelson (1969) and Merton (1971), the optimal portfolio problem has been well studied. First, we review the optimal portfolio rule in this simple setting of complete information about the parameters of the return distribution. Then we consider a number of extensions; multivariate, exchangeable and predictable returns. In the next section we discuss in detail simulation-based approaches for finding optimal portfolios in the presence of parameter uncertainty.

The original work of Samuelson and Merton shows that if asset returns are i.i.d., an investor with power utility who rebalanced optimally should choose the *same* asset allocation regardless of the time horizon. If returns are predictable there will be an advantage to exploit it. In many cases, investors with a longer horizon will allocate more aggressively to stocks. Jacquier, Kane, and Marcus (2005) show that parameter uncertainty produces the exact opposite, but much stronger, results. Namely, on account of parameter uncertainty, the long-term investor will invest much less in stocks.

The investor who optimally re-balances his portfolio at regular intervals faces a dynamic programming problem. The use of power utility for sequential investment problems with Bayesian learning goes back to Bellman and Kalaba (1958). Ferguson and Gilstein (1985) and Bruss and Ferguson (2002) provide extensions. In this case the utility function is given by $U(W) = W^{1-\gamma}/(1-\gamma)$ with utility defined over current wealth. The special case of $\gamma = 1$ corresponds to log-utility and the Kelly criterion. Browne and Whitt (1986) discuss Bayesian learning in this context. Barberis (2000) extends this analysis and shows that this leads to horizon effects where, in particular, people with large time horizons are willing to hold more stocks.

11.3.1.1 Single Period

We start in a univariate one-period setting. This can be generalized in a number of ways, to a cross-section of returns or to a multivariate set of returns.

The optimal portfolio weight ω can be determined as follows: the investors' wealth is $W = (1-\omega)r_f + \omega R$ with risky free rate r_f and risky return R. The problem is to choose ω to maximize the expected utility, $\max_\omega E(U(\omega))$. If U is twice

differentiable, increasing and strictly concave in ω, the optimal allocation is characterized by the first order condition:

$$E\left[U'(W)(R-r_f)\right]=0$$

which yields $\text{Cov}\left[U'(W),R-r_f\right]+E\left[U'(W)\right]E\left[R-r_f\right]=0$. Stein's lemma (Berger, 1985) equates the covariance of a function of normal random variables to the underlying covariance times a proportionality constant. If X denotes a normal random variable, $X\sim\mathcal{N}(\mu,\sigma^2)$ with mean μ and variance σ^2 and $g(X)$ is differentiable such that $E|g'(X)|<\infty$, then $\text{Cov}\left[g(X),X\right]=E\left[g'(X)\right]\sigma^2$. In the bivariate case for normal random variables (X,Y), Stein's lemma becomes $\text{Cov}\left[g(X),Y\right]=E\left[g'(X)\right]\text{Cov}\left[X,Y\right]$. Applying this identity to the first order condition yields:

$$\omega E\left[U''(W)\right]\text{Var}(R)+E\left[U'(W)\right](E(R)-r_f)=0.$$

Hence, the optimal allocation $\omega^\star$ has a simple closed form

$$\omega^\star=\frac{1}{\gamma}\left(\frac{\mu-r_f}{\sigma^2}\right),$$

where $\mu=E[R]$ and $\sigma^2=\text{Var}(R)$. The parameter γ is the agent's global absolute risk aversion: $\gamma=-E\left[U''(W)\right]/E\left[U'(W)\right]$. This approach can be extended to the case of stochastic volatility, see Gron, Jorgensen, and Polson (2004).

To illustrate this basic result the average real return for quarterly US data over the period 1947.2 to 1998.4 shows a return of 8.1%. The average riskless real interest rate is 0.9% per year. Stocks are volatile with an annualized standard deviation of 15.6% for this period. A reasonable risk-aversion of $\gamma=4$ would then lead to an allocation of 71% stocks.

The combination of a risk-free asset with any risky asset occur on a straight line, denoted capital allocation line, in this (mean, standard deviation) space. So the introduction of a risk-free asset to the investment opportunity set, brings in the tangency portfolio T, with the highest slope, aka Sharpe ratio, the ratio of its expected premium over the risk-free rate μ_T-R_f by its standard deviation σ_T. All investors allocate their wealth along that line according to their attitude to risk.

In an i.i.d. log-normal risky asset $R_t\sim N(\mu,\sigma^2)$ has T-period log-return given by \$1 is $\log(1+R_T)\sim N(\mu T,\sigma^2 T)$. A common choice is power utility of final wealth, namely $U(W_T)=\frac{1}{1-\gamma}\exp[(1-\gamma)\log(1+R_T)]$. This can be greatly affected by estimation risk as illustrated in Jacquier, Kane, and Marcus (2005).

We now consider a number of extensions: the multivariate mean-variance case; exchangeability in the cross-section and time series dimensions and finally how allocation rules are affect by return predictability.

11.3.1.2 Mean-Variance

Mean-variance portfolio theory was pioneered by de Finetti (1940) and Markowitz (1959, 2006). The basic mean-variance problem for the investor reduces to finding portfolio weights that solve the quadratic programming problem:

$$\min_{\omega} \omega' \Sigma \omega \quad \text{subject to} \quad \begin{matrix} \omega' \iota = 1, \\ \omega' \mu = \mu_P, \end{matrix}$$

where ι is a vector of ones. If asset returns are jointly normal $N(\mu, \Sigma)$, for computational convenience our expected utility depends only on its moments. The efficient frontier, with no short sales constraint, has a long history and is well understood.

We can then identify the *mean/variance efficient portfolio*,

$$\omega_{EV} = \frac{1}{\iota' \Sigma^{-1} \mu} \Sigma^{-1} \mu$$

with expected return $\mu_{EV} = \mu' \Sigma^{-1} \mu / \iota' \Sigma^{-1} \mu$. We can also define the *minimum variance* portfolio $\omega_{MV}(\Sigma)$ by $\omega_{MV}(\Sigma) = \frac{1}{\iota' \Sigma^{-1} \iota} \Sigma^{-1} \iota$ which has expected return $\mu_{MV} = \iota' \Sigma^{-1} \mu / \iota' \Sigma^{-1} \iota$. The global minimum variance portfolio just depends on the variance-covariance matrix Σ and so, from a statistical viewpoint, becomes a good portfolio to study as we change the input Σ.

As discussed in Perold (1988) and Chopra and Ziemba (1993), implementation of portfolio choice in higher dimensions tends to result in extreme weights on securities. One strategy to approach this issue is to introduce constraints in the optimization problem. We introduce upper and lower constraints in the optimization problem by letting $c_i(x_i) = -\infty$ if $x_i < l_i$ or $x_i > u_i$ and consider the problem of $\min_{\omega} \frac{1}{2} \omega' \Sigma \omega - \omega' \mu - \lambda c(\omega)$ where $c(\omega) = \sum_{i=1}^{k} c_i(x_i)$. In the indexing problem we will typically choose $c_i(x_i) = c$ on $l_i < x_i < u_i$ and $l_i = 0 \; \forall i$ (reflecting a no short-sales constraint) and $u_i = u_0$ a predetermined constant upper bound. The higher the level of u_0 the more aggressive the portfolio in the sense of the few numbers of securities held and the higher tracking error of the portfolio. Other choices could depend on the benchmark index weights or the individual volatilities of the securities. For the implications of higher-order moments on optimal portfolio rules see Harvey et al. (2004).

It has long been known that "plug-in" estimates of variance-covariance matrices can be very noisy estimates of the underlying parameters. Moreover, the optimizer tends to focus on these estimation errors and can lead to extreme weights. Specifically, using $\mu_{EV} = \hat{\mu}' \hat{\Sigma}^{-1} \hat{\mu} / \iota' \hat{\Sigma}^{-1} \hat{\mu}$, where $\hat{\mu}$ and $\hat{\Sigma}$ are the MLEs, can lead to poor performance. Jobson and Korkie (1980) provide a simulation study and illustrate these effects. There are a number of ways of dealing with this, the most popular being shrinkage based estimation (Black and Litterman, 1991). One approach is to use Bayesian estimators derived from prior information (for example, shrinking towards market equilibrium) use the posterior distribution $p(\mu, \Sigma | Z_t)$. Polson and Tew (2000) argue instead for the use of posterior predictive moments instead of plug-in

estimates of means and variance-covariances which naturally accounts for parameter uncertainty.

In the cross-section, we can extend the independence assumption by assuming that the multivariate return distribution is exchangeability. Then its joint distribution is invariant to permutation. The order of the variables leads to the same joint. There are two cases, either exchangeable in the cross-section or in the time series. In a one period setting, if the conditional distribution of the returns R is exchangeable, then the optimal portfolio rule is $\omega = 1/N$. Hence the diversified equally-weighted rule is optimal.

One can see this as follows: suppose you invest ω_i in ith asset, with $\sum \omega_i = 1$. Then your expected utility of wealth tomorrow of the portfolio $\omega \cdot R$ is

$$EU(\omega \cdot R) = EU(\pi(\omega) \cdot X)$$

for any permutation π, where $\pi(\omega)_i = \omega_{\pi(i)}$ due to exchangeability. Hence

$$\begin{aligned} EU(\omega \cdot X) &= \frac{1}{N!} \sum_{\pi} EU(\pi(\omega) \cdot X) \leq EU\left(\frac{1}{N!} \sum_{\pi} \pi(\omega) \cdot X \right) \quad \text{(concavity of U)} \\ &= EU\left(N^{-1} 1 \cdot X \right). \end{aligned}$$

So whatever the concave utility, under this exchangeability hypothesis you're best to use the $1/N$-rule. DeMiguel, Garlappi, and Uppal (2009) provide empirical out-of-sample performance for this rule and document its surprisingly good performance against other strategies.

11.3.1.3 Estimation Risk without Predictability

Incorporating parameter uncertainty or estimation risk is important for a number of reasons. First, it can dramatically affect the optimal holds when the investors' time horizon is taken into account. Second, it more realistically models historical returns data. Maximizing expected utility can be computationally intractable. A strain of literature concentrates on a function of the first (four at most) moments. We discuss Harvey et al. (2004) for a Bayesian implementation.

Classical mean variance optimization requires estimates of the mean and variance-covariance matrix of all assets in the investment universe. Maximum likelihood estimates suffer from having poor sampling properties such as mean squared error in high dimensional problems. An advantage of the Bayesian approach is that it naturally allows for regularization through the choice of prior. Estimation risk is then seamlessly taken into account and one can also combine market equilibrium information with an investors investment views as in the popular Black-Litterman (1991) model.

To illustrate what happens when you take into account the estimation risk, consider the time series exchangeable case. Here the predictive distribution of returns is given by

$$p(R_{t+1}|Y^t) = \int p(R_{t+1}|\mu,\Sigma)p(\mu,\Sigma|Y^t)d\mu d\Sigma.$$

Estimation risk is taken account of by marginalizing (integrating) out the uncertainty in the parameter posterior distribution. The mean-variance approach leads to a further simplification and under elliptical distributions such as the multivariate normal with the posterior mean being the predictive mean.

The Bayesian investor learns the mean and variance via the updating formulas

$$\mu_{t+1} = E\left(R_{t+1}|Y^t\right) = E\left(\mu|Y^t\right),$$
$$\Sigma_{t+1} = E\left(\Sigma|Y^t\right) + \mathrm{Var}\left(\mu|Y^t\right).$$

A portfolios excess return is defined as the rate of return on the portfolio minus the Treasury bill rate. Polson and Tew (2000) show that with full information and a longer time series than assets that a standard non-informative prior $p(\mu,\Sigma) \equiv |\Sigma|^{\frac{(m+1)}{2}}$ that the predictive variance-covariance is proportional to the maximum likelihood estimate $\hat{\Sigma}$ and so there is no effect of estimation risk.

The differences appear with large assets and with the common situations of missing data. Gramacy, Lee, and Silva (2009) develop predictive distributions with missing data and extend the pricing errors to fat-tailed t-errors and regularization penalties tailored to high dimensional problems. The myopic rule obtained by plug-in these predictive means and covariances ignores the inter-temporal hedging demands that exist as the investor re-balances his posterior distributions. Campbell and Viceira (1999) provide a discussion of this and show that in many cases the hedging demand is negligible.

A popular practitioners' approach is Black and Litterman (1991) who note that modifying one element of the vector of means, for which one has better information, can have an enormous and unwanted impact on all the portfolio weights. They combine investor views and market equilibrium, in the spirit of shrinkage, by shrinking to equilibrium expected returns.

One nice feature of the Bayesian approach is that one can incorporate individual views via shrinkage (Black and Litterman, 1991). Specifically, suppose that excess log-returns have a multivariate normal distribution

$$(R_{t+1}|\mu,\Sigma) \sim N(\mu,\Sigma) \text{ and } (\mu|\bar{\mu},\lambda) \sim N(\mu,\Lambda)$$

with a corresponding multivariate normal prior. We can use this to place restrictions on a linear combination (or portfolio) of returns which yields

$$(P\mu|\bar{\mu},\lambda) \sim N\left(P\mu, P\Lambda P'\right)$$

for a $K \times N$-matrix P. We can choose $P\mu$ to be equilibrium market weights. Let $\Omega = P\Lambda P'$. Then Bayes rule gives the updated weights

$$E(\mu|y) = \left(\Lambda^{-1} + P'\Omega P\right)^{-1}\left(\Lambda^{-1}\bar{\mu} + P'\Omega^{-1}PR\right),$$
$$\mathrm{Var}(\mu|y) = \left(\Lambda^{-1} + P'\Omega P\right)^{-1}.$$

The Ω matrix can be found, for example, by exploiting the use of a factor model.

11.3.1.4 Estimation Risk with Predictability

When predictability is present, it is common to model excess returns using a vector auto-regression (VAR) of the form

$$Y_t = Bx_t + \Sigma^{\frac{1}{2}}\varepsilon_t,$$

where Y_t contains both the stock return information as its first component and the remaining components are variables that are thought to be useful for predicting returns. Let $\beta = \mathrm{vec}(B)$, a $T \times k$ by 1 vector. We need to be able to simulate from the joint posterior distribution $p(\beta, \Sigma|y)$. The likelihood function is given by

$$p(Y|\beta,\Sigma) = (2\pi)^{-\frac{Tk}{2}}|\Sigma|^{-\frac{T}{2}}\exp\left(-\frac{1}{2}\sum_{t=1}^{T}(Y_t - Bx_t)'\Sigma^{-1}(Y_t - Bx_t)\right).$$

The prior distribution can either be diffuse or the usual conjugate matrix normal-inverse Wishart.

Barberis (2000) quantifies the magnitude of estimation risk, parameter learning and optimal decisions. Various predictability regression models have been proposed to predict future excess market returns. The basic model is

$$r_{t+1} = \alpha + \beta x_t + \sigma\varepsilon^r_{t+1},$$
$$x_{t+1} = \alpha_x + \beta_x x_t + \sigma_x\varepsilon^x_{t+1},$$

where r_{t+1} are monthly returns on the CRSP value-weighted portfolio in excess of the risk-free rate, and the predictor variable x_t is the payout yield, defined as the time t payouts over the past year divided by the current price. The errors are jointly standard normal, with $\mathrm{Corr}(\varepsilon^r_t, \varepsilon^x_t) = \rho < 0$. Typical estimates are in the range of -0.7 depending on the sample period. The effect of a negative correlation is that it is more likely that a drop in the regressor (dividend yield) is associated with a positive shock to stock returns. This in turn has the effect, since dividend yields are lower, of inducing a stock return forecast that is lower in the future.

The intuition of the effect of estimation risk is then as follows. Time variation in expected returns induces mean reversion into returns slowing the growth of cumulative variances of long-horizon returns. This makes equities look less risky. Portfolio decision makers, therefore, allocate more to stocks even though they face substantial parameter uncertainty. Similar statements can be said about model risk, except clearly these effects can be greater. Johannes, Korteweg, and Polson (2009) provide a sequential Bayesian analysis of this portfolio problem including investors updat-

ing beliefs about model probabilities of a variety of models that also incorporate stochastic volatility.

Models with multiple predictors have been analyzed in a Bayesian setting by Avramov (2002), Cremers (2002) and Boudoukh et al. (2007). We also do not impose economic restrictions on the regressions as in Campbell and Thompson (2008). There is also a large literature has tackled the issue of testing the efficiency of portfolios. Kandel, McCulloch, and Stambaugh (1995) find the posterior distribution of the maximum correlation between the tested portfolio tested and any portfolio on the efficient frontier.

Brandt (2009) provides the following example. The underlying dynamics are given by a VAR model of the form

$$\begin{pmatrix} \ln(1+R_{t+1}) \\ \ln dp_{t+1} \end{pmatrix} = \beta_0 + \beta_1 \ln dp_t + \varepsilon_{t+1}, \tag{11.3.1}$$

where dp_{t+1} is the dividend-to-price ratio and the errors are assumed to be homoscedastic normals. Brandt finds that solving the optimal portfolio problem at the median dividend yield leads to the following weights. The optimal allocation to stocks is 58% for a one-quarter horizon, 66% for a one-year horizon and 96% for a five-year horizon. At a single-period horizon the allocation to stocks at the $(25, 50, 75)$th quantiles of the dividend-to-price ratio are 23%, 58%, and 87%, respectively.

11.3.1.5 Assessing Model Risk

Model selection can be performed as follows. Let $\{\mathscr{M}_j\}_{j=1}^J$ be a collection of models and $X^t = (X_1, \ldots, X_t)$ a vector of state variables. Consider a factorization of the posterior distribution of states and models as

$$p\left(X^t, \theta, \mathscr{M}_j | y^t\right) = p\left(X^t, \theta | \mathscr{M}_j, y^t\right) p\left(\mathscr{M}_j | y^t\right), \tag{11.3.2}$$

which dissects the inference problems into two components. First, $p\left(X^t, \theta | \mathscr{M}_j, y^t\right)$ solves the parameter and state "estimation" problem conditional on a model. Bayes theorem implies that the posterior given a model is

$$p\left(X^t, \theta | \mathscr{M}_j, y^t\right) = \frac{p\left(y^t | X^t, \theta, \mathscr{M}_j\right) p\left(X^t | \theta, \mathscr{M}_j\right) p\left(\theta | \mathscr{M}_j\right)}{p\left(y^t | \mathscr{M}_j\right)},$$

where $p\left(y^t | \theta, X^t, \mathscr{M}_j\right)$ is the full-information likelihood (conditional on the latent states), $p\left(X^t | \theta, \mathscr{M}_j\right)$ is the stochastic specification for the dynamics of the latent variables (e.g., the specifications for the dynamics of V_t^r, V_t^x, and β_t), and $p\left(\theta | \mathscr{M}_j\right)$ is the prior distribution of the parameters in model j. It is important to note that all of these components are subjective modeling assumptions. Bayesian statistical inference involves summarizing $p\left(X^t, \theta | \mathscr{M}_j, y^t\right)$ in useful ways.

The second component of (11.3.2) consists of $p(\mathscr{M}_j|y^t)$, or more aptly, comparing $p(\mathscr{M}_j|y^t)$ to $p(\mathscr{M}_k|y^t)$. This portion of the inference problem is called model choice or model discrimination. Jeffreys (1961) introduced Bayesian model comparison, which weighs the relative strength of one model to another via the posterior odds ratio of model j to model k :

$$odds\left(\mathscr{M}_j \text{ vs. } \mathscr{M}_k|y^t\right) = odds_t^{j,k} = \frac{p(\mathscr{M}_j|y^t)}{p(\mathscr{M}_k|y^t)} = \frac{p(y^t|\mathscr{M}_j)}{p(y^t|\mathscr{M}_k)}\frac{p(\mathscr{M}_j)}{p(\mathscr{M}_k)}.$$

The priors odds ratio is $p(\mathscr{M}_j)/p(\mathscr{M}_k)$ and the Bayes factor is the marginal likelihood ratio.

Johannes, Korteweg, and Polson (JKP, 2009) consider an extension of the basic dividend-yield regression to five model specifications incorporating stochastic volatility and drifting coefficients:

$$r_{t+1} = \alpha + \beta_t x_t + \sigma\sqrt{V_t^r}\varepsilon_{t+1}^r, \tag{11.3.3}$$

$$x_{t+1} = \alpha_x + \beta_x x_t + \sigma_x\sqrt{V_t^x}\varepsilon_{t+1}^x, \tag{11.3.4}$$

where V_t^r, V_t^x are stochastic volatility factors each with their respective dynamics. Hence we have a list of models $\mathscr{M}_i$ including the benchmark model as well as

1. 'DC' (for drifting coefficients) denotes the extension of the standard model with volatility still constant.
2. 'SV' denotes the extension which assumes that volatility is stochastic.
3. 'SVDC' denotes the most general specification.

JKP provides an illustration of the evolution of sequential posterior model probabilities as marginal distributions from the full joint distribution $p(X_t, \mathscr{M}_i, \theta_i|y^t)$ where θ_i are model specify parameters over time.

11.3.2 Optimum Portfolios by MCMC

We now describe an algorithm introduced by Jacquier, Johannes, and Polson (JJP, 2007), which produces the optimum of expected utility, and d^*, without using gradient methods. Consider again the generic problem of an agent solving

$$\max_d E_X\left[U(X,d)\right],$$

where $U(X,d)$ represents the agent's utility given a decision d, and X is the random variable directly relevant for computing utility. The expectation E_X is taken with respect to the distribution $p(X)$, which is the predictive density of the X after marginalizing the other parameters and state variables. Draws of $p(X)$ can be made, either directly or via any of the many MCMC algorithms that appear in the recent literature. For example in a portfolio problem, X is the vector of future returns on the

vector of investment opportunities, and the marginalized parameters and state variables can be means and variances of portfolio returns or an unobserved time varying covariance matrix. One may want to characterize the optimal decision as a function of conditioning variables Y. Then one needs to consider $p(X|Y)$ and the associated optimal decision rule $d^*(Y)$. This is the functional optimization framework. For the portfolio problem, Y, could be a latent variable such as the future volatilities or any other estimated parameter.

The simulation-based algorithm used exploits desirable properties of MCMC simulation of $p(X)$, so that the algorithm also produces the optimal decision rule. Specifically, the method produces the optimal decision d^* without having to compute a simulated expected utility and its derivatives or implement a gradient method on that simulated function.

One simply *randomizes* the decision rule and makes draws that concentrate on the optimal decision. It is crucial to note that this is consistent with decision theory within which decision variables are not random, as in Berger (1985), Chamberlain (2000), and Zellner (2002). The variability in the draws of d in the algorithm is purely of a computational nature, it represents in no way an econometric or economic uncertainty about the decision variable for the agent. In fact JJP show that the algorithm collapses on the optimal decision as some choice variable increases.

The algorithm proposed can serve to find the global optimal decision d^*, the optimal functional decision rule $d^*(Y)$, or the optimal sequential decision (d_1^*, d_2^*). The algorithm constructs draws from a joint distribution on d and J replications of X, denoted $\pi_J(\widetilde{X}^J, d)$. Specifically, the joint density of the parameters and the decision is defined by

$$\pi_J(\widetilde{X}^J, d) \propto \prod_{j=1}^{J} U(X^j, d) p(X^j) \mu(X, d), \tag{11.3.5}$$

where $\mu(X,d)$ is a measure, typically uniform, that will be used to enforce the needed regularity conditions in the standard utility framework. Typical restrictions on the portfolio weights d and the predictive density of returns X can be imposed.

JJP first show that the marginal density on the decision variable $\pi_J(d)$, obtained by integrating out $\widetilde{X}^J$ from the joint density in (11.3.5) is

$$\pi_J(d) = C(d) \exp\{J \ln E(U(\theta, d))\}$$

for an appropriate normalizing constant $C(d)$. Then, $\pi_J(d)$ collapses on the optimum decision $d^* = \arg\max E_X[U(X,d)]$ as J becomes large. This happens for practical, i.e., low enough, values of J. An asymptotically normal limiting result in J under extra suitable regularity conditions provides a diagnostic for selecting J.

Well known Markov chain Monte Carlo (MCMC) methods can be used to make draws $\{(\tilde{X}^{J,(g)}, d^{(g)}), \quad g = 1, \ldots, G\}$ from $\pi_J(\widetilde{X}^J, d)$. Therefore, we draw from the $J+1$ conditionals, $X^j|d$ and $d|\widetilde{X}^J$, which can be shown to be

$$X^j|d \sim U\left(X^j,d\right)p\left(X^j\right) \text{ for } \quad j=1,\ldots,J, \tag{11.3.6}$$

$$d|X^1,\ldots,X^J \sim \prod_{j=1}^{J} U\left(X^j,d\right)p\left(X^j\right). \tag{11.3.7}$$

Note that the draws from X^j are tilted away from the predictive density $p(X^j)$ toward $p(X^j)U(X^j,d)$, while, as we will see, the algorithm has d converging to d^*. So the draws of X^j concentrate on the regions of the domain of X^j with a higher utility. So, in the spirit of importance sampling, the algorithm concentrates on "smart" values of X^j. Here the importance function is the utility, which itself tightens around d^* as the algorithm converges. Sampling the X^j's in a utility-tilted way helps converge quicker to the relevant region of the decision space using $d|\widetilde{X}^J$.

This differs from *standard* expectation-optimization algorithms for two reasons. First we draw efficiently from $p(X|d)$ as just discussed. In contrast, expectation-optimization algorithms, at every step of d, draw G samples from $X^{(g)} \sim p(X)$, the predictive density of X. Second, they approximate the expected utility $E_X[U(X,d)]$ by $\frac{1}{G}\sum_{g=1}^{G} U\left(X^{(g)},d\right)$, as well as all required derivatives similarly, and an optimization step over d is performed typically via a gradient-based method. The process is repeated until convergence. For functional optimization and sequential problems this can be computationally intractable.

11.3.2.1 MCMC Algorithm for Maximizing Expected Utility

An agent wants to find the global optimal decision d^* or study the optimal functional decision $d^*(Y)$ for a wide range of values of Y, where Y is a parameter or state variable of interest. For example, the agent may want to understand the sensitivity of the portfolio to potential variations in volatility or revisions of expected returns.

The uncertainty of state variable X is described by the conditional predictive distribution $p(X|Y)$. This distribution follows from the integration of the other state variables and parameters not directly relevant to the agent. However, as the agent wants to study the optimal decision as a function of the state variable or parameter of interest Y, we do not integrate it out. Specifically, the agent wants to solve

$$\max_{d(Y)} E_{X|Y}\left[U\left(X,d\right)\right].$$

Define the augmented joint distribution

$$\pi_J(\widetilde{X}^J,d,Y) \propto \prod_{j=1}^{J} U(X^j,d)p(X^j|Y)\mu(d,X,Y) \tag{11.3.8}$$

for some measure $\mu(d,X,Y)$, typically a uniform to ensure that the regularity conditions hold true. We drop μ from the rest of the section to lighten the notation.[3]

We now consider the following MCMC algorithm that simulates from $\pi_J(\widetilde{X}^J,d,Y)$:

$$X^j|d,Y \propto U\left(X^j,d\right)p\left(X^j|Y\right) \text{ for } j=1,...,J, \tag{11.3.9}$$

$$d|\widetilde{X}^J,Y \propto \prod_{j=1}^{J} U\left(X^j,d\right)p\left(X^j|Y\right), \tag{11.3.10}$$

$$Y|\widetilde{X}^J,d \propto p(\widetilde{X}^J|Y)=\prod_{j=1}^{J} p(X^j|Y). \tag{11.3.11}$$

Clearly, upon integrating out Y, the above MCMC set of conditionals reduces to the conditionals given in (11.3.6), (11.3.7). Hence, the results presented below specialize to the simple global maximum expected utility problem, using (11.3.6), (11.3.7), instead of (11.3.9)-(11.3.11). The slice sampler can also be used to draw efficiently from (11.3.9)-(11.3.11).

For the purpose of functional optimization, one could conceivably use an algorithm based upon (11.3.9) and (11.3.10), for a selected value of Y, repeating the procedure for a a discrete grid of values of Y. This brute force procedure, while correct, is not efficient as possibly uninteresting values of Y may be selected. Indeed, the complete algorithm in (11.3.9), (11.3.10) and (11.3.11) has two advantages. First, it gives the global optimum d^* as a by-product. Second, it draws Y more frequently where it has higher expected utility, as per the conditional in (11.3.11). This is an efficiency gain in the spirit of importance sampling. Note however that if J gets extremely large, the algorithm collapses around d^*, as per (11.3.10), hence around some Y^* as can be seen from (11.3.11). It becomes then impractical as a means to describe the function $d^*(Y)$. This is however not likely in practice. Take for example, an already high $J=200$, and Y being an unknown variance. Then (11.3.11) is akin to a draw from the variance of X given 200 observations. Clearly, it would take a much higher value of J to collapse $Y|\widetilde{X}^J,d$ on one value.

Practically, the algorithm produces joint draws d,Y that can be plotted. The optimal function $d^*(Y)$ can then found with one of the many known kernel-smoothing techniques. The optimal value d^* is found by averaging the draws of d as in any MCMC algorithm.

We now show that the marginal of $\pi_J(d|Y)$ collapses on the functional relationship $d^*(Y)$ as J gets large. First, it follows from (11.3.8) that:

$$\pi_J(d|Y)=C(d)e^{J\log\left(E_{X|Y}[U(X,d)]\right)}.$$

In turn, as $J\to\infty$, we have that

[3] The families of utilities most always used in financial economics are the power and the exponential. Both are negative. One remedies this problem by shifting the utility. We proceed in this section under the assumption that $U\geq 0$.

$$\pi_J(d|Y) \longrightarrow d^*(y) = \arg\max_{d(y)} E_{X|y}[U(X,d)].$$

The problem then is to find an efficient MCMC algorithm to sample this joint distribution over the space (d,X,Y). The Markov Chain produces draws $\left\{(\widetilde{X}^{(j,g)}, d^{(g)}, Y^{(g)}), g = 1,\ldots,G\right\}$.

In the global optimization problem, we know that maximizing $U(d)$ is equivalent to simulating from the sequence of densities $\pi_J(d) \propto \exp(J \log U(d))$, as J becomes large, see Pincus (1968). Simulated annealing uses this result to construct a Markov Chain over d to sample from π_J, see Aarts and Korst (1989) for example. Unfortunately, a key assumption of simulated annealing is that $U(d)$ can be exactly evaluated. This is not the case here as $U(d) = \int U(d,X)\, p(X)\, dX$ is not analytic. In contrast, our approach relies on the following key result from evolutionary Monte Carlo: $\pi_J(d)$ can be viewed as the marginal distribution of $\pi_J\left(d, \widetilde{X}^J\right) \propto \prod_{j=1}^{J} U(d,X^j) p(X^j)$. This suggests that the Markov Chain should operate in the higher dimensional space of $\left(d, \widetilde{X}^J\right)$. MCMC is then the natural methodology to sample these variables. This is why we draw iteratively from $p\left(d|\widetilde{X}^J\right)$ and $p\left(\widetilde{X}^J|d\right)$ and eventually the simulated $d^{(G)} \longrightarrow d^*$.

Recall that standard simulation draws X from $p(X)$. In contrast, this approach samples the J random variables, $X^j|d \propto U(d,X^j)p(X^j)$. This is why the approach will work well for large dimensions, complicated distribution and potentially non-smooth utility. For example, in the case where the maximizing decision depends critically on the tail behavior, it will generate more samples from the high-utility portions of the state space.

A key advantage of this joint optimization and integration approach is that it delivers Monte Carlo error bounds in high dimensions. This is due to the fact that using MCMC sampling can result in fast convergence such as geometric convergence λ^G in nearly all cases and polynomial time in some cases. Contrast this approach with even sophisticated Monte Carlo strategies such as importance sampling that generates the standard central limit theorem type $\sqrt{G}$ convergence. Aldous (1987) and Polson (1996) discuss why this is insufficient for high dimensional problems and consider random polynomial time convergence bounds.

11.3.3 Discussion

This section provides a discussion of Bayesian methods in portfolio selection. Simulation-based methods are particularly suited to solving the integration problem for estimation risk and the optimization problem to find the portfolio weights. A major problem for future research are dynamic asset allocation setting in many dimensions, see, for example, Brandt et al. (2005). Extending these methods to higher dimensions is challenging. One alternative avenue for future research is to apply

Q-Learning techniques to solving the dynamic multi-period problem under uncertainty, see Polson and Sorensen (2009).

11.4 Bayesian Multidimensional Scaling and Its Applications in Marketing Research

Duncan K.H. Fong

Multidimensional scaling (MDS) refers to a set of multivariate procedures that provide spatial representations for displaying the structure in various types of empirical data (proximity, dominance, profile, etc) in one or more latent dimensions. The technique is widely used in social and behavioral sciences. It has its roots in the mathematical psychology literature (cf. Shepard, 1980), and was originally used to assess subjects' judgments and attitudes to various presented stimuli. It has now become quite popular generically and has extended into areas other than its traditional place in the behavioral sciences. The topic is also gaining attention in statistics, and it is covered in many graduate level multivariate statistics textbooks (e.g., Johnson and Wichern, 2007). Indeed, there is a very wide variety of multivariate visualization and dimension reduction techniques related to classical MDS (e.g., Diaconis, Goel, and Holmes, 2008; Chen and Buja, 2009).

The MDS technique has been widely used in marketing research for positioning, market segmentation, optimal product/service design, etc. (see Carroll and Green, 1997; DeSarbo and Wu, 2001; Lilien and Rangaswamy, 2004). In particular, MDS procedures that produce joint space maps representing both brands and consumers are most useful for representing the relationships between the two entities in examining the underlying structure of preference data (Johnson, 1971; Green, 1975). Classical MDS procedures for two-way preference data abound in terms of either unfolding representations, vector representations, or correspondence analysis and optimal scaling type approaches (e.g., DeSarbo and Rao, 1986; Gifi, 1990; Cox and Cox, 2001; Borg and Groenen, 2005). In unfolding MDS procedures (e.g., ALSCAL; see Takane, Young, and de Leeuw, 1977), both row (e.g., consumer) and column (e.g., brand) elements of the input data matrix are represented by points in a reduced dimensional space and the Euclidean distance between these row and column points are indicative of the dominance relations shared between them in the data. In a vector MDS procedure (e.g., MDPREF; see Carroll, 1980), one set of entities (either row or column) are represented by vectors emanating from the origin of the derived joint space while the alternative set of entities are represented by points. Here, the orthogonal projection of the points onto these vectors in the reduced dimensional space renders information on the dominance relationships contained in the data. Correspondence analysis (e.g., Benzecri, 1992; Shin, Fong, and Kim, 1998) is an exploratory technique typically used to analyze contingency tables containing some measure of correspondence between the rows and columns. The graphic re-

sults from correspondence analysis provide information which is similar in nature to those produced by principal component analysis techniques. However, unlike the unfolding or vector model, associations between row and column entities are often more difficult to assess directly.

There have been a plethora of applications of these types of two-way MDS procedures in various disciplines. For example, Herche and Swenson (1991) used MDPREF to measure classroom performance where vectors represented teaching attributes and faculty was represented by points. In political science, such MDS techniques were used to assess the dimensions of political perception and individual differences in the importance of each dimension in preferential choice (Matsusaka and McCarty, 2001). These techniques have also been used to assess citizen attitude toward various public policies and preference for candidates based on policies (Aragones and Palfrey, 2002). A study in sports and substance abuse (Pan and Baker, 1998) used MDPREF to investigate student athletes' perceptions of banned substances (alcohol, steroids) relative to their selected attributes (drowsiness, hair loss). The tourism industry has applied MDS programs, such as ALSCAL, to explore relationships between stimulus points (destinations) to property vectors or destination attributes (Kim, Guo, and Agrusa, 2005). Research in agriculture has estimated individual preference functions for food safety attributes in an attempt to segment consumers on their willingness to pay for safety standards utilizing such MDS techniques (Baker and Crosbie, 1993). Another study used these techniques to assess farmers' goals in relation to the decision being made, and understand the differences among groups of farmers (Patrick, Blake, and Whitaker, 1983). In the area of nutrition and food sciences, sensory studies employing these MDS methods have been used to evaluate and group species of fish and determine preferences for various product groups (Elmore et al., 1999). See Borg and Groenen (2005) for a survey of many other applications of such MDS procedures across a number of different domains of science.

Although MDS is popular in practice, there are a number of limitations associated with classical MDS procedures: (1) There is generally little basis for determining the dimensionality of the underlying space for the multidimensional representation of data. Some commercially available software restricts the analysis to two dimensions, while other approaches encourage the use of ad-hoc scree plots examining variance accounted for (VAF) or stress (S) versus the dimensionality; (2) For most procedures, it is possible to obtain point estimates only for parameters of interest which do not help to assess the stability of the MDS solution; (3) Many procedures require data preprocessing (row, column, or matrix centering or standardization) to avoid degenerate or non-informative solutions which can substantially affect the results obtained; (4) External information (e.g., brand attributes, consumer demographics) are commonly used in a post-hoc analysis for interpretation purposes only. It is more desirable to incorporate them directly into the analysis to derive the MDS solution.

Some Bayesian MDS models have recently been proposed to improve upon existing MDS procedures. DeSarbo et al. (1998) develop a Bayesian MDS model that estimates spatial market structures from pick-any/J choice data, provides for individual level parameters, and allows for correlations among the choice alternatives

across individuals. DeSarbo, Kim, and Fong (1999) propose a Bayesian formulation of a vector MDS procedure for the spatial analysis of binary choice data. Oh and Raftery (2001) provide a Bayesian metric MDS procedure for object configuration and a simple Bayesian criterion for dimension determination. Lee (2001) has employed the Bayesian Information Criterion (BIC) to determine the dimensionality of MDS representation for cognitive process models when similarity measures among stimuli are available. Martin and Quinn (2002) describe a dynamic Bayesian measurement model of ideal points for all justices serving on the U.S. Supreme Court from 1953 through 1999. Clinton, Jackman, and Rivers (2004) develop a Bayesian procedure for estimation and inference for spatial models of roll call voting. Jackman (2004) uses a Bayesian model for measuring graduate school applicant quality, combining the information in the committee members' ordinal ratings with the information in applicants' GRE scores. Gormley and Murphy (2007) analyze Irish election data using a Bayesian latent space ideal point model where both candidates and voters had positions in a latent space. Park, DeSarbo, and Liechty (2008) propose a Bayesian MDS model that combines both the vector model and the ideal point model in a generalized framework for modeling metric dominance data. Fong et al. (2010) offer a Bayesian vector MDS model to provide a joint spatial representation of ordered preference data. Furthermore, their proposed Bayesian procedure allows external information in the form of an intractable posterior distribution derived from a related data set to be incorporated as a prior in deriving the spatial representation of the preference data.

11.4.1 Bayesian Vector MDS Models

Preference data can be in the form of ratings, rankings, pick-any/J, etc. We start with ratings data and then explain how the Bayesian model can be adapted to handle other forms of preference data. Let y_{ij} denote the preference rating for consumer i $(i = 1, 2, ..., N)$ pertaining to j $(j = 1, 2, ..., J)$. We assume a vector MDS model for the data:

$$y_{ij} = \underline{a}_i' \underline{b}_j + e_{ij}, \tag{11.4.1}$$

where the T-dimensional latent consumer vector $\underline{a}_i$ and the T-dimensional latent brand vector $\underline{b}_j$ are assumed to be random and the error terms are independent and normally distributed, $e_{ij} \sim N(0, \sigma^2)$. Note that the model in (1) is under-identified. In particular, one can obtain the identical scalar products by multiplying $\underline{a}_i$ and $\underline{b}_j$ by a non-singular orthogonal matrix $\mathbf{M}$ as $\underline{a}_i' \mathbf{M}' \mathbf{M} \underline{b}_j = \underline{a}_i' \underline{b}_j$. To address the identification problem (see Gustafson, 2005), we assume informative proper priors (which may involve covariates like brand attributes and consumer demographics) for $\underline{a}_i$ and $\underline{b}_j$. For example, suppose A_{ik}, $k = 1, \ldots, K$, is the value of the k^{th} descriptor variable for consumer i and B_{jm}, $m = 1, \ldots, M$, is the value of the m^{th} attribute for brand j, and we let

$$\underline{a}_i' = \underline{\Theta}\underline{A}_i + \underline{\tau}_i, \text{ and } \underline{b}_j = \underline{P}\underline{B}_j + \underline{\delta}_j, \tag{11.4.2}$$

where $\underline{A}_i = (A_{ik})'$, $\underline{B}_j = (B_{jm})'$, $\underline{\Theta}$ and $\underline{P}$ are random coefficient matrices, $\underline{\tau}_i \sim N(\underline{0}, \underline{\Xi})$ and $\underline{\delta}_j \sim N(\underline{0}, \underline{\Sigma})$ are random error vectors. Proper hyper-priors are then assumed on σ^2, $\underline{\Theta}$, $\underline{P}$, $\underline{\Xi}$ and $\underline{\Sigma}$ (e.g., Brown, Vannucci, and Fearn, 1998). Using the Bayesian approach one can compute posterior interval as well as point estimates for various parameters of interest. Also, the probability based criterion, Bayes factor, can be evaluated to determine the optimal dimension T of the derived joint space map. Furthermore, variable selection can be performed to identify significant demographics and attribute variables that affect the consumer and brand vectors.

For ranking data, we introduce latent utilities and latent cutoff points to model the data (cf. Johnson and Albert, 1999). For each observation y_{ij}, we assume the existence of a latent utility z_{ij} and latent cutoff points $-\infty = \gamma_{i,0} < \gamma_{i,1} < \gamma_{i,2} \ldots < \gamma_{i,c-1} < \gamma_{i,c} = +\infty$ such that:

$$y_{ij} = c \text{ when } \quad \gamma_{i,c-1} \le Z_{ij} \le \gamma_{i,c}, i = 1,2,...,N, j = 1,...,J, c = 1,2,...,C. \tag{11.4.3}$$

Equation (11.4.1) then becomes:

$$Z_{ij} = \underline{a}_i' \underline{b}_j + e_{ij}^*, \tag{11.4.4}$$

where the error terms are independent and $e_{ij}^* \sim N(0,1)$ is assumed to eradicate an identification problem. Fong et al. (2010) performed a Bayesian factor analysis (e.g., Rowe, 2002; Lopes and West, 2004) on a related attribute ratings data set to develop an informative prior for $\underline{b}_j$. Specifically, let $\underline{x}_j$ be the K-dimensional (mean corrected) attribute rating vector for brand j. They assume a Bayesian factor model with

$$\underline{x}_j = \Lambda \underline{f}_j + \underline{\varepsilon}_j, \tag{11.4.5}$$

where $\underline{\varepsilon}_j \sim N(\underline{0}, \underline{\Psi})$ independently, Λ is $K \times T$ matrix of unobserved factor loadings, the factors $\underline{f}_j$ are independent among themselves as well as independent of the error terms, and $\underline{\Psi}$ is the $K \times K$ error variance matrix. The posterior distribution of the factors, $\pi_f(\underline{f}_1, ..., \underline{f}_J | \underline{X})$, is used as a prior for $(\underline{b}_1, ..., \underline{b}_J)$, or

$$p(\underline{b}_1, ..., \underline{b}_J | \underline{X}) = \pi_f(\underline{b}_1, .., \underline{b}_J | \underline{X}), \tag{11.4.6}$$

where $p(\underline{b}_1, .., \underline{b}_J | \underline{X})$ is the derived prior distribution and $\underline{X}$ is the attribute data matrix. Although the prior distribution in (11.4.6) is a high dimensional integral which cannot be evaluated in closed form, it is noted in Fong et al. (2010) that the prior can be expressed in a hierarchical form with the first stage prior $p(\underline{b}_1, .., \underline{b}_J | \Lambda, \Psi, \underline{X})$ which is a product of multivariate normal densities, and the second stage prior $p(\Lambda, \Psi | \underline{X})$, which can be sampled using MCMC methods. This observation greatly reduces the computational burden associated with the Bayesian analysis. If demographic information is available, proper priors as described in Equation (11.4.2) will be assigned to $\underline{a}_i$. To complete the prior specification, a vague prior for the individual level cutoff points in Equation (11.4.3) may be used.

Finally, for pick-any/J choice data, we introduce correlated latent utility variables as in DeSarbo, Kim, and Fong (1999) and employ an equation similar to (11.4.4) to relate the latent utility with the corresponding consumer and brand vectors. Here, Z_{ij} is specified such that choice j is observed (i.e., $y_{ij} = 1$) if the Z_{ij} is larger than a threshold parameter which may vary by individual or can be constant across consumers. Also, the utility error terms are assumed to be correlated for any fixed i, instead of independent. In addition, for identification purposes, a correlation matrix may be preferred over a covariance matrix in the specification of the error distribution. If a correlation matrix is assumed, the conventional Wishart prior assumption for the precision matrix will be inappropriate and more extensive calculation is expected to perform the Bayesian analysis. However, we note that the parameter expansion technique (e.g., Hobert and Marchev, 2008) can be applied to alleviate the computational burden in this case.

11.4.2 A Marketing Application

Suppose J brands are ranked by N consumers, MDS can be applied to the $N \times J$ data matrix to provide a joint spatial representation of the brands and consumers in a reduced dimensional space to address various positioning issues. Frequently, in a survey, respondents are asked to rank brands as well as to rate brand attributes. Thus, in addition to the preference data set, a related data set on attribute ratings is also obtained. Traditionally, classical factor analysis is used to analyze the attribute ratings to yield a "perceptual map" where the factor score matrix gives the location of each brand in the map. The procedure PREFMAP3 then introduces for each respondent either an ideal brand or a preference vector into the perceptual map in a manner that ensures maximal correspondence between the input preference ratings (or rankings) for the brands and the preference relationships among the brands in the resulting joint space map (cf. Lilien and Rangaswamy, 2004). However, there are two issues with this approach. One, the uncertainty around the factor score estimates is not incorporated into the analysis. Two, the method ignores possible contribution from the preference data set on the estimation of brand locations.

To overcome these two problems as well as other limitations associated with classical MDS procedures mentioned in the Introduction section, Fong et al. (2010) offer a Bayesian solution for the analysis of such data sets. They perform a Bayesian factor analysis on the attribute ratings data and use the posterior distribution of the factor scores as the prior for brand locations in the joint space map. To illustrate their methodology, they employ data from a December 2002 survey of consumers, who have plans to purchase sport utility vehicles (SUVs) within the next 6-12 months. Considering the top ten luxury SUV brands (Lexus RX300, Acura MDX, BMW X5, Mercedes Benz M-Class, Cadillac Escalade, Lincoln Navigator, Hummer H2, Land Rover Discovery, Lexus LX470, and Range Rover) and nine associated attributes (good vehicle for family use, good ride/handling off-road, built rugged and tough, technically advanced, prestigious, high trade-in value, excellent cargo space,

good interior passenger room, and good gas mileage) in that study, they obtain a three-dimensional solution as shown in Figure 11.4. Here the posterior means of the distributions for the brand coordinates and consumer vectors are plotted in pair-wise dimension fashion. (The consumer vectors are normalized to allow better visualization of the brand locations.) The three dimensions are labeled as follows. Dimension 1 is a *Practicality* dimension as it can be characterized as built rugged and tough, good ride/handling off road, and not a good vehicle for family use; Dimension 2 is a *Size* dimension as it can be described as excellent cargo space, good interior passenger room, and poor gas mileage; and Dimension 3 is a *Quality* dimension as it is characterized by technically advanced, prestigious, and high trade-in value. Using the Bayesian results, one can provide answers to various marketing research questions such as: What are the underlying dimensions that consumers utilize to form their consideration to buy? What specific competitive brands appear to be most threatening to a particular brand? How does intended consumer demand vary along these dimensions?

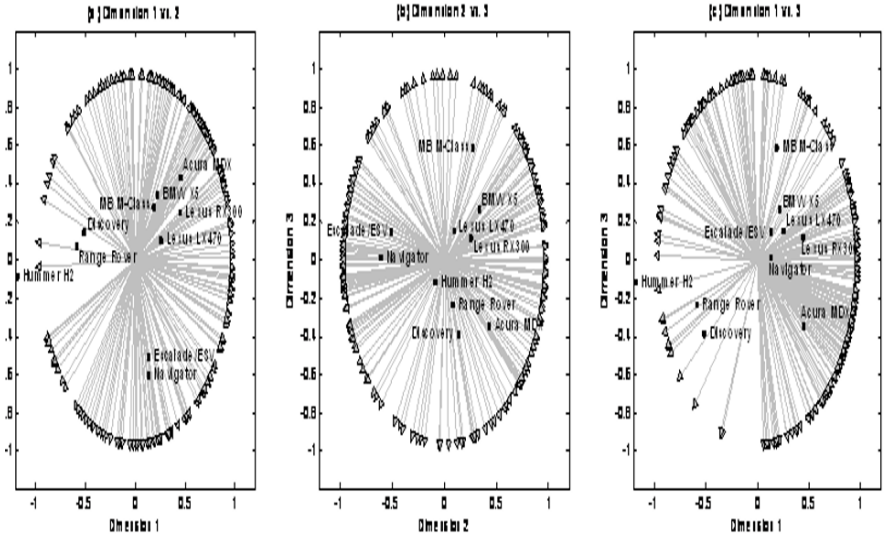


FIGURE 11.4. The plots of the posterior means of the distributions for the brand coordinates and consumer vectors.

The Bayesian vector MDS model is shown to outperform several traditional vector MDS approaches in Fong et al. (2010). To compare with other traditional MDS procedures, we report in Table 11.1 the Bayesian results as well as those from the popular ALSCAL unfolding MDS analysis for the luxury SUV data. The same measure of fit given in Fong et al. (2010) is used:

$$F_j = \left[1 - \frac{\sum_{i=1}^{N} |y_{ij} - \hat{y}_{ij}|}{N(C-1)}\right] \times 100\%$$

for each brand j where $\hat{y}_{ij}$ is the predicted response. As shown in Table 11.1, the Bayesian method provides a uniform improvement over all 10 brands in terms of the fit measure. The overall fit measure for the Bayesian vector MDS procedure is 83% versus 73% for ALSCAL. Furthermore, when ALSCAL is used to analyze the data, there is a more than 34% increase in the overall mean square errors (MSE).

TABLE 11.1. Measure of fit by brand for the two models.

Brand	**Bayesian Vector MDS**	**ALSCAL**
Land Rover Discovery	82%	78%
Range Rover	84%	81%
BMW X5	80%	69%
Mercedes Benz M-Class	83%	72%
Lincoln Navigator	85%	69%
Lexus LX 470	82%	80%
Lexus RX 300	80%	72%
Acura MDX	83%	67%
Hummer H2	86%	73%
Esscalade/ESV	83%	69%
Overall Fit	**83%**	**73%**
Overall MSE	**0.96**	**1.29**

11.4.3 Discussion and Future Research

A Bayesian approach to MDS offers a number of advantages over similar classical MDS procedures. Some of them are: (1) A probability based criterion such as Bayes factor can be used to determine the number of dimensions of the derived joint space map; (2) A Bayesian model can provide interval estimates for various parameters of interest which are useful to assess the stability of the MDS solution; (3) Data preprocessing is generally not needed to perform a Bayesian analysis; and (4) External information like statistical results from a related data set can be easily incorporated as prior input for the current study in a Bayesian analysis.

The Bayesian approach to MDS is promising and various extensions of existing work are possible. For example, segmentation is an important topic in marketing and one may want to modify the model in Fong et al. (2010) to perform the task. As shown in Figure 11.4, consumers (represented by vectors) are not homogeneous and it is useful to separate them into different groups (segments) for marketing purposes. Thus, if segmentation is one of the research goals, it will be desirable to obtain group memberships as part of the output (see DeSarbo et al., 2004). It is also desirable to predict market share for any product at any location on the joint space map in Figure 11.4. This can be done formally in a Bayesian analysis. To accommodate heterogeneity which is an important concept in marketing, one may explore the possibility of unknown and unequal error variances in (11.4.1). For ranking data,

in addition to changing the variance assumption for (11.4.4), one may impose an exchangeable prior on the cutoff points in (11.4.3). Finally, it is desirable to develop multi-mode, multi-way Bayesian models as well as dynamic Bayesian MDS procedures to analyze various types of data.

Acknowledgments: Dr. Fong's work was sponsored in part by a research grant from the Smeal College.

Chapter 12
Bayesian Categorical Data Analysis

Some interesting research challenges for Bayesian inference arise from binary and categorical data, including more traditional inference problems like contingency tables with sparse data and case-control studies as well as more recent research frontiers like non-standard link function for binary data regression.

12.1 Good Smoothing

James H. Albert

In the analysis of categorical data, a general problem is the handling of small counts. In the estimation of a proportion p, samples consisting of all successes or all failures are problematic. The corresponding proportions of successes, 1 and 0, are clearly unsatisfactory estimates at the proportion. If $\hat{p}$ is the sample proportion and n is the sample size, the standard frequency confidence interval (the Wald interval) of the form $(\hat{p} - 1.96SE, \hat{p} + 1.96SE)$, where $SE = \sqrt{\hat{p}(1-\hat{p})/n}$ is an unsatisfactory interval estimate. In the standard test of independence, students are very aware of the problems with tables with small observed counts (so-called *sparse* tables) and software programs routinely will give warnings for the use of the Pearson test of goodness of fit in these situations.

A simple ad-hoc solution to the small count problem is to simply add small counts to the data and apply frequentist methods to the adjusted data. In describing exploratory methods for fraction data, John Tukey in Tukey (1977) recommends "starting" counts by adding the constant 1/6 as a way of handling "none seen below" when it is possible to observe these counts. Agresti and Coull (1998) describe the good performance of the procedure that applies the Wald interval estimate to counts adjusted by adding two successes and two failures. (This procedure is commonly used and incorporated in introductory statistics texts such as Moore, 2006.) In the analysis of contingency tables, one solution to the small count problem is to simply

M.-H. Chen et al. (eds.), *Frontiers of Statistical Decision Making and Bayesian Analysis*, DOI 10.1007/978-1-4419-6944-6_12,

add a constant, say 1/2, to each cell of the table before applying standard inferential methods.

Adding imaginary counts to categorical data essentially corresponds to prior information incorporated into the analysis. So any discussion about the proper choice of imaginary counts leads one to naturally to a Bayesian analysis. One of the first people to think very seriously about the choice of imaginary counts in the smoothing of categorical data was I. J. Good. One goal of this section is to give a historical perspective on the Bayesian smoothing of tables by discussing Good's famous 1967 JRSS paper (Good, 1967). The topic of his paper, constructing a Bayesian test of equiprobability of a multinomial parameter, seems narrow in scope. But this paper gives a nice view on Good's approach to categorical data problems and his testing procedure is naturally linked with an appropriate choice of "flattening constant" in the estimation of a multinomial probability vector. In Section 12.1.2, we apply Good's general smoothing strategy for problems of smoothing of a single proportion, a 2 by 2 table, and a general two-way contingency table. For each problem, a two-stage prior is proposed where a conjugate prior is assigned at the first stage with unknown parameters, and a vague prior is assigned to the unknown parameters at the second stage. We illustrate the posterior estimates for a collection of examples and contrast these Bayesian smoothing estimates with frequentist methods.

In Section 12.1.3, we consider several data mining applications where the goal is to simultaneously learn about a large number of parameters. There is a challenge in simultaneously estimating a collection of hitting rates of baseball players due to the high variability of the rates for players with few opportunities to hit. By fitting an exchangeable model, one is smoothing the observed hitting rates towards a combined estimate, where the degree of shrinkage depends on the suitability of a model that assumes equal hitting probabilities. Two graphs are used to judge the suitability of this exchangeable model. In sports, there is much interest in the ability of a player to hit at games played at home relative to his ability to hit at games played away from home. An odds ratio can be used to measure the association between hitting and the venue and the problem is to simultaneously estimate the odds ratios for a collection of players. One is interested in smoothing the observed odds ratios towards a model that assumes that all players have the same hitting average at home versus away games.

12.1.1 Good's 1967 Paper

12.1.1.1 The Testing Problem and Priors

Good (1967) describes a Bayesian testing procedure for multinomial data. Suppose we observe the vector of counts $y = (y_1, y_2, \ldots, y_t)$ from a multinomial distribution with sample size n and cell probabilities $p = (p_1, p_2, \ldots, p_t)$. One wishes to test the hypothesis that the probabilities are equiprobable:

$$H: p_1 = p_2 = ... = p_t = 1/t$$

against the alternative hypothesis A that there are some differences in the values of $\{p_j\}$. The standard test procedure is the Pearson statistic

$$X^2 = \sum_{j=1}^{t} \frac{(y_j - n/t)^2}{(n/t)}$$

which has an asymptotic chi-square distribution with $t-1$ degrees of freedom. If X^2_{obs} is the observed values of X^2 and U is the chi-squared$(t-1)$ random variable, then one typically makes a decision about H on the basis of the p-value $P(U > X^2_{obs})$. The accuracy of the chi-square approximation is questionable for tables with small counts, so Good wishes to develop an "exact" Bayesian test that is free from the asymptotic theory and can be used with confidence for all values of t and n.

The Bayes factor against the null hypothesis H is the ratio of the marginal densities of y under the hypotheses A and H. Under the equiprobable hypothesis H, y is simply the multinomial distribution with cell probabilities $p_1 = p_2 = \cdots = p_t = \frac{1}{t}$ and so the marginal density of y is given by

$$m(y|H) = \frac{n!}{\prod_{j=1}^{t} y_j!}(1/t)^n.$$

Under the alternative hypothesis A, suppose p is assigned the proper prior $g(p)$. Then the marginal density of y under the alternative hypothesis is given by the integral

$$m(y|A) = \frac{n!}{\prod_{j=1}^{t} y_j!} \int \prod_{j=1}^{t} p_j^{y_j} g(p) dp,$$

and the Bayes factor against the null hypothesis is given by

$$BF = \frac{m(y|A)}{m(y|H)} = t^n \int \prod_{j=1}^{t} p_j^{y_j} g(p) dp.$$

Much of Good's paper is devoted to the construction of a "suitable" prior for the multinomial probabilities under the general hypothesis A. Here is an outline of Good's approach.

1. Good begins with a "sufficiency" postulate that is a modification of an argument by the philosopher W. E. Johnson. This postulate states that the prior for p is a linear combination of priors indexed by a parameter κ such that if we knew values of t, n, y_j and κ, then the knowledge of the other multinomial counts $\{y_k, k \neq j\}$ would have no effect on the posterior mean of p_j. Equivalently

$$E(p_j|y_j, t, n, \kappa) = E(p_j|y, t, n, \kappa).$$

2. Good that shows that this sufficiency postulate leads to a general expression for the posterior mean of p_j: $E(p_j|y_j,t,n,\kappa) = (y_j+k)/(n+tk)$, where the flattening constant k depends on the values of t and κ.
3. This expression for the posterior mean can be shown equivalent to the assumption that p follows a symmetric Dirichlet distribution with parameter k:

$$g(p|k) = \frac{\Gamma(tk)}{\Gamma(k)^t} \prod_{j=1}^{t} p_j^{k-1}.$$

   Let H_k denote the model that assumes p has this Dirichlet(k) distribution. The model H_k essentially says to adjust the multinomial counts by means of the flattening constant k to estimate the probabilities $\{p_j\}$. The maximum likelihood estimate y_k/n chooses $k = 0$, a uniform prior would correspond to $k = 1$, and a Jeffreys prior would choose $k = 1/2$.
4. Good argues that no specific symmetric Dirichlet distribution is adequate and places a distribution $\phi(k)$ on the hyperparameter k resulting in the prior distribution

$$g(p) = \int_0^\infty \frac{\Gamma(tk)}{\Gamma(k)^t} \prod_{j=1}^{t} p_j^{k-1} \phi(k) dk.$$

What does Good suggest for the second-stage prior density $\phi(k)$? The traditional noninformative prior of the form$\phi(k) = 1/k$ is inappropriate since the marginal density will not be defined with the use of an improper prior. Good wishes to use a proper distribution that approximates the density $1/k$, leading to the choice of log-Cauchy density

$$\phi(k) = \frac{1}{\pi k} \frac{\lambda}{\lambda^2 + \{\log(k/\mu)\}^2}, \; k > 0.$$

This prior density depends on two parameters μ and λ. The parameter μ is the median and λ is a scale parameter that is connected to the quartiles q_L and q_U by the relationship $\lambda = \log(q_U/\mu) = \log(\mu/q_L)$. In practice, Good recommends that μ are λ are chosen based on beliefs about the multinomial probability vector p when the equiprobable assumption is false. A useful parameter to think about is the *repeat rate*

$$\rho = \sum_{j=1}^{t} p_j^2.$$

The mean of ρ for a symmetric Dirichlet density is given by $E(\rho) = \frac{k+1}{tk+1}$. Good recommends specifying values of μ and λ by guessing at values of the quartiles of the repeat rate ρ. By setting these guesses to $(1+q_L)/(tq_L+1)$ and $(1+q_U)/(tq_U+1)$, one obtains estimates for q_L and q_U, which can used to get the log Cauchy parameters μ and λ.

12.1.1.2 Bayes Factors

Recall the model H_k was the assumption of a Dirichlet symmetric distribution with parameter k. As k approaches infinity, the model H_k approaches the equiprobable model H. The Bayes factor in support of H_k over H (or equivalently H_∞) is given by

$$BF(k) = \frac{m(y|A)}{m(y|H)} = t^n \frac{D(y+k)}{D(k)},$$

where $D(a) = \prod \Gamma(a_j)/\Gamma(\sum a_j)$ is the Dirichlet function evaluated at the vector $a = (a_1, a_2, \ldots, a_t)$. If the smoothing parameter k is assigned a density $\phi(k)$ over the alternative hypothesis A, then the Bayes factor in support of A over H is given by

$$BF = \int_0^\infty BF(k)\phi(k)dk.$$

Good believes it is useful to plot the Bayes factor $BF(k)$ as a function of the flattening constant k; it is analogous to plotting the likelihood function. One useful test statistic is the maximum of the Bayes factor over k: $BF_{max} = \max_k BF(k)$.

12.1.1.3 Smoothing Estimates

Although the focus of Good's paper is on the development of a Bayesian test procedure, this analysis gives different estimates of the flattening or "smoothing" estimate k. One can estimate k from the "Type-II likelihood" proportional to $BF(k)$; the Type-II maximum likelihood estimate is given by k_{max}. Alternatively, when the prior $\phi(k)$ is assigned to k, one can estimate k from its posterior distribution

$$\phi(k|y) \propto \frac{D(y+k)}{D(k)}\phi(k),\ k > 0.$$

12.1.1.4 An Example

To illustrate Good's methodology, we consider the counts of new visits to the website `http://bayes.bgsu.edu/bcwr` recorded during the week March 8 to March 14, 2009. Are visits more likely to occur during particular days of the week? Table 12.1 displays the number of hits during each of the seven days and we wish to test the hypothesis that the probabilities are equiprobable. From a frequentist perspective, the observed chi-square test statistic is 16.96. The p-value, the probability that a chi-square(6) exceeds 16.96, is equal to 0.0094, indicating substantial support to reject the equiprobable model. If we assume that H and A have equal prior probabilities, and the p-value is interpreted as the posterior probability of H, then the corresponding approximate Bayes factor (on the log 10 scale) is equal to $\log_{10} BF = -\log_{10}(0.0094) = 2.23$.

TABLE 12.1. Counts of the number of new visits to a particular website during a week in March, 2009.

Sun	Mon	Tue	Wed	Thu	Fri	Sat
14	25	16	11	22	12	6

Figure 12.1 displays values of $\log_{10} BF(k)$ plotted against values of $\log k$. This is a typical graph where the Bayes factor in support of the hypothesis A is small for small values of $\log k$, increases until it reaches a maximum value, and then approaches zero for large $\log k$. The Bayes factor is maximized at $\log k = 2.05$ and $\log_{10} BF(k_{max}) = 1.06$. This implies that the $\log_{10} BF(k)$ is smaller than 1.06 for all k, indicating that the evidence against the equiprobable hypothesis is much smaller than implied by the p-value where $\log_{10} BF = 2.23$. To compute the Bayes factor BF, one averages values of $BF(k)$ over the prior density $\phi(k)$. Suppose we assign k the standard log Cauchy density — the corresponding density on $\log k$ is displayed in Figure 12.1. The resulting value of the Bayes factor on the log 10 scale is given by $\log_{10} BF = 0.23$, indicating modest evidence against the equiprobable assumption.

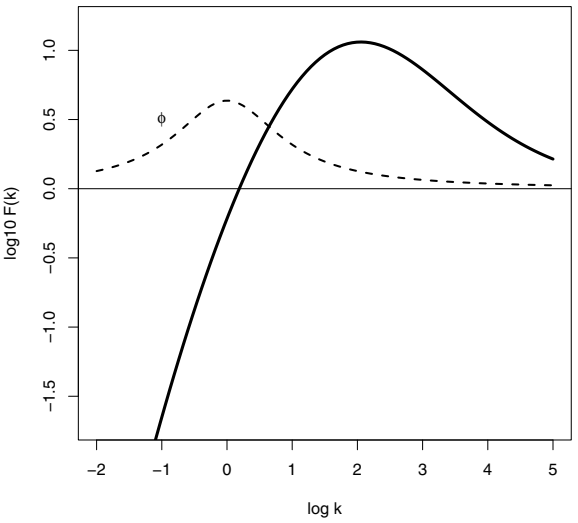


FIGURE 12.1. Logarithm of the Bayes factor against the equiprobable hypothesis plotted as a function of the logarithm of the hyperparameter k. The dotted curve represents the Cauchy prior on $\log k$.

As a by-product of this testing procedure, we get estimates at the flattening constant k. The type-II MLE, found by maximizing $BF(k)$ is given by $k_{max} =$

$\exp(2.05) = 7.8$ which corresponds to adding 8 to each of the counts. If one assigns a standard log Cauchy density to k, the posterior mode is given by $E(k|y) = 3.97$ which corresponds to adding 4 to each of the observed counts.

12.1.1.5 Bayes/Non-Bayes Compromise Statistics

Good makes a considerable effort to reconcile the Bayesian test statistics with classical measures of evidence. One interesting approximation relates the Bayes factor $BF(k)$ with the chi-square statistic X^2:

$$2\log BF(k) \approx X^2 + A(k) + O(n^{-1/2}),$$

where $\exp\{A(k)\} = \left[\Gamma(tk)^2(2\pi)^{t-1}\right] / \left[\Gamma(k)^{2t} t^{(2k-1)t} n^{t-1}\right]$. Another rule of thumb (developed from a number of earlier papers) is that the Bayes factor against H will typically falls within $1/(30P)$ and $3/(10P)$, where P is the tail-area probability (the p-value) for the analogous classical test statistic such as X^2. Good makes a conjecture that $BF(k)$ has at most one local maximum, and the probability that the maximum is $k = \infty$ is approximately equal to $1 - c_t$, where c_t is the probability a chi-square(t) statistic will exceed t.

Good provides many illustrations of the use of his Bayesian statistics for 18 examples which vary with respect to the number of components t and sample size n. For each example, the Bayes factor $BF(k)$ is presented for a range of values of k, and the integrated Bayes factor F is given for a variety of Type-III distributions $\phi(k)$. He also computes p-values for each example using a range of testing statistics. A "reliable" p-value P is computed for each example and he finds the empirical formula relating P with the integrated Bayes factor BF:

$$\frac{1}{6P\sqrt{2\pi n}} < BF < \frac{6}{P\sqrt{2\pi n}}.$$

12.1.1.6 General Features of Good's Approach

There are general notable aspects of Good's approach to the sparse multinomial data problem. First, smoothing a table of counts is directly related to a test of a hypothesized model. One wishes to smooth the observed counts towards fitted counts under the model and the degree of smoothing depends on the agreement of the data with the model. Specifically, the optimal choice of flattening constant is a function of the Bayes factor against the equiprobability hypothesis.

Good makes an effort to relate the Bayesian measures and associated estimates of the flattening constant with standard frequentist procedures. He wanted to develop "general all-purpose" methods that make sense and relate to classical methods. The posterior probability of the equiprobability model is related to the chi-square testing statistic and associated p-value. The Bayesian methods can suggest new methods such as the statistic F_{max} for measuring deviances from the null hypothesis. This

particular statistic is viewed as a "Bayesian/non-Bayesian compromise" since one uses a frequentist procedure, maximum likelihood, to summarize the marginal likelihood from a Bayesian model.

Good was one of the first people to advocate the use of hierarchical priors in Bayesian inference. He understood that it would be a bit arbitrary to assume a fixed first-stage prior such as a Dirichlet (k) to represent one's opinion about the multinomial proportion vector and advocates a distribution placed on k to reflect one's uncertainty about this parameter. One achieves a degree of robustness by the use of this second-stage parameter in that the Bayesian procedures are relatively insensitive to the choices of parameters at the second stage of the prior.

Last, Good discourages the use of noninformative priors and advocates priors that reflect one's beliefs about the problem. In the multinomial testing problem, Good describes the construction of a prior $\phi(k)$ that represents one's opinions about the flattening constant when the equiprobability assumption is false.

12.1.2 Examples of Good Smoothing

12.1.2.1 Estimating a Proportion

In this simple setting, one observes y from a binomial(n,p) distribution. In the case where y is observed to be 0 or n, the typical estimate y/n is undesirable and one wishes to move this estimate away from the boundaries of the parameter. Suppose one assigns p a beta(a,b) proportional to $p^{a-1}(1-p)^{b-1}$. It is convenient to reparameterize the shape parameters a and b by the mean $\eta = a/(a+b)$ and the precision $K = a+b$. If these hyperparameters were known, then the smoothed estimate of p is $(y+K\eta)/(n+K)$. What about the common situation where the hyperparameters are unknown?

TABLE 12.2. Estimates at (η, K) for all binomial samples for a sample size $n = 20$.

y	0	1	2	3	4	5	6	7	8	9	10
$\hat{\eta}$	0.12	0.29	0.33	0.36	0.38	0.40	0.42	0.44	0.46	0.48	0.50
$\hat{K}$	0.89	1.16	1.23	1.28	1.32	1.35	1.37	1.39	1.40	1.41	1.41
y	11	12	13	14	15	16	17	18	19	20	
$\hat{\eta}$	0.52	0.54	0.56	0.58	0.60	0.62	0.64	0.67	0.71	0.88	
$\hat{K}$	1.41	1.40	1.39	1.37	1.35	1.32	1.28	1.23	1.16	0.89	

We take η and K independent; the prior mean η is assigned the Jeffreys prior proportional to $\eta^{-1/2}(1-\eta)^{-1/2}$ and the precision K is assigned the standard log-Cauchy density used by Good. The posterior density of $(\text{logit}\,\eta, \log K)$ is given by

$$g(\text{logit}\,\eta, \log K|y) \propto \frac{B(K\eta+y, K(1-\eta)+n-y)}{B(K\eta, K(1-\eta))} \frac{\eta^{1/2}(1-\eta)^{1/2}}{(1+(\log K)^2)}.$$

If this posterior is summarized by the posterior mode, we obtain estimates for (η, K) that are displayed in Table 12.2 for all samples with sample size $n = 20$. Note that the estimate of η shrinks the observed proportion y/n towards 0.5. The estimates of K correspond approximately to "add a half count to the numbers of successes and failures."

Suppose we look at "add two successes and two failures" algorithm from a Bayesian perspective. The strategy is to add these pseudo counts to the data and apply a standard algorithm to the adjusted data. From a Bayesian perspective, this strategy is equivalent to assigning the proportion a beta(2, 2) prior and the posterior distribution is beta$(y+2, n-y+2)$. Suppose we observe $y = 0$ successes in a sample of size $n = 10$. Then the posterior density of p is beta(2, 12) which can be summarized by the 90% "equal-tails" interval estimate (0.028, 0.316).

But the choice of adding two successes and two failures to the data is arbitrary. From a Bayesian perspective, a preferable approach to allow for flexible smoothing by assigning the beta parameters (K, η) a second-stage distribution. Suppose $\log K$ is assigned a Cauchy density with location $\log 4$ and scale 1 — this reflects the belief that you want to add 4 observations to the data. Then η is assigned a beta prior with mean 0.5 and precision $K_0 = 80$ — this reflects a strong belief that you want to split the pseudo counts equally between successes and failures. For the same data $y = 0, n = 10$, the posterior median of $\log K$ is $\log 1.67$, indicating that 1.67 rather than 4 observations will be added to the data. A straightforward calculation gives a 90% interval estimate of (0.000, 0.336) which is substantially wider than the interval implied by the "add two successes and two failures" algorithm.

12.1.2.2 Smoothing a 2 by 2 Table

Suppose we observe independent binomial samples y_1 distributed binomial(n_1, p_1), y_2 distributed binomial(n_2, p_2), and we wish to smooth the counts in the two by two table $(y_1, n_1 - y_1; y_2, n_2 - y_2)$ in the case where small counts are observed. Suppose p_1, p_2 are assigned a common beta prior with mean η and precision K, which essentially adds the "prior counts" $(K\eta, K(1-\eta))$ to each row of the table.

We assign the pair (η, K) a vague prior reflecting little information about the location of these parameters. The parameters are assumed independent with prior mean η assigned the Jeffreys prior $\eta^{-1/2}(1-\eta)^{-1/2}$ and K assigned the standard log Cauchy density. In practice, one can assign $\log K$ a general log Cauchy density, where the location and scale are assessed based on knowledge about the association in the table.

The posterior estimates of K and η provide the smoothing of the table. The posterior estimate at the proportion p_1 can be represented approximately as

$$\hat{p}_1 = \frac{y_1}{n_1}\left(1 - \frac{\hat{K}}{n_1 + \hat{K}}\right) + \hat{\eta}\frac{\hat{K}}{n_1 + \hat{K}},$$

where $\hat{\eta}$ is a pooled estimate of the proportion when the table has the independence structure where $p_1 = p_2$, and the estimate $\hat{K}$ reflects the agreement of the observed counts with an independence structure. Generally larger estimates of K correspond to more "independent" tables and more shrinkage of the observed proportion estimate y_1/n_1 towards an independence estimate.

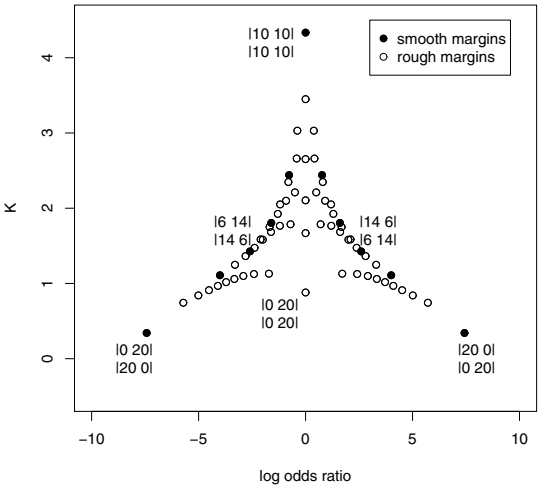


FIGURE 12.2. Estimates of the precision parameter K for all two by two tables, plotted as a function of the log odds ratio estimate of the table.

The posterior estimates of the precision parameter K were computed for all two by two tables where $n_1 = n_2 = 20$. Figure 12.2 displays the estimates $\hat{K}$ for all tables as a function of the log odds ratio estimate

$$\log\left[\frac{(y_1+0.5)/(n_1-y_1+0.5)}{(y_2+0.5)/(n_2-y_2+0.5)}\right]$$

in the table. In Figure 12.2, some of the points are labeled with the corresponding table of counts and the dark points correspond to tables with smooth row and column margins of (20, 20). To help understand this figure, six points are labeled with the corresponding two by two table. Generally, we see that the smoothing parameter K increases as the table moves towards an independence structure. For the table (20, 0; 0, 20) that is far from independence, the estimate of K is 0.34. In contrast, the estimate of K is between 3 and 4 for tables close to independence. Also, the estimate of K depends on the smoothness of the margins $(y_1+y_2, n_1+n_2-y_1-y_2)$. The points where the margins are the smooth values $(20, 20)$ are indicated by black plotting points. For a given independence structure (as measured by the log odds

ratio estimate), the estimate of K is an increasing function of the smoothness of the margins.

12.1.2.3 Smoothing a I by J Table

Suppose we observe cell counts $\{y_{ij}\}$ in a two-way contingency table that are distributed Poisson with respective means λ_{ij}. We are interested in smoothing the counts towards a hypothesized log-linear model $\log \lambda_{ij} = x_i\beta$. If one is interested in smoothing by adding the same pseudo-count to each cell, that would be equivalent to a uniform log-linear model of the form $\log \lambda_{ij} = \beta_0$. Alternatively, one may be interested in smoothing counts towards the model of independence $\log \lambda_{ij} = \beta_0 + u_i + v_j$.

One attractive way of model this smoothing assumes the λ_{ij} are independent Gamma$(\alpha, \alpha/\mu_{ij})$, where the prior means $\{\mu_{ij}\}$ satisfy the log-linear model $\log \mu_{ij} = x_i\beta$. Assume that α and β are independent with β distributed uniform and α distributed according to the log Cauchy density with location $\log \mu_\alpha$ and scale σ_α:

$$\phi(\alpha) = \frac{1}{\pi\alpha}\left(\frac{\sigma_\alpha}{\sigma_\alpha^2 + \{\log(\alpha/\mu_\alpha)\}^2}\right),\ \alpha > 0.$$

The posterior estimates at the expected counts are given by $\hat{\lambda}_{ij} = (y_{ij} + \hat{\alpha})/(1 + \hat{\alpha}/\hat{\mu}_{ij})$, where $\hat{\mu}_{ij}$ and $\hat{\alpha}$ are respectively posterior estimates at μ_{ij} and α. In this model, α plays the same role as the precision parameter K in the binomial/beta model; the estimate $\hat{\alpha}$ can be viewed as the number of pseudo-counts added to each cell of the table.

To illustrate the use of this smoothing model, Table 12.3 displays a crosstabulation of 72 student teachers who were rated by two supervisors from Bishop, Fienberg, and Holland (2007). One wishes to smooth the table to remove the one observed zero in the table. Suppose we apply the hierarchical model assuming the ratings of the two supervisors are independent. A standard log Cauchy density with location 0 and scale 1 is applied to the shrinkage parameter α.

TABLE 12.3. Crosstabulation of student teachers rated by two supervisors.

		Rating of Supervisor 2		
		Authoritarian	Democratic	Permissive
Rating of Supervisor 1	Authoritarian	17	4	8
	Democratic	5	12	0
	Permissive	10	3	13

In this example, there is a clear pattern of dependence in the table and one would anticipate only modest shrinkage of the counts towards independence. As expected, the posterior estimate at α is the small value $\hat{\alpha} = 1.84$ and the posterior estimates of the expected cell counts are displayed in Table 12.4.

TABLE 12.4. Bayes smoothed estimates at expected counts for student teacher example.

		Rating of Supervisor 2		
		Authoritarian	Democratic	Permissive
Rating of Supervisor 1	Authoritarian	16.3	4.8	7.9
	Democratic	5.5	10.2	1.3
	Permissive	10.2	4	11.8

In practice, one can assign a general Cauchy($\log \mu, \sigma$) prior to $\log \alpha$ that reflects prior knowledge about the smoothing parameter. For datasets such as this one that are not compatible with independence, the posterior estimates will be relatively insensitive to the choice of the hyperparameters $\log \mu$ and σ.

12.1.3 Smoothing Hitting Rates in Baseball

12.1.3.1 Introduction

In baseball, a variety of batting measures are collected for players. Some of the measures are helpful in understanding the abilities of the players and other statistics are largely affected by chance variation and are less useful in measuring player abilities. The jth player gets n_j opportunities to bat and one collects the number of "successes" y_j (different definitions of success are given below). If we have N players, we assume that $y_1, y_2, \ldots, y_N$ are binomial distributed with respective success probabilities $p_1, p_2, \ldots, p_N$.

In the usual exchangeable model, we assume that the probabilities $p_1, p_2, \ldots, p_N$ are a random sample from a beta distribution with parameters K and η. At the final stage, η and K are assumed independent; η is assigned the Jeffreys prior proportional to $\eta^{-1/2}(1-\eta)^{-1/2}$ and K is assigned a log logistic density of the form $g(K) = \frac{1}{1+K^2}$ for $K > 0$.

In the examples presented below, this model is fit to batting statistics for all players for a particular baseball season. There are substantial differences in the $\{n_j\}$ corresponding to the number of opportunities for part-time and regular players, making it difficult to interpret the success rates $\{y_j/n_j\}$. The posterior estimates of K and η are helpful in understanding the variation of abilities (the probabilities $\{p_j\}$) of the ballplayers. As will be seen below, the ability distributions have different spreads depending on the definition of success. By using the posterior estimates of K and η, we get flattened estimates of the batting probabilities — the degree of shrinkage will be greatest for the part-time players with a limited number of opportunities.

After the model is fit, one desires to look at residuals to detect general patterns of model misfit and to find particular players who have statistics that vary substantially from the general smoothing. We will illustrate two definitions of residuals. The predictive residuals are based on the posterior predictive distribution of the rate y_j/n_j; a standardized predictive residual is given by

$$r_j = \frac{y_j/n_j - \hat{\eta}}{\sqrt{1/n_j + 1/(\hat{K}+1)}},$$

where $\hat{\eta}$ and $\hat{K}$ are estimates from the posterior distribution.

A second residual is based on the definition of an outlier model. If the jth observation is unusual, then one would want to limit the shrinkage of the rate y_j/n_j towards the common mean. The beta parameter K controls the degree of shrinkage; for the "jth outlier" model that we denote by M_j, we suppose that the precision parameter for this component is given by CK, where C is a given constant smaller than 1 such as 1/2 or 1/3. Then one can compare the models "jth observation outlier" with the "no outlier model" M by use of the Bayes factor $BF_j = m(y|M_j)/m(y|M)$. Following Good, we express BF_j using a log10 scale.

In the examples, we illustrate plotting both the predictive residual r_j and $\log 10 BF_j$ as a function of the square root of the number of opportunities n_j.

12.1.3.2 Batting Averages, Home Run Rates, and In-Play Hit Rates

Batting data were collected for all 487 nonpitchers in the 2008 season. We collect the number of at-bats (AB), the number of hits (H), and the number of strikeouts (SO) for all batters.

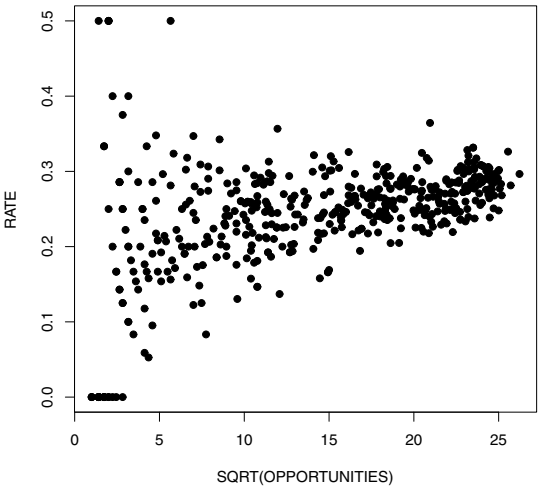


FIGURE 12.3. Plot of batting average against square root of at-bats for all nonpitchers in the 2008 baseball season.

The traditional measure of batting performance is the batting average $AVG = H/AB$. Figure 12.3 plots the observed batting average against the square root of the number of at-bats for all players. Note the high variability of the batting averages for small values of AB, indicating that the AVG is a relatively poor estimate of the players' true batting probability. It is certainly desirable to use an exchangeable model to get a better estimate at the batting probabilities for all players.

The exchangeable binomial/beta model was fit using a standard logistic prior for $\log K$. The estimates of the second-stage parameters are $\hat{K} = 379$ and $\hat{\eta} = 0.263$. The shrinkage fraction is given by $\hat{K}/(n_j + \hat{K}) = 379/(n_j + 379)$. For a player with 100 AB, the posterior mean estimate of the batting probability shrinks the observed AVG $379/(100+379) = 79\%$ towards the common estimate of 0.263. For a regular player with 500 AB, the Bayesian estimate shrinks the AVG $379/(500+379) = 45\%$ towards 0.263. Figure 12.4 displays the posterior mean estimates as a function of the square root of AB. We see substantial shrinkage, especially for part-time players with a small number of AB.

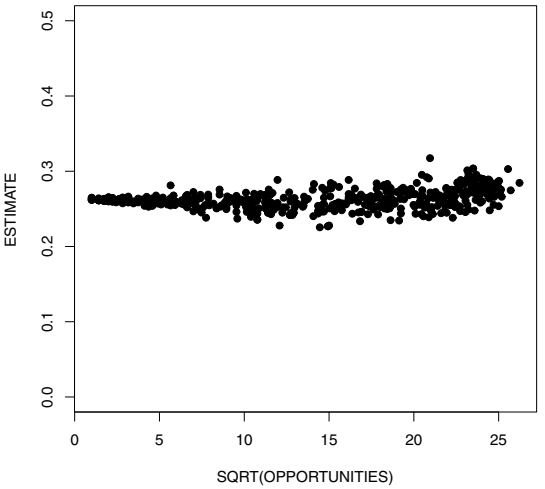


FIGURE 12.4. Plot of posterior mean estimates against square root of at-bats for all nonpitchers in the 2008 baseball season.

To check if this is a reasonable model, we look at residuals. Figure 12.5 plots the standardized predictive residuals as a function of $\sqrt{AB}$. Several outlying points stand out, including Chipper Jones who won the batting crown in 2008 with an average of 0.364. Also, by fitting a lowess curve, we see an increasing pattern in the residuals, suggesting that players with more at-bats tend to have higher batting averages. Figure 12.6 plots the log (base 10) Bayes factors for the "j-out" models; large values of the Bayes factors correspond to players who have averages that are

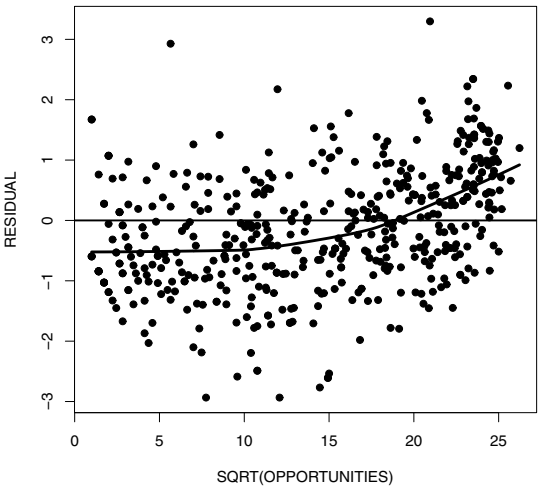


FIGURE 12.5. Plot of standardized predictive residuals against square root of at-bats for all non-pitchers in the 2008 baseball season.

not consistent with the fitted model. This graph of Bayes factors seems to do a better job than the predictive residual plot in distinguishing the players that had unusual batting averages.

One criticism of a batting average is that it is not a good estimate of the batting ability of a player. Some hits such as a home run clearly reflect the strength of a hitter and are ability-driven. Other types of hits, such as a ground ball hit between two fielders, seems to be more lucky and is not reflective of the ability or talent of the hitter. One way of providing support for these comments is to fit this binomial/beta model to alternative batting rates that may be less or more reflective of batting abilities than a batting average. One alternative measure of batting is the home run rate $HR/(AB-SO)$ — this is the proportion of home runs out all of the batting plays where the batter makes contact. Another measure of batting is the "in-play hit rate" $(H-HR)/(AB-SO-HR)$ — this is the proportion of hits when the batter makes contact (puts the ball in play) and does not hit a home run. Home run rates are generally thought to be ability-driven and the in-play hit rates are more driven by chance variation.

Table 12.5 displays the estimates for (η, K) when this binomial/beta model is fit to all batters using these alternative batting rates. The estimate for K for the home run rates is the small value $\hat{K} = 60.7$ indicating there is broad variability in the ability to hit a home run. The Bayesian estimates at the home run probabilities perform modest shrinkages of the observed rates $\{y_i/n_i\}$ towards the common rate of 0.0356. For example, the home run rate for a player with 500 opportunities is

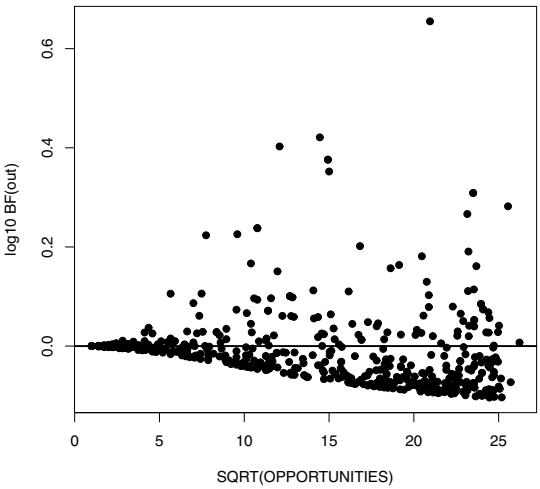


FIGURE 12.6. Plot of log Bayes factors for "j-out" models against square root of at-bats for all nonpitchers in the 2008 baseball season.

shrunk only 11% towards the overall mean. In contrast, the estimate of K for the in-play hit rates is $\hat{K} = 586$, reflecting that there are small differences in the players' abilities to get a hit from a pitch that is placed in-play. The corresponding probability estimates shrink the observed rates strongly towards the common rate of 0.3021. Even for a player who places 500 balls in play, his probability estimate shrinks the observed rate 54% towards the common rate.

TABLE 12.5. Estimates at (η, K) and representative shrinkages for three baseball hitting rates.

	Batting average	Home run rate	In play hit rate
Estimate at η	0.263	0.0356	0.3021
Estimate at K	379	60.7	586
Shrinkage, 100 opportunities	79%	38%	85%
Shrinkage, 500 opportunities	45%	11%	54%

12.1.3.3 Home versus Away Effects

In baseball, there is much interest in how players perform in different situations such as home and away games, left and right-handed pitchers, and "clutch" situations. Albert and Bennett (2003) describe that there are clear biases in particular situations. For example, batters tend to hit better against pitchers of the opposite arm. An in-

teresting question is whether players possess different abilities to perform better in particular situations.

For the jth player, one observes the numbers of hits s_{jH} and outs f_{jH} for home games, and the numbers of hits and outs s_{jA} and f_{jA} for away games. Assume that the hit numbers come from independent binomial distributions with probabilities p_{jH} and p_{jA}. Define the odds ratio and odds product of the probabilities

$$\alpha_j = \frac{p_{jH}/(1-p_{jH})}{p_{jA}/(1-p_{jA})},\ \beta_j = \frac{p_{jH}}{1-p_{jH}} \times \frac{p_{jA}}{1-p_{jA}}.$$

One is interested in simultaneously estimating the N odds-ratios $\alpha_1, \alpha_2, \ldots, \alpha_N$ corresponding to the association patterns in the associated N two by two tables.

Here we focus on estimating the odds ratios for the 195 regular players in the 2008 season who had at least 400 at-bats. Using the approximation described by Lindley (1964), the observed log odds-ratio

$$y_j = \log\left(\frac{s_{jH} f_{jA}}{f_{jA} s_{jH}}\right)$$

is normally distributed with mean $\theta_j = \log \alpha_j$ and variance $v_j = 1/s_{jH} + 1/f_{jH} + 1/s_{jA} + 1/f_{jH}$. Using the familiar exchangeable model, we assume the two-stage hierarchical model where $\theta_1, \theta_2, \ldots, \theta_N$ are a random sample from a normal distribution with mean μ and variance τ^2, and (μ, τ^2) are assigned the prior $g(\mu, \tau^2) = \frac{M}{M+\tau^2}$. Here we assign $M = 1000$, reflecting little knowledge about the value of τ^2. The posterior estimate of μ is $\hat{\mu} = 0.0746$, indicating that players tend to hit better at home. The posterior estimate of τ is $\hat{\tau} = 0.0769$, and applying standard normal/normal Bayesian calculations, the posterior estimate at the log odds ratio for the jth player is given by

$$\hat{\theta}_j = \frac{y_j/v_j + \hat{\mu}/\hat{\tau}^2}{1/v_j + 1/\hat{\tau}^2}.$$

Here the shrinkage values $(1/\hat{\tau}^2)/(1/v_j + 1/\hat{\tau}^2)$ range from 82% to 93%, indicating substantial shrinkage of the observed log odds-ratios towards a common value. Out of the 195 Bayesian log odds-ratio estimates, 194 are positive and half of the estimates fall between 0.058 and 0.090. The conclusion is that all regular players have approximately the same hitting advantage at home versus away games.

12.1.4 Closing Comments

Observed zeros in a table are problematic and there are many frequentist and Bayesian approaches for smoothing these counts to remove the observed zeros. The addition of imaginary counts corresponds to the use of prior information and the Bayesian approach is a natural way of deciding on the appropriate choice of these new counts. Good's strategy consists first of thinking of an appropriate model of

interest and developing a Bayesian test of the model hypothesis. The choice of optimal flattening constants is related to the Bayesian measure of goodness of fit of the model. The goal of this section is to review Good's contribution in the context of the problem of estimating a multinomial vector and apply Good's approach to other categorical data problems.

The use of hierarchical priors are very suitable for these smoothing problems. Prior information consists of a hypothesized model and the strength of the belief in this model and one can model this information by a hierarchical prior. The posterior estimates provide good smooths, where the degree of smoothing depends on the agreement of the data with the hypothesized model. The baseball examples demonstrate a similar benefit of the use of hierarchical priors for smoothing a collection of rates or a collection of odds-ratios.

12.2 Bayesian Analysis of Matched Pair Data

Malay Ghosh and Bhramar Mukherjee

In many studies, especially in biomedical applications, data are collected from *matched pairs*. For example, in case-control studies, cases may be matched with controls on the basis of demographic characteristics such as age, gender, ethnicity or other potential confounders. In other experimental instances, both elements of a matched pair may refer to the same subject, such as measurements on the left and right eyes, observations recorded at two time points under a typical pre-post study design, or responses under two treatments in a crossover experiment. This leads to data with fine degree of stratification. Often it is possible to summarize the data within each stratum in the form of a 2×2 table. The landmark paper by Mantel and Haenszel (1959) considered a series of s 2×2 tables of the following pattern:

Disease Status	Exposed	Not Exposed	Total
Case	n_{11i}	n_{10i}	n_{1i}
Control	n_{01i}	n_{00i}	n_{0i}
Total	e_{1i}	e_{0i}	N_i

Assuming a common odds ratio θ across strata, the Mantel-Haenszel (MH) estimator of the common odds ratio is

$$\hat{\theta}_{MH} = \frac{\sum_{i=1}^{s} n_{11i} n_{00i}/N_i}{\sum_{i=1}^{s} n_{01i} n_{10i}/N_i}.$$

Homogeneity of the odds ratios across tables is tested by

$$\sum_{i=1}^{s} \{n_{11i} - E(n_{11i}|\hat{\theta}_{MH})\}^2 / \mathrm{Var}(n_{11i}|\hat{\theta}_{MH}),$$

which follows an approximate χ^2 distribution with $s-1$ degrees of freedom under the null hypothesis. The derivation of the variance of the MH estimator posed a challenge, and was addressed in several subsequent papers (see Breslow, 1996 for details). Regression models for the analysis of such matched pairs data were introduced later by a number of authors in different contexts. In particular, Breslow et al. (1978) introduced such models for the analysis of matched case-control data.

This section will consider the analysis of matched pairs data specifically in the context of item response and case-control problems. This will be the content of the following subsections.

12.2.1 Item Response Models

Item response models were developed in educational testing to describe how the probability of a correct answer depends on the subject's ability and the question's level of difficulty. Specifically, the models assume that subject i has a parameter θ_i describing that subject's ability and question j has a parameter α_j such that its negative describes its level of difficulty $(i = 1,\ldots,n,\ j = 1,\ldots,k)$. The response X_{ij} denotes the outcome for the ith subject on the jth question, where $X_{ij} = 1$ for a correct response and 0 for an incorrect response.

Let $p_{ij} = P(X_{ij} = 1)$. The simplest and most widely quoted model for p_{ij} is the *Rasch model* (Rasch, 1961), for which

$$p_{ij} = \exp(\theta_i + \alpha_j)\,[1 + \exp(\theta_i + \alpha_j)]^{-1}. \tag{12.2.1}$$

This is also referred to as the *one-parameter logistic model*, since as a function of θ_i it has the form of the distribution function of a one-parameter logistic distribution with location parameter -α_j. A related popular model is the *probit model*, $p_{ij} = \Phi(\theta_i + \alpha_j)$, where Φ is the standard normal cumulative distribution function. These models are special cases of the generalized linear model

$$F^{-1}(p_{ij}) = \theta_i + \alpha_j$$

where the link function F^{-1} is the inverse of an arbitrary continuous distribution function.

A systematic development of item response theory from the classical point of view owes much to the pioneering work of Lord (e.g., Lord, 1953), Rasch (1961) and their colleagues. Among the many noteworthy contributions in the same vein are Andersen (1970) and Bock and Lieberman (1970).

Bayesian methods originally proposed for item response models were link-specific. For the Rasch model and its two-parameter extension, relevant work includes Birnbaum (1969), Owen (1975), Swaminathan and Gifford (1982, 1985), Leonard and Novick (1985), Mislevy and Bock (1984), Kim et al. (1994), and Tsutakawa and various co-authors listed in the references (e.g., Tsutakawa and Lin, 1986). However, these methods are primarily approximate Bayes due to analyti-

cally intractable posteriors. Albert (1992) conducted a full Bayesian analysis for the two-parameter probit model, but his parameter augmentation technique applies only to the probit link. A very general Bayesian approach to handle one-parameter item response models with arbitrary link functions were due to Ghosh et al. (2000b). They began with the likelihood:

$$L(\theta,\alpha)|x) = \prod_{i=1}^{n}\prod_{j=1}^{k}\left[F^{x_{ij}}(\theta_i+\alpha_j)\bar{F}^{1-x_{ij}}(\theta_i+\alpha_j)\right], \tag{12.2.2}$$

where $\bar{F} = 1-F$ and $x = (x_{11},\ldots,x_{1k}\ ,\ldots,x_{n1},\ldots,x_{nk})$. Define $t_i = \sum_{j=1}^{k} x_{ij}$, $i = 1,\ldots,n$ and $y_j = \sum_{i=1}^{n} x_{ij}$, $j = 1,\ldots,k$. Also, make the one-to-one parameter transformation $\eta_i = \theta_i + \alpha_k$ $(i = 1,\ldots,n)$, $\xi_j = \alpha_j - \alpha_k$ $(j = 1,\ldots,k-1)$. Write $\eta = (\eta_1,\ldots,\eta_n)'$, $\xi = (\xi_1,\ldots,\xi_{k-1})$. Then (θ,α) is one to one with (η,ξ,α_k) and the likelihood function given in (12.2.2) can be rewritten as

$$L(\eta,\xi,\alpha_k|x) = \prod_{i=1}^{n}\prod_{j=1}^{k-1}\left[F^{x_{ij}}(\eta_i+\xi_j)\bar{F}^{1-x_{ij}}(\eta_i+\xi_j)\right]\prod_{i=1}^{n}\left[F^{x_{ik}}(\eta_i)\bar{F}^{1-x_{ik}}(\eta_i)\right],$$

which is non-identifiable in α_k. However, the following theorem provides sufficient conditions under which the posterior can still be proper.

Theorem 12.1. *Suppose (i)* $0 < t_i < k$ *for all i, (ii)* $0 < y_j < n$ *for all j, and* $\int_{-\infty}^{\infty}|z|^{n+k-1}dF(z) < \infty$. *Then, under the prior* $\pi(\eta,\xi) \propto 1$, $\pi(\eta,\xi|x)$ *is proper.*

As an alternative prior, one considers a flat prior for α, but a multivariate t-prior for θ. The latter can be viewed also as $\theta|\sigma^2 \sim N(0,\sigma^2)$ and $\sigma^2 \sim \text{IG}(a,b)$, where IG refers to an inverse gamma distribution.

TABLE 12.6. Cell counts for cross-over study comparing treatments for relief of primary dysmenorrhea.

		Treatment A	
Treatment B	Treatment C	Relief	No Relief
Relief	Relief	8	45
Relief	No Relief	4	4
No Relief	Relief	7	9
No Relief	No Relief	3	6

To illustrate, consider the Bayesian approaches to item response models using Table 12.6. Frequentist analyses are given in Jones and Kenward (1987) and Agresti (1993). The data result from a three-period cross-over trial designed to compare placebo (treatment *A*) with a low-dose analgesic (treatment *B*) and high-dose analgesic (treatment *C*) for relief of primary dysmenorrhea. At the end of each period, each subject rated the treatment as giving either some relief (1) or no relief (2). Let

p_{ij} denote the probability of relief for subject i using treatment j $(j = A, B, C)$. The treatment effects are estimated for the logit, probit, and log-log links. The interest focuses specifically on the posterior means and standard deviations of treatment differences $\alpha_j - \alpha_k$. Results are displayed in Table 12.7 with a variety of choices of (a, b). The conclusions are fairly insensitive to the choice, and regardless of the prior one may conclude that treatments B and C are substantially better than placebo, with only mild evidence that C is better than B.

TABLE 12.7. Bayes treatment comparison estimates (standard errors in parentheses) for logit model with Table 12.6, using a variety of parameters (a, b) for t priors.

	Parameters for t Prior					
a	0.001	0.010	0.100	1.0	3.0	2.0
b	0.0	0.0	0.0	0.0	5.0	4.0
$\alpha_B - \alpha_A$	2.08	2.09	2.11	2.19	2.18	2.16
	(0.38)	(0.38)	(0.38)	(0.39)	(0.38)	(0.38)
$\alpha_C - \alpha_A$	2.62	2.64	2.66	2.75	2.74	2.72
	(0.42)	(0.43)	(0.43)	(0.43)	(0.41)	(0.41)
$\alpha_C - \alpha_B$	0.54	0.55	0.55	0.56	0.56	0.56
	(0.37)	(0.38)	(0.38)	(0.38)	(0.38)	(0.38)

12.2.2 Bayesian Analysis of Matched Case-Control Data

The frequentist backdrop. The contribution of statisticians to the development of case-control methodology is perhaps the most important contribution that they have made to public health and biomedicine. The central theme of a case-control study is to compare a group of subjects (cases) having the outcome (typically a disease) to a control group (not having the outcome or disease) with regard to one or more of the disease's potential risk factors. The method gained popularity in the 1920s for studying rare diseases, especially cancer, where following a healthy population or a cohort over time is impractical. In a case-control set-up, matching is often used for selecting "comparable" controls to eliminate bias due to confounding factors.

As mentioned in the introduction, while the Mantel-Haenszel (1959) approach was possibly the first in handling matched data, regression techniques for analyzing matched case-control data were first developed in Breslow et al. (1978). In the simplest setting, the data consist of s matched sets and there are M_i controls matched with a case in each matched set or stratum. We denote the i-th matched set by S_i, $i = 1, \cdots, s$. As before, one assumes a prospective stratified logistic disease incidence model, namely,

$$P(D=1|z,S_i) = H(\alpha_i + \beta' z), \tag{12.2.3}$$

where $H(u) = \{1+\exp(-u)\}^{-1}$ and α_i's are stratum-specific intercept terms. The stratum parameters α_i are eliminated by conditioning on the unordered set of exposures for the cases and controls in each stratum. This is equivalent to conditioning on the number of cases in each stratum which is a complete sufficient statistic for the nuisance parameters α_i. The generated conditional likelihood is free of the nuisance parameters and yields the optimum estimating function (Godambe, 1976) for estimating β. Assuming, without loss of generality, the first subject in each stratum is a case and rest of the subjects are controls, the derived conditional likelihood is

$$L_c(\beta) = \prod_{i=1}^{s} \frac{\exp(\beta' z_{i1})}{\sum_{j=1}^{M_i+1} \exp(\beta' z_{ij})}.$$

The above method is known as conditional logistic regression (**CLR**).

The classical methods for analyzing data from matched studies suffer from loss of efficiency when the exposure variable is partially missing. Lipsitz, Parzen, and Ewell (1998) proposed a pseudo-likelihood method to handle missing exposure variable. Rathouz, Satten, and Carroll (2002) developed a more efficient semiparametric method of estimation in presence of missing exposure in matched case control studies. Satten and Kupper (1993), Paik and Sacco (2000), and Satten and Carroll (2000) addressed the missingness problem from a full likelihood approach assuming a distribution of the exposure variable in the control population. Recently, McShane et al. (2001) proposed a conditional score method for estimating bias-corrected estimates of log odds ratio parameters in matched case control studies.

The above account of the frequentist development of matched case-control analysis is nowhere near complete and is beyond the scope of the current article. Here we only attempt to review the Bayesian contributions in this field. In spite of the enormously rich literature in the frequentist domain, Bayesian modeling for case-control studies did not start until the late 1980s. Bayesian analysis of case-control data seems particularly appealing with the rapid development of Markov chain Monte Carlo techniques. The possibilities include introduction of random effects, measurement error, missingness, flexibility to incorporate hierarchical structure, and existing prior information in modeling the relative risk/log odds-ratio parameters.

Altham (1971) is possibly the first Bayesian paper which considered several 2×2 contingency tables with a common odds ratio, though not in the case-control context, and performed a Bayesian test of association based on this common odds ratio. Later, Zelen and Parker (1986), Nurminen and Mutanen (1987), Marshall (1988) and Ashby, Hutton, and McGee (1993) considered a Bayesian approach to *unmatched* case-control problems involving a single binary exposure. Müller and Roeder (1997) considered bivariate continuous exposures with measurement error and proposed an elegant retrospective modeling of case-control data in a semiparametric Bayes framework. Müller et al. (1999) considered retrospective modeling with any number of continuous and binary exposures. Seaman and Richardson (2001) extended the binary exposure model of Zelen and Parker to any number of categorical expo-

sures. Gustafson, Le, and Vallee (2002) treated the problem of measurement errors in exposure by allowing both discrete and continuous exposures using a Dirichlet prior that only places support on the grid of possible exposure values. While all the above papers except Altham (1971) focused on unmatched data, Diggle, Morris, and Wakefield (2000) presented a more general regression based Bayesian analysis for situations when cases are individually matched to the controls. They considered matched data when exposure of primary interest was defined by the spatial location of an individual relative to a point or line source of pollution. The classical conditional likelihood was treated as a genuine likelihood for carrying out hierarchical Bayesian inference.

Ghosh and Chen (2002) developed general Bayesian inferential techniques for matched case-control problems in the presence of one or more binary exposure variables. The model considered was more general than that of Zelen and Parker (1986). Also, unlike Diggle, Morris, and Wakefield (2000), the analysis was based on an unconditional likelihood rather than a conditional likelihood after elimination of nuisance parameters. The general Bayesian methodology based on the full likelihood as proposed by Ghosh and Chen worked beyond the usual logit link. Their procedure included not only the probit and the complementary log links but also some new symmetric as well as skewed links. The propriety of posteriors was proved under a very general class of priors which need not always be proper. Also, some robust priors such as multivariate t-priors were used. The Bayesian procedure was implemented via Markov chain Monte Carlo.

To illustrate their general methodology, Ghosh and Chen considered the L.A. Cancer data as given in Breslow and Day (1980), and studied the effect of gall-bladder disease on the risk of endometrial cancer as a single exposure variable, and both gall-bladder disease and hypertension as multiple exposure variables. In the case of one case and one control, it was demonstrated that the Bayesian procedure could yield conclusions different from the conditional likelihood regarding association between the exposures and the disease. This is because, while the conditional likelihood ignores all concordant pairs, the Bayesian procedure based on the full likelihood utilizes these pairs as well.

Recently Sinha et al. (2004, 2005, 2007) have proposed a unified Bayesian framework for matched case-control studies with missing exposure data and a semiparametric alternative for modeling varying stratum effects on the exposure distribution. They considered s, $1:M$ matched strata with a disease indicator D, a vector of completely observed covariate Z and an exposure variable X with possible missing values. The missingness is assumed to be at random (MAR in the sense of Little and Rubin (2002)). A prospective model for the disease status is assumed, namely,

$$P(D_{ij}=1|X_{ij},Z_{ij},S)=H(\beta_0(S_i)+\beta_1'Z_{ij}+\beta_2X_{ij}),$$

where ij refers the jth individual in the ith stratum, $j=1,\cdots,M+1$ and $i=1,\cdots,s$. The distribution of the exposure variable X in a control population is assumed to be a member of general exponential family, namely

$$p(X_{ij}|Z_{ij},S_i,D_{ij}=0)=\exp[\xi_{ij}\{\theta_{ij}X_{ij}-b(\theta_{ij})\}+c(\xi_{ij},X)],$$

where the canonical parameter θ_{ij} is modeled in terms of the completely observed covariate Z_{ij} and stratum specific varying intercept terms, namely $\theta_{ij}=\gamma_{0i}+\gamma' Z_{ij}$. Here γ_{0i} are the nuisance parameters due to stratification. Sinha et al. describe two important lemmas which are useful to write down the likelihood.

Lemma 12.1. $\frac{\text{pr}(D_{ij}=1|Z_{ij},S_i)}{\text{pr}(D_{ij}=0|Z_{ij},S_i)}=\exp\left[\beta_0(S_i)+\beta_1' Z_{ij}+\xi_{ij}\{b(\theta^*_{ij})-b(\theta_{ij})\}\right]$, *where* $\theta^*_{ij}=\theta_{ij}+\xi_{ij}^{-1}\beta_2$.

Lemma 12.2. *Based on the above two model assumptions, one can derive the distribution of the exposure variable in the case population as:*

$$p(X_{ij}|D_{ij}=1,Z_{ij},S_i)=\exp[\xi_{ij}\{\theta^*_{ij}X_{ij}-b(\theta^*_{ij})\}+c(X_{ij},\xi_{ij})]. \tag{12.2.4}$$

Using the above two lemmas and assuming data missing at random, the prospective joint conditional likelihood of the disease status and the exposure variable given the completely observed covariate, the stratum effect and the conditional event that there is exactly one case in each stratum can be shown to be proportional to,

$$\begin{aligned}
&\prod_{i=1}^{s}P(D_i=1,D_{i2}=\cdots D_{iM+1}=0,\{X_{ij},\Delta_{ij}\}_{j=1}^{M+1}|S_i,\{Z_{ij}\}_{j=1}^{M+1},\sum_{j=1}^{M+1}D_{ij}=1)\\
&\propto\prod_{i=1}^{s}\Bigg\{P(D_{i1}=1,D_{i2}=\cdots=D_{iM+1}=0|\{Z_{ij}\}_{j=1}^{M+1}\sum_{j=1}^{M+1}D_{ij}=1,S_i)\\
&\quad\times p^{\Delta_{i1}}(X_{i1}|Z_{i1},D_{i1}=1,S_i)\times\prod_{j=2}^{M+1}p^{\Delta_{ij}}(X_{ij}|Z_{ij},D_{ij}=0,S_i)\Bigg\}\\
&=\prod_{i=1}^{s}\Bigg\{\frac{P(D_{i1}=1|Z_{i1},S_i)/P(D_{i1}=0|Z_{i1},S_i)}{\sum_{j=1}^{M+1}P(D_{ij}=1|Z_{ij},S_i)/P(D_{ij}=0|Z_{ij},S_i)}\\
&\quad\times p^{\Delta_{i1}}(X_{i1}|Z_{i1},D_{i1}=1,S_i)\times\prod_{j=2}^{M+1}p^{\Delta_{ij}}(X_{ij}|Z_{ij},D_{ij}=0,S_i)\Bigg\}.
\end{aligned}$$

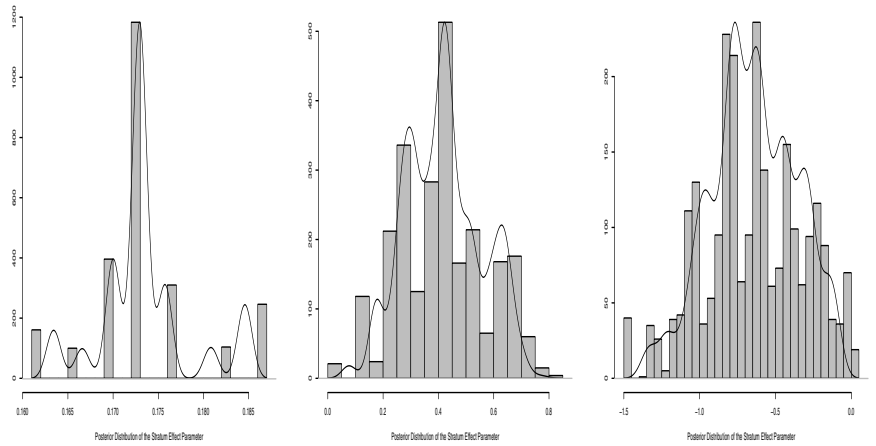
In the above expression it is assumed that the first observation in each stratum is the one coming from case population, Δ_{ij} is the missing value indicator and takes value 0 if X_{ij} is missing and 1 otherwise.

The parameters were estimated in a Bayesian framework by using a non-parametric Dirichlet Process prior on the stratum specific effects γ_{0i} in the distribution of the exposure variable and parametric priors for all other parameters. Since the number of these stratum specific parameters grow with the sample size, estimating these effects from a conditional MLE perspective may lead to the Neyman-Scott phenomenon. The novel feature of the Bayesian semiparametric method is that it can capture unmeasured stratum heterogeneity in the distribution of the exposure variable in a robust manner. The method is appealing as a unified framework to handle

missingness, measurement error and misclassification in the exposure variable. The method may also be extended to take into account the possible association schemes that may exist in a mixed set of continuous and categorical multiple exposures with partial missingness (Sinha, Mukherjee, and Ghosh, 2007). The examples and simulation results in these papers indicate that in presence of missingness, if association among exposures truly exists, one gains efficiency in estimating the relative risk parameters by modeling the stratum heterogeneity instead of ignoring it.

Sinha et al. (2005) reanalyzed the LA endometrial cancer data. They also reanalyzed a dataset related to equine epidemiology earlier analyzed in Kim, Cohen, and Carroll (2002) where the exposure was continuous. The data consisted of 498 strata with 1 : 1 matching in each stratum. Participating veterinarians were asked to provide data monthly for one horse treated for colic and one horse that received emergency treatment for any condition other than colic, between March 1, 1997, and February 28, 1998. A case of colic was defined as the first horse treated during a given month for signs of intra-abdominal pain. A control horse was defined as the next horse that received emergency treatment for any condition other than colic. Age was considered as a single exposure variable (X) measured on a continuous scale with one binary covariate (Z) indicating whether the horse experienced recent diet changes or not. For scaling purposes the authors linearly transformed age so that X was on the interval $[0, 1]$. Another example provided in Sinha et al. (2005) involves a matched case-control dataset coming from a low birth weight study conducted by the Baystate Medical Center in Springfield, Massachusetts. The dataset is discussed in Hosmer and Lemeshow (2000, Section 1.6.2) and is used as an illustrative example of analyzing a matched case-control study in Chapter 7 of their book. Low birth weight, defined as birth weight less than 2500 grams, is a cause of concern for a newborn as infant mortality and birth defect rates are very high for low birth weight babies. The data was matched according to the age of the mother. A woman's behavior during pregnancy (smoking habits, diet, prenatal care) can greatly alter the chances of carrying the baby to term. The goal of the study was to determine whether these variables were "risk factors" in the clinical population served by Baystate Medical Center. The matched data contain 29 strata and each stratum has one case (low birthweight baby) and 3 controls (normal birthweight baby). Figure 12.7 captures the essence of Bayesian non-parametric modeling of finer stratification effects that could affect the distribution of the missing exposure. The figure shows a sample draw from the posterior distribution of the stratum specific nuisance parameters, that are assumed to follow a Dirichlet Process prior. The flexible semiparametric Bayesian procedure adapts itself to the needs of the three examples and captures unmeasured heterogeneity in stratification effects in a data adaptive way. Whereas there is no stratification effect in the equine epidemiology example (a), in graphs labeled (b) and (c), the LA cancer data and the low birthweight data show substantial variability in the distribution of the stratification effects (Figure 12.7).

Sinha, Mukherjee, and Ghosh (2004) extended the Bayesian semiparametric methods to a matched case-control setting when there is finer subclassification within the cases and applied the method to the low birthweight data under a polytomous logistic regression model. They divided the cases into two categories, very



(a) Equine data (b) LA Cancer data (c) Low Birthweight data

FIGURE 12.7. Posterior Density of stratum specific nuisance parameters for the Three Examples discussed in Sinha et al. (2005). The histogram is based on 5,000 generated values from the posterior distribution of stratification parameters whose distribution is governed by a Dirichlet Process prior. The dashed line corresponds to the density of the base measure of the Dirichlet process prior, which was normal(0, 10) in all examples.

low (weighing less than 2000 gms) and low (weighing between 2000 to 2500 gms) and tried to assess the impact of smoking habits of mother on the chance of falling in the two low birth-weight categories relative to the baseline category (normal birth-weight, weighing more than 2500 grams). Presence of uterine irritability in mother and mother's weight at last menstruation period were considered as relevant covariates. It was noted that smoking mothers had a higher relative risk of having a low birth weight child when compared to a non-smoking mother. However, the risk of having a *very* low birth weight child did not depend on smoking significantly.

For this multi-category analysis with K categories of the disease state, the conditional probabilities of the disease variable given the covariate, exposure and the stratum are given by,

$$P(D_{ij} = k|S_i, \mathbf{Z}_{ij}, X_{ij}) = \frac{\exp\{\beta_{0k}(S_i) + \beta'_{1k}\mathbf{Z}_{ij} + \beta_{2k}X_{ij}\}}{1 + \sum_{r=1}^{K} \exp\{\beta_{0r}(S_i) + \beta'_{1r}\mathbf{Z}_{ij} + \beta_{2r}X_{ij}\}} \text{ for } k = 1, \cdots, K.$$

The binary variable representing smoking status is assumed to follow a Bernoulli distribution. In their example $K = 2$. Three analyses were performed, namely Bayesian Semiparametric (BSP), Parametric Bayes (PB) assuming constant stratum effect coming from a single distribution and the third one is iid parametric (PBV) where they assumed all the stratum effects are different and each of the n stratum

effect parameters γ_{0i}, $i = 1, \cdots, n$ are coming from n independent normal distributions.

TABLE 12.8. Analysis of low birth weight data with two disease states, using full dataset. BSP stands for Bayesian semiparametric method whereas PBC and PBV stand for parametric Bayes methods assuming constant and varying stratum effects respectively.

		BSP			PB			PBV		
Logit	Parameter	Mean	SD	HPD Interval	Mean	SD	HPD Interval	Mean	SD	HPD Interval
1	SMOKE	1.42	0.60	(0.33 ,2.72)	1.26	0.56	(0.25 ,2.50)	1.48	0.65	(0.26,2.08)
	LWT	-0.86	1.39	(-3.78 ,1.81)	-1.03	1.35	(-3.58 ,1.86)	-0.73	1.36	(-3.40,2.01)
	UI	0.15	0.67	(-1.27 ,1.46)	0.10	0.67	(-1.19 ,1.52)	0.18	0.67	(-1.14,1.52)
2	SMOKE	0.37	0.83	(-1.35 ,2.05)	0.23	0.66	(-1.10 ,1.54)	0.38	0.85	(-1.30,2.17)
	LWT	-0.52	1.61	(-3.76 ,2.52)	-0.55	1.59	(-3.73 ,2.41)	-0.55	1.62	(-3.65,2.79)
	UI	1.81	0.83	(0.18 ,3.51)	1.78	0.83	(0.30 ,3.59)	1.81	0.87	(0.27,3.72)

Table 12.8 contains the posterior means, posterior standard deviations and 95% HPD credible intervals for the parameters of interest under the proposed Bayesian Semiparametric method (BSP) and the parametric Bayes (PB the PBV) methods as discussed before. The analysis indicates that smoking of mother is a significant risk factor for low birth-weight category (category 1) and is not very significant in the *very* low birth-weight category (category 2). UI on the other hand shows an opposite association, showing significance in category 2 and almost no significance in category 1. LWT does not seem to be a significant covariate in any of the categories. The BSP and the PBV methods are in closer agreement whereas the PB estimates show some numerical differences. Obviously, without the finer classification into two weight categories, the fact that smoking is not so significant for category 2 and UI is appreciably significant for category 2 cannot be concluded from looking at the overall analysis. The work of Sinha et al. (2004, 2005, 2007) has contributed to recent advances in Bayesian methodology for matched case-control data with complex exposure and outcome structures.

12.2.3 Some Equivalence Results in Matched Case-Control Studies

An interesting class of theoretical and foundational issues have recently been studied for matched pair data that we describe below.

12.2.3.1 Equivalence of Retrospective and Prospective Analysis

Prentice and Pyke (1979) showed that if the disease risk is modeled by logistic regression function, and the subjects are selected into the study irrespective of their exposure value, prospective and retrospective analysis of the case-control data yield

the same estimate of the association parameter β. Moreover, the asymptotic standard errors of the estimate are the same in both the methods. However, the intercept parameter of the prospective model of the disease risk is not identifiable in a retrospective likelihood unless we know the disease prevalence in the population. Generally, the prospective model involves fewer parameters, and hence is easy to implement. Prentice and Pyke (1979) provided the theoretical validity for prospective analysis of data collected retrospectively.

Roeder, Carroll, and Lindsay (1996) proved a similar result when exposure variables are measured with error. They showed that the profile likelihood function of the association parameter obtained from a retrospective likelihood is the same as the one obtained through the joint distribution of the disease variable, true exposure variable, and its error prone surrogate variable. Carroll, Wang, and Wang (1995) extended Prentice-Pyke type results to situations with missingness and measurement error in the exposure variable.

Recently, Seaman and Richardson (2004) proved a Bayesian analogue of the Prentice-Pyke equivalence result. They considered a discrete exposure vector X with J support points $\{z_1,\ldots,z_J\}$. Let n_{0j} and n_{1j} be the number of cases and controls having the exposure value $X = z_j,\ j = 1,\cdots,J$. Now if $\text{pr}(X = z_j|D = 0) = \theta_j/\sum_{j=1}^{J}\theta_j$, and odds of disease associated with $X = x$ is $\exp(\beta' x)$, then the natural retrospective likelihood for the case-control data is

$$L_{MR} = \prod_{d=0}^{1}\prod_{j=1}^{J}\Big\{\frac{\theta_j \exp(d\beta' z_j)}{\sum_{k=1}^{J}\theta_k \exp(d\beta' z_k)}\Big\}^{n_{dj}}.$$

If one assumes that the data came from a cohort study, then the natural prospective likelihood is

$$L_{MP} = \prod_{j=1}^{J}\prod_{d=0}^{1}\Big\{\frac{\alpha^d \exp(d\beta' z_j)}{\sum_{k=0}^{1}\alpha^k \exp(d\beta' z_j)}\Big\}^{n_{dj}},$$

where α is the baseline odds of disease when exposure $X = \mathbf{0}$. These authors proved a Prentice-Pyke type equivalence result in the Bayesian context for a specific class of priors.

The equivalence result of Seaman and Richardson (2004) is a significant contribution to the Bayesian literature on case-control studies. An extension of their results to matched case-control studies involving missingness is due to Ghosh, Zhang, and Mukherjee (2006). We now present their results.

Suppose there are I strata where each stratum has s cases and t controls in a stratified case-control study. Let S_i denote the i-th stratum. Let $D_{ij}(= 1 \text{ or } 0)$ correspond to the presence or absence of a disease for the jth individual in ith stratum, and let x_{ij} denote the vector of discrete exposure variables for the jth observed subject in the ith stratum. We assume that each x_{ij} can take one of the K possible values $\{z_1,\cdots,z_K\}$. Suppose now

$$P(D_{ij}=1|X_{ij}=z_k,S_i)=\frac{\alpha_i\exp(\beta' z_k)}{1+\alpha_i\exp(\beta' z_k)},$$
$$P(X_{ij}=z_k|D_{ij}=0,S_i)=\frac{\gamma_{ik}}{\sum_{l=1}^{K}\gamma_{il}}.$$

Also, let $P(X_{ij}=z_k|D_{ij}=0,S_i)=\gamma_{ik}$. An application of a formula due to Satten and Kupper (1993) now yields $P(X_{ij}=z_k|D_{ij}=1,S_i)=\frac{\gamma_{ik}\exp(\beta' z_k)}{\sum_{l=1}^{K}\gamma_{il}\exp(\beta' z_l)}$.

Let Δ_{ij} denote the missingness indicator for the ith stratum (0 indicating missingness) with

$$P(\Delta_{ij}=1|S_i)=1-P(\Delta_{ij}=0|S_i)=\eta_i.$$

Let $\eta=(\eta_1,\cdots,\eta_I)'$. With the missing completely at random assumption, η_i does not depend on the parameters γ_{ik}, α_i or β.

Let $y_{idk}=\sum_{j=1}^{s+t}\{I[X_{ij}=z_k]I[D_{ij}=d]I[\Delta_{ij}=1]\}$, $d=0,1$, i.e, y_{i0k} and y_{i1k} are the respective numbers of non-diseased and diseased subjects having $X=z_k$ in the ith stratum, and I denotes the usual indicator function. Now, the prospective likelihood is

$$L_P=\prod_{i=1}^{I}\prod_{j=1}^{s+t}\Big[P(D_{ij}|x_{ij},S_i)\Big]^{\Delta_{ij}}=\prod_{i=1}^{I}\prod_{d=0}^{1}\prod_{k=1}^{K}\Big[\frac{\alpha_i^d\exp(d\beta' z_k)}{\sum_{l=0}^{1}\alpha_i^l\exp(l\beta' z_k)}\Big]^{y_{idk}},$$

and the retrospective likelihood is

$$L_R=\prod_{i=1}^{I}\prod_{d=0}^{1}\prod_{k=1}^{K}\Big[\frac{\gamma_{ik}\exp(d\beta' z_k)}{\sum_{l=1}^{K}\gamma_{il}\exp(d\beta' z_l)}\Big]^{y_{idk}}.$$

We now have the following equivalence theorem.

Theorem 12.2. *Suppose $\sum_{k=1}^{K}y_{i1k}\geq 1$ and $\sum_{k=1}^{K}y_{i0k}\geq 1$, for all $i=1,\cdots,I$. Assume mutually independent priors for the α_i, γ_{ik}, η and β, where $p(\alpha_i)\propto\alpha_i^{-1}$, $p(\gamma_{ik})\propto\gamma_{ik}^{-1}$, while η and β have proper priors $\pi_1(\eta)$ and $\pi_2(\beta)$. Then the posterior distribution of β derived from the prospective likelihood is equivalent to that from the retrospective likelihood.*

However, these results are limited in the sense that they only apply to certain specific types of prior. The equivalence results for any other type of prior are still an open question. Another limitation is that the exposure variable is assumed to be discrete. For continuous exposure Seaman and Richardson recommended categorization which is not an ideal solution.

12.2.3.2 Equivalence between Conditional and Marginal Likelihood for Analyzing Matched Case-Control Data

The disease risk model (12.2.3) for matched case-control study involves a set of nuisance parameters α_i's which capture the stratification effect on the disease risk. A natural way of analyzing matched case-control data is the conditional likelihood approach, where one considers conditional likelihood of the data conditioned on an

approximately ancillary statistic for the parameter of interest (see e.g. Godambe, 1976). In a matched case-control study we condition on the complete sufficient statistic for the nuisance parameters α_i, namely, the number of cases in each matched set. Alternatively, one can work with a marginal likelihood obtained by integrating the nuisance parameters over a mixing distribution, say G. Rice (2004) showed that the full conditional likelihood can exactly be recovered via the marginal likelihood approach by integrating the nuisance parameter with respect to a particular mixing distribution G. He derived the necessary and sufficient conditions on the class of distributions G for the two approaches to agree. The conditions invoke certain invariance properties of the distribution G and such invariant distributions are shown to exist under certain mild natural condition on the odds ratios. In the light of this agreement, in a Bayesian framework, the posterior distribution of the disease-exposure association parameter, β, of a matched case-control study using an invariant mixture distribution as a prior on the nuisance parameter and a flat prior on the association parameter, is proportional to the conditional likelihood. Rice (2008) extends the equivalence results to multiple covariates. This foundational argument is an important advance in Bayesian methods for matched pair data that provides a validation of using informative priors together with conditional likelihood as a basis of Bayesian inference (as proposed in Diggle, Morris, and Wakefield (2000), without a formal justification).

12.2.4 Other Work

Giron, Martinez, and Moreno (2003) considered some matched pair models which in the case control framework correspond to binary response and binary exposure. These models can accommodate some nuisance parameters. They considered location and scale families of distributions. Among other things, they found priors which provide the same posteriors for the parameters of interest both with full data as well as data based on individual paired differences.

Rice (2003) considered a matched case-control analysis where a binary exposure is potentially misclassified. His method can be interpreted in a Bayesian framework with prior information incorporated on odds ratios, misclassification rates and other model parameters and is tied to the equivalence theory of Rice (2004). Prescott and Garthwaite (2005) presented Bayesian methods for analyzing matched case-control studies in which a binary exposure variable is sometimes measured with error, but whose correct values have been validated for a random sample of the matched case-control sets. Three models are considered. Model 1 makes few assumptions other than randomness and independence between matched sets, while Models 2 and 3 are logistic models, with Model 3 making additional distributional assumptions about the variation between matched sets. With Models 1 and 2 the data are examined in two stages. The first stage analyses data from the validation sample and is easy to perform; the second stage analyses the main body of data and requires MCMC methods. All relevant information is transferred between the stages by using the pos-

terior distributions from the first stage as the prior distributions for the second stage. With Model 3, a hierarchical structure is used to model the relationship between the exposure probabilities of the matched sets, which gives the potential to extract more information from the data. All the methods that are proposed are generalized to studies in which there is more than one control for each case. The Bayesian methods and a maximum likelihood method are applied to a data set for which the exposure of every patient was measured using both an imperfect measure that is subject to misclassification, and a much better measure whose classifications may be treated as correct. To test the methods, the latter information was suppressed for all but a random sample of matched sets. Rice (2006) provided a nice comparison of the full likelihood based method of Rice (2003) to the two-stage approach presented by Prescott and Garthwaite (2005). Liu et al. (2009) proposed a Bayesian adjustment for the misclassification of a binary exposure variable in a matched case-control study. The method admits a priori knowledge about both the misclassification parameters and the exposure-disease association. The standard Dirichlet prior distribution for a multinomial model is extended to allow separation of prior assertions about the exposure-disease association from assertions about other parameters. The method is applied to a study of occupational risk factors for new-onset adult asthma.

Mukherjee, Liu, and Sinha (2007) again considered matched case-control studies where there is one control group, but there are multiple disease states with a *natural ordering* among themselves. They adopted a cumulative logit or equivalently, a proportional odds model to account for the ordinal nature of the data. The important distinction of this model from a stratified dichotomous and polychotomous logistic regression model is that the stratum specific parameters cannot be eliminated in this model via the conditional likelihood approach. They considered several choices for handling the stratum-specific nuisance parameters, appealing to the literature available for handling stratified ordinal response with the proportional odds model, including a random effects and a Bayes approach. They pointed out difficulties with some of the standard likelihood-based approaches for the cumulative logit model when applied to case-control data. A simulation study compared the different ordinal methods with methods ignoring sub-classification of the ordered disease states. Ahn et al. (2009) illustrated Bayesian analysis of matched case-control data with multiple disease states using the stereotype regression model.

12.2.5 Conclusion

In this section we have focused on Bayesian modeling of data arising in case-control studies as well as educational testing data requiring some item response analysis. The Bayesian paradigm offers a great deal of flexibility to accommodate unusual, unorthodox data situations and incorporating prior information on risk related parameters but comes with certain computational challenges. The popularity and use of these methods is highly dependent on developing user friendly softwares for implementing the analysis.

Item response theory has been the cornerstone of research in psychometry and educational statistics. For example, the Educational Testing Service in Princeton, New Jersey needs a constant evaluation and assessment of questions they put on different placement tests. If most of the students do very poorly in an exam, then the items in the exam are generally too difficult, and may not be helpful in assessing the real ability of students. A similar problem arises if the majority of students are able to answer all questions correctly. Thus an ideal placement test needs to have a broad range of questions with varying difficulty levels. Moreover, even if the test is effective in measuring the students' abilities, the next concern is whether the individual items on the test are effective in discriminating among students of different abilities. The item response models described in this review are quite widely used in assessing the difficulty levels of questions, but two-parameter item response models which contain "discrimination parameters" in addition to subject and item parameters also require a comprehensive Bayesian analysis. Swaminathan and Gifford (1985), and later Patz and Junker (1999a, 1999b) provided a Bayesian analysis for the logistic model, but a more general hierarchical Bayesian model generalizing Ghosh et al. (2000a) seems to be a good topic for future research.

There are other issues besides analysis, like that of Bayesian variable selection (Raftery and Richardson, 1996), sample size determination (De Santis, Perone Pacifico, and Sambcini, 2001) both for item response and case-control studies which are not discussed in this article but are interesting in their own right.

Matched pair studies bear enough promise for further research, both methodological and applied. One potential area of research is to extend the methodology when longitudinal data are available for both cases and controls. Hierarchical Bayesian modeling, because of its great flexibility, should prove to be a very useful tool for the analysis of such data. A second important area of research is to adapt the method for the analysis of stratified survival data. Some work in this regard has been initiated in a frequentist framework by Prentice and Breslow (1978), but Bayesian inference for such problems seems to be largely unexplored. For example the methods could potentially be adapted to bivariate survival models in family based study design (see Oakes, 1986, 1989; Shih and Chatterjee, 2002).

Finally, genetic case-control study is a new emerging area of research where one of the objectives is to study the association between a candidate gene and a disease. In many such instances, the population under study is assumed to be homogeneous with respect to allele frequencies, but comprises subpopulations that have different allele frequencies for the candidate gene (see Satten, Flanders, and Yang, 2001). If these subpopulations also have different disease risks, then an association between the candidate gene and disease may be incorrectly estimated without properly accounting for population structures. Zhang et al. (2006) present a Bayesian model averaging approach to handle population stratification. Accounting for population stratification may also pose some interesting statistical challenges in case-control studies of disease-gene or disease-gene-environment association and the Bayesian route remain to be explored.

A Bayesian approach could also be useful to analyze data coming from other designs used in an epidemiologic context (Khoury, Beaty, and Cohen, 1993). Family

based association studies also lead to the issue of dealing with family specific random effects that are very weakly identifiable in the likelihood. Zhang et al. (2009) consider a Bayesian semiparametric approach to family based genetic association study which deals with family specific random effects in a non-parametric manner. Another frequently used design is a nested case-control design (Wacholder, 1996) where a case control study is nested within a cohort study. This type of study is useful to reduce the cost and labor involved in collecting the data on all individuals in the cohort as well as to reduce the computational burden associated with a time-dependent explanatory variable. Unlike the case-control design this design allows us to estimate the absolute risk of the disease by knowing the crude disease rate from the cohort study. Some other study designs along this line are the case cohort design (Prentice, 1986), proband design (Gail, Pee, and Carroll, 2001) and case-only design (Armstrong, 2003). In a recent article, Cheng and Chen (2005) propose Bayesian analysis of case-control studies of genetic association, specifically for assessing gene-environment interaction via case-only design under independence of genetic and environmental factors. They use informative power prior (Ibrahim and Chen, 2000) which in their analysis is taken to be the retrospective likelihood based on historic data raised to a suitable power.

It is important to note that case-control designs are choice-based sampling designs in which the population is stratified on the values of the response variable itself. Among others, this was noticed in Scott and Wild (1986) who compared in such cases a maximum likelihood based approach with some other ad hoc methods of estimation. Breslow and Cain (1988) considered in such cases a two-phase sampling design. In later years, there is a long series of publications on inference for such designs, but any Bayesian approach to these problems is still lacking, and will be a worthwhile future undertaking.

12.3 Bayesian Choice of Links and Computation for Binary Response Data

Ming-Hui Chen, Sungduk Kim, Lynn Kuo, and Wangang Xie

The analysis of binary response data frequently arises in clinical investigations for chronic disease, including AIDS and cancer. For example, a binary response variable may be the presence or absence of cancer. In some other occasion, a binary response variable may be an indicator of the severeness of the disease such as whether the cancer has spread out around a certain organ. In this type of application, a binary response is often related to a patient's health conditions or certain other baseline characteristics such as age, weight, gender, and so on.

In medical and epidemiologic studies, one of the most commonly used models for binary response data is the logistic regression model. The logit link in a binary regression model gives a simple interpretation of the regression parameters, because

in this case, it allows for a simple representation of the odds ratio, which helps in the interpretation of the results. Other commonly used links include the probit and complementary log-log links. However, these popular links do not always provide the best fit for a given data set. The link could be misspecified, which can yield substantial bias in the mean response estimates (see Czado and Santner, 1992).

Chen, Dey, and Shao (1999) proposed to use the rate at which the expectation of a given binary response approaches 1 (or 0) to describe a link. Under this notion, a link is symmetric if the expectation of a given binary response approaches zero at the same rate as it approaches one. On the other hand, the link is skewed or asymmetric if the expectation of a given binary response approaches zero at a different rate than it approaches one. A skewed link can be further characterized as a positively skewed link or a negatively skewed link. A link is called *positively skewed* if the rate approaching to 1 is faster than the rate approaching to 0 and a link is called *negatively skewed* if the rate approaching to 0 is faster than the rate approaching to 1. One popular approach to guard against the misspecification of links is to embed the symmetric links, such as the logit or probit, into a wide parametric class of links. Many such parametric classes for binary response data have been proposed in the literature. Aranda-Ordaz (1981), Guerrero and Johnson (1982), Morgan (1983), and Whittmore (1983) proposed one-parameter families. Stukel (1988) extended these links by proposing a class of generalized logistic models. However, Chen, Dey, and Shao (1999) observed that in the presence of covariates, the Stukel's model yields improper posterior distributions for many types of noninformative improper priors, including the improper uniform prior for the regression coefficients. By using a latent variable approach of Albert and Chib (1993), Chen, Dey, and Shao (1999) proposed a class of skewed link models, where the underlying latent variable has a mixed effects model structure. Kim, Chen, and Dey (2008) proposed a rich class of symmetric generalized t-links. They showed that the symmetric generalized t-link is naturally resulted from the symmetric t-link with unknown degrees of freedom via reparameterization. Although the generalized t-link is essentially equivalent to the usual t-link in terms of model fitting, the symmetric generalized t-link model allows us to achieve faster convergence and better mixing of the Gibbs sampling algorithm, which is used to sample from the posterior distribution, than the t-link model with unknown degrees of freedom. In addition, Kim, Chen, and Dey (2008) developed a class of skewed generalized t-links. Most recently, Wang and Dey (2009) proposed the new generalized extreme value (GEV) link.

The choice of links and the selection of covariates are a crucial part of the analysis of binary response data. Chen, Dey, and Ibrahim (2004) proposed a Bayesian model assessment criterion, called weighted L measure, which is constructed from the posterior predictive distribution of the data, for binary regression, ordinal regression, multivariate correlated categorical data, and discrete choice data. Kim, Chen, and Dey (2008) used the Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002) to guide the choice of links. In this section, we consider a new strategy to carry out the model assessment. First, we randomly split the data into the testing cohort and validation cohort. Second, we use the DIC and the marginal likelihood to guide the choice of links and the selection of covariates in the testing data. Third,

we use the predictive mean squared error to validate the choice of links and variable selection in the validation data. Due to the complexity of the skewed generalized t-link models, the computation of marginal likelihoods is quite challenging. For this, we develop a version of the Stepping-Stone method proposed by Xie et al. (2009), which is particularly suitable for Bayesian binary regression models with latent variables.

The rest of this section is organized as follows. Section 12.3.1 presents binary regression models. In Section 12.3.2, we discuss the prior and posterior distributions and provide necessary formulas of DIC and marginal likelihoods. The detailed development of the Stepping-Stone method and the corresponding MCMC sampling algorithm are given in Section 12.3.3. Section 12.3.4 is a case study, while Section 12.3.5 concludes.

12.3.1 The Binary Regression Models

Consider a binary response y_i, $i = 1, 2, \ldots, n$. Let $\mathbf{x}_i = (x_{i0}, x_{i1}, \ldots, x_{ip})'$ be the corresponding $(p+1)$-dimensional vector of covariates, where $x_{i0} = 1$ corresponds to an intercept, let $\mathbf{y} = (y_1, y_2, \ldots, y_n)'$ be the vector of the observed binary responses, and let $\beta = (\beta_0, \beta_1, \ldots, \beta_p)'$ be a $(p+1)$-dimensional vector of regression coefficients. We assume that the y_i are independent. Then, the binary regression model for $[y_i|\mathbf{x}_i]$ assumes

$$f(y_i|\mathbf{x}_i, \beta) = \{F(\mathbf{x}_i'\beta)\}^{y_i}\{1 - F(\mathbf{x}_i'\beta)\}^{1-y_i},\ i = 1, 2, \ldots, n, \tag{12.3.1}$$

where $F(\cdot)$ denotes a cumulative distribution function (cdf) and F^{-1} is called the link function. The likelihood function of β is

$$L(\beta|X, \mathbf{y}) = \prod_{i=1}^{n}\{F(\mathbf{x}_i'\beta)\}^{y_i}\{1 - F(\mathbf{x}_i'\beta)\}^{1-y_i}, \tag{12.3.2}$$

where $X = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)'$ is the $n \times (p+1)$ design matrix. The logistic, probit, and complementary log-log (C log-log) regression models are special cases of (12.3.1) by taking $F(w) = \exp(w)/\{1 + \exp(w)\}$, $F(w) = \Phi(w)$, and $F(w) = 1 - \exp\{-\exp(w)\}$, respectively, where $\Phi(w)$ denotes the $N(0, 1)$ cdf evaluated at w.

Using the latent variable approach of Albert and Chib (1993), a binary regression model (12.3.1) can be described as follows. Let w_i be a latent variable such that

$$y_i = \begin{cases} 1 & \text{if } w_i > 0 \\ 0 & \text{if } w_i \le 0 \end{cases} \text{ and } w_i = \mathbf{x}_i'\beta + \varepsilon_i,\ \varepsilon_i \sim F, \tag{12.3.3}$$

where F is a cdf. When F is symmetric, i.e., $F(w) = 1 - F(-w)$, the binary regression model defined by (12.3.3) is equivalent to (12.3.1). Chen, Dey, and Shao (1999)

proposed a class of skewed links. Kim, Chen, and Dey (2008) extended the skewed link of Chen, Dey, and Shao (1999) to a class of flexible skewed generalized t-links. Let $f_{gt,\nu_1,\nu_2}(w)$ denote the probability density function (pdf) of the generalized t-distribution introduced by Arellano-Valle and Bolfarine (1995) is given by

$$f_{gt,\nu_1,\nu_2}(w) = \frac{1}{\sqrt{\pi}} \frac{\Gamma\left(\frac{\nu_1+1}{2}\right)}{\sqrt{\nu_2}\,\Gamma\left(\frac{\nu_1}{2}\right)} \times \frac{1}{\left(1+\frac{w^2}{\nu_2}\right)^{\frac{\nu_1+1}{2}}},$$

where ν_1 is a shape parameter (or degrees of freedom) and ν_2 is a scale parameter. The skewed generalized t-link model takes the form:

$$y_i = \begin{cases} 1 & \text{if } w_i > 0 \\ 0 & \text{if } w_i \le 0 \end{cases} \text{ and } w_i = \mathbf{x}_i'\beta + \delta\{z_i - E(z)\} + \varepsilon_i, \tag{12.3.4}$$

where $\delta > 0$, $\varepsilon_i \sim f_{gt,\nu_1,\nu_2=1}$, $z_i \sim G$, where G is the cdf of a skewed distribution, $E(z) = \int z dG(z)$, and z_i and ε_i are independent. We assume $0 \le \delta \le 1$ so that the resulting generalized t-link in (12.3.4) has heavier tails than the Cauchy-link. Notice that this constraint can be relaxed to $\delta > 0$. In (12.3.4), the parameter ν_1 purely controls the heaviness of the tails of the link and the parameter δ controls the scale of the link. The skewed generalized t-link model reduces to a symmetric generalized t-link model when $\delta = 0$ or G is a degenerate distribution at 0. To ensure model identifiability, we assume that G is known. The following G's are considered: (i) G is degenerated at 0, denoted by $\Delta_{\{0\}}$, i.e., $P(z_i = 0) = 1$; (ii) G is the standard exponential distribution $\mathscr{E}$ with pdf $g_{\mathscr{E}}(z_i) = \exp(-z_i)$ if $z_i > 0$ and 0 otherwise, which gives a positively skewed link; and (iii) G is the negative standard exponential distribution $\mathscr{NE}$ with pdf $g_{\mathscr{NE}}(z_i) = \exp(z_i)$ if $z_i < 0$ and 0 otherwise, which leads to a negatively skewed link. See Kim, Chen, and Dey (2008) for other attractive properties of the generalized t-link and other choices of G.

Let $F_{gt,\nu_1,\nu_2=1}(w) = \int_{-\infty}^{w} f_{gt,\nu_1,\nu_2=1}(u)du$, which is the cdf of the generalized t-distribution, and $g(z) = \frac{dG(z)}{dz}$. Then, the likelihood function of β, δ, and ν_1 is given by

$$\begin{aligned} L(\beta,\delta,\nu_1|X,\mathbf{y}) = \prod_{i=1}^{n} \Big\{ \int \Big[F_{gt,\nu_1,\nu_2=1}(\mathbf{x}_i'\beta + \delta\{z_i - E(z)\})\Big]^{y_i} \\ \times \Big[1 - F_{gt,\nu_1,\nu_2=1}(\mathbf{x}_i'\beta + \delta\{z_i - E(z)\})\Big]^{1-y_i} dG(z_i) \Big\}. \end{aligned} \tag{12.3.5}$$

12.3.2 Prior and Posterior Distributions

To carry out Bayesian inference, we need to specify a prior distribution for the model parameters. For the skewed generalized t-link model, we assume that β, δ

and ν_1 are independent *a priori*. Thus, the joint prior for (β, δ, ν_1) is of the form $\pi(\beta, \delta, \nu_1) = \pi(\beta)\pi(\delta)\pi(\nu_1)$. We further assume that $\beta \sim N_{p+1}(0, \tau_0 I_{p+1})$, $\pi(\delta)$ is proper, and

$$\pi(\nu_1) = \pi(\nu_1|\zeta_0, \gamma_0) = \frac{1}{1 - \Gamma(1|\zeta_0, \gamma_0)} \frac{\gamma_0^{\zeta_0}}{\Gamma(\zeta_0)} \nu_1^{\zeta_0 - 1} \exp(-\gamma_0 \nu_1), \quad \nu_1 > 1,$$

i.e., a Gamma(ζ_0, γ_0) distribution truncated at $\nu_1 > 1$, where ζ_0 and γ_0 are two pre-specified hyperparameters and

$$\Gamma(1|\zeta_0, \gamma_0) = \int_0^1 \frac{\gamma_0^{\zeta_0}}{\Gamma(\zeta_0)} \nu_1^{\zeta_0 - 1} \exp(-\gamma_0 \nu_1) d\nu_1.$$

For the logistic and C log-log regression models, we take $\beta \sim N_{p+1}(0, \tau_0 I_{p+1})$. In Section 12.3.4.2, we use $\tau_0 = 10$ for $\pi(\beta)$, $\zeta_0 = 1$ and $\gamma_0 = 2$ for $\pi(\nu_1)$, and $\pi(\delta) = 1$ for $0 < \delta < 1$, which corresponds to the uniform distribution $U(0, 1)$.

We consider the DIC and marginal likelihood criteria to compare the skewed generalized t-link model, the symmetric generalized t-link model, and logistic and C log-log regression models. We do not consider the probit model as it is a special case of the symmetric generalized t-link model with $\nu_1 \to \infty$. Let $D_o = (\mathbf{y}, X)$ denote the observed data. Under the logistic regression model or the C log-log regression model, the posterior distribution of β and the marginal likelihood are given by

$$\pi(\beta|D_o) \propto L(\beta|X, \mathbf{y})\pi(\beta) \tag{12.3.6}$$

and

$$m(D_o) = \int L(\beta|X, \mathbf{y})\pi(\beta) d\beta, \tag{12.3.7}$$

where $L(\beta|X, \mathbf{y})$ is defined by (12.3.2). Under the skewed generalized t-link model, the posterior distribution of (β, δ, ν_1) and the marginal likelihood are given by

$$\pi(\beta, \delta, \nu_1|D_o) \propto L(\beta, \delta, \nu_1|X, \mathbf{y})\pi(\beta)\pi(\delta)\pi(\nu_1) \tag{12.3.8}$$

and

$$m_{sgt}(D_o) = \int L(\beta, \delta, \nu_1|X, \mathbf{y})\pi(\beta)\pi(\delta)\pi(\nu_1) d\beta d\delta d\nu_1, \tag{12.3.9}$$

where $L(\beta, \delta, \nu_1|X, \mathbf{y})$ is defined by (12.3.5).

Let $\theta = \beta$ or $\theta = (\beta, \delta, \nu_1)$. Then, DIC is defined as

$$\text{DIC} = D(\overline{\theta}) + 2p_D,$$

where $D(\theta) = -2\log L(\theta|X, \mathbf{y})$, which is either $-2\log L(\beta|X, \mathbf{y})$ or $-2\log L(\beta, \delta, \nu_1|X, \mathbf{y})$, is a deviance function and $\bar{\theta} = E(\theta|D_o)$ is the posterior mean of θ with respect to the posterior distribution defined in either (12.3.6) or (12.3.8). In DIC, p_D is the effective number of model parameters, which is calculated as $p_D = \overline{D(\theta)} - D(\overline{\theta})$, where $\overline{D(\theta)} = E[D(\theta)|X, \mathbf{y}]$.

12.3.3 Computational Development

In this section, we discuss the Stepping-Stone methods and the MCMC sampling algorithm for computing marginal likelihoods.

12.3.3.1 Stepping-Stone Methods

From (12.3.2), (12.3.5), (12.3.7), and (12.3.9), it is easy to see that it does not appear possible to evaluate the marginal likelihoods $m(D_o)$ and $m_{sgt}(D_o)$ analytically. Thus, a Monte Carlo (MC) method is needed to compute these analytically intractable marginal likelihoods. Several Monte Carlo methods have been developed in the Bayesian computational literature, including, for example, Newton and Raftery (1994), Chib (1995), Meng and Wong (1996), Raftery (1996b), Chen and Shao (1997), and Chib and Jeliazkov (2001). However, these methods may not be directly applicable or efficient for computing $m_{sgt}(D_o)$. Thus, we develop an extension of the MC method independently proposed by Lartillot and Philippe (2006) and Friel and Pettitt (2008) via the power posterior.

We first discuss how to compute $m(D_o)$ given in (12.3.7). Define the kernel of the power posterior as follows

$$q_b(\beta) = L^b(\beta|X,\mathbf{y})\pi(\beta). \tag{12.3.10}$$

Let $m_b(D_o) = \int q_b(\beta)d\beta$. Then, the power posterior is given by

$$p_b(\beta|D_o) = \frac{q_b(\beta)}{m_b(D_o)}.$$

It is easy to show that $m_{b=1}(D_o) = m(D_o)$ and $m_{b=0}(D_o) = 1$. Observe that

$$\log m(D_o) = \log\left[\frac{m_{b=1}(D_o)}{m_{b=0}(D_o)}\right] = \int_0^1 E_b[U(\beta,b)]db, \tag{12.3.11}$$

where $U(\beta,b) = \frac{d}{db}\log q_b(\beta) = \log L(\beta|X,\mathbf{y})$ and the expectation is taken with respect to $p_b(\beta|D_o)$. Note that an identity similar to (12.3.11) is also established in Gelman and Meng (1998). Both Lartillot and Philippe (2006) and Friel and Pettitt (2008) advocated the trapezoidal approximation for computing $\log m(D_o)$ given in (12.3.11). Specifically, let $b_0 = 0 < b_1 < b_2 < \cdots < b_K = 1$. Then, the trapezoidal approximation to $\log m(D_o)$ is given by

$$\log m(D_o) \approx \sum_{k=1}^{K}(b_k - b_{k-1})\frac{\left[E_{b_k}\{U(\beta,b_k)\} + E_{b_{k-1}}\{U(\beta,b_{k-1})\}\right]}{2}.$$

Let $\{\beta_{kj},\ j = 1,2,\ldots,J\}$ denote a Markov chain Monte Carlo (MCMC) sample from $p_{b_k}(\beta|D_o)$ for $k = 0,1,2,\ldots,K$. Then, an Monte Carlo estimate of $\log m(D_o)$

is given by

$$\log \hat{m}(D_o) = \frac{1}{2}\sum_{k=1}^{K}(b_k - b_{k-1})\Big[\frac{1}{J}\sum_{j=1}^{J}\Big\{U(\beta_{kj}, b_k) + U(\beta_{k-1,j}, b_{k-1})\Big\}\Big]. \quad (12.3.12)$$

Note that under the logistic and C log-log regression models, it is easy to show that the power posterior $p_b(\beta|D_o)$ is log-concave. Thus, the MCMC samples from $p_{b_k}(\beta|D_o)$ can be easily obtained via the adaptive rejection algorithm of Gilks and Wild (1992).

In (12.3.12), Lartillot and Philippe (2006) chose $b_k - b_{k-1} = \frac{1}{K}$ while Friel and Pettitt (2008) set $\beta_k = a_k^4$, where the a_k's are 11 equally spaced points in the interval $[0,1]$. As discussed in Xie et al. (2009), $\log \hat{m}(D_o)$ is always biased regardless the MC sample size J. To obtain a more efficient MC estimate of $\log m(D_o)$, we consider the Stepping-Stone method (SSM) proposed by Xie et al. (2009). The SSM is based on the following identity:

$$m(D_o) = \frac{m_{b=1}(D_o)}{m_{b=0}(D_o)} = \prod_{k=1}^{K}\frac{m_{b_k}(D_o)}{m_{b_{k-1}}(D_o)}.$$

Let $r_k = \frac{m_{b_k}(D_o)}{m_{b_{k-1}}(D_o)}$. Using the identity given in Chen, Shao, and Ibrahim (2000, Chapter 5), we have

$$r_k = \int \frac{q_{b_k}(\beta)}{q_{b_{k-1}}(\beta)} p_{b_{k-1}}(\beta|D_o)d\beta.$$

Using the MCMC sample $\{\beta_{k-1,j},\ j = 1,2,\ldots,J\}$, an unbiased estimate of r_k is given by

$$\hat{r}_k = \frac{1}{J}\sum_{j=1}^{J}\frac{q_{b_k}(\beta_{k-1,j})}{q_{b_{k-1}}(\beta_{k-1,j})}. \quad (12.3.13)$$

for $k = 1,2,\ldots,K$. For the logistic, probit, and C log-log regression models, it can be shown that $p_{b_{k-1}}(\beta|D_o)$ has heavier tails than $p_{b_k}(\beta|D_o)$. Thus, as discussed in Chen, Shao, and Ibrahim (2000, Chapter 5), the use of the MCMC sample from a heavier tailed importance sampling density, namely, $p_{b_{k-1}}(\beta|D_o)$, leads to an efficient estimate $\hat{r}_k$ of r_k. Using (12.3.13), an MC estimate of $\log m(D_o)$ based on the SSM is thus given by

$$\log \hat{m}(D_o) = \sum_{k=1}^{K}\log \hat{r}_k = \sum_{k=1}^{K}\log\Big[\frac{1}{J}\sum_{j=1}^{J}\frac{q_{b_k}(\beta_{k-1,j})}{q_{b_{k-1}}(\beta_{k-1,j})}\Big]. \quad (12.3.14)$$

As shown in Xie et al. (2009), the estimate $\log \hat{m}(D_o)$ given by (12.3.14) based on the SSM is more efficient than the one given by (12.3.12) based on the trapezoidal approximation. Xie et al. (2009) proposed to choose b_k as the $(k/K)^{th}$ quantile of the beta distribution, $Beta(\alpha, 1)$, with $\alpha = 0.3$ to improve the efficiency of the SSM estimate $\log \hat{m}(D_o)$. Several other desirable properties of the SSM are also discussed in details in Xie et al. (2009).

Next, we discuss how to compute the marginal likelihood under the skewed generalized t-link model. Computing $m_{sgt}(D_o)$ in (12.3.9) is much more challenging than $m(D_o)$ in (12.3.7). We note that the power posterior similar to (12.3.10) does not work as it is difficult to sample (β, δ, ν_1) from the power posterior $p_b(\beta, \delta, \nu_1|D_o) \propto L^b(\beta, \delta, \nu_1|X, \mathbf{y})\pi(\beta)\pi(\delta)\pi(\nu_1)$ due to an unknown shape parameter ν_1, an unknown skewness parameter δ, and n analytically intractable integrals in (12.3.5). To this end, we will construct a novel power posterior via the introduction of several latent variables. Using a known fact that the generalized t-distribution can be represented as a gamma mixture of normal distributions, we introduce a mixing variable λ_i such that

$$\varepsilon_i \mid \lambda_i \sim N(0, 1/\lambda_i) \text{ and } \lambda_i \sim \text{Gamma}\,(\nu_1/2, 1/2) \text{ for } i = 1, 2, \ldots, n.$$

Let $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_n)'$, $\mathbf{w} = (w_1, w_2, \ldots, w_n)'$, and $\mathbf{z} = (z_1, z_2, \ldots, z_n)'$. Also let $\pi(\lambda_i|\nu_1) = \frac{\left(\frac{1}{2}\right)^{\nu_1/2}}{\Gamma\left(\frac{\nu_1}{2}\right)} \lambda_i^{\nu_1/2-1} \exp\left(-\frac{\lambda_i}{2}\right)$. Then, the augmented joint posterior distribution of $(\beta, \delta, \nu_1, \mathbf{w}, \lambda, \mathbf{z})$ based on the observed data D_o is given by

$$\begin{aligned}\pi(\beta, \delta, \nu_1, \mathbf{w}, \lambda, \mathbf{z}|D_o) \propto \prod_{i=1}^{n} \Big\{ & [1\{y_i = 0\}1\{w_i \le 0\} + 1\{y_i = 1\}1\{w_i > 0\}] \\ & \times \frac{1}{\sqrt{2\pi}} \lambda_i^{1/2} \exp\Big(-\frac{\lambda_i}{2}[w_i - \mathbf{x}_i'\beta - \delta\{z_i - E(z)\}]^2 \Big) \pi(\lambda_i|\nu_1) g(z_i) \Big\} \\ & \times \pi(\beta)\pi(\delta)\pi(\nu_1), \end{aligned} \tag{12.3.15}$$

where $g(z_i)$ is the density of z_i and the indicator function $1\{A\}$ is defined as $1\{A\} = 1$ if A is true and 0 otherwise. Now, we construct the kernel of an augmented power posterior as follows

$$\begin{aligned} q_b(\beta, \delta, \nu_1, \mathbf{w}, \lambda, \mathbf{z}) = \prod_{i=1}^{n} \Big(& [1\{y_i = 0\}1\{w_i \le 0\} + 1\{y_i = 1\}1\{w_i > 0\}] \\ & \times \frac{1}{\sqrt{2\pi}} \lambda_i^{1/2} \exp\Big(-\frac{b\lambda_i}{2}[w_i - \mathbf{x}_i'\beta - \delta\{z_i - E(z)\}]^2 \Big) \exp\Big\{ -\frac{(1-b)c_0\lambda_i}{2} w_i^2 \Big\} \\ & \times \pi(\lambda_i|\nu_1) g(z_i) \Big) \pi(\beta)\pi(\delta)\pi(\nu_1), \end{aligned} \tag{12.3.16}$$

where $0 \le b \le 1$ and $0 < c_0 < 1$ are constants. Let $m_{b,sgt}(D_o)$ denote the normalizing constant and the corresponding power posterior is thus given by

$$p_b(\beta, \delta, \nu_1, \mathbf{w}, \lambda, \mathbf{z}|D_o) = \frac{1}{m_{b,sgt}(D_o)} q_b(\beta, \delta, \nu_1, \mathbf{w}, \lambda, \mathbf{z}). \tag{12.3.17}$$

We are led to the following theorem.

Theorem 12.3. *The augmented power posterior given in (12.3.17) have the following properties:*

(i) $m_{b=1,sgt}(D_o) = m_{sgt}(D_o)$, *where* $m_{sgt}(D_o)$ *is given by (12.3.9);*

(ii) $m_{b=0,sgt}(D_o) = \left(\frac{1}{2\sqrt{c_0}}\right)^n$; *and*

(iii) *for* $0 \le b_{k-1} < b_k$ *and* $0 < c_0 < 1$, $p_{b_{k-1}}(\beta, \delta, \nu_1, \mathbf{w}, \lambda, \mathbf{z}|D_o)$ *has heavier tails than* $p_{b_k}(\beta, \delta, \nu_1, \mathbf{w}, \lambda, \mathbf{z}|D_o)$, *that is,*

$$\lim_{||\beta|| \text{ or } ||\mathbf{w}|| \text{ or } ||\lambda|| \text{ or } ||\mathbf{z}|| \to \infty} \frac{q_{b_k}(\beta, \delta, \nu_1, \mathbf{w}, \lambda, \mathbf{z})}{q_{b_{k-1}}(\beta, \delta, \nu_1, \mathbf{w}, \lambda, \mathbf{z})} = 0,$$

where $||\cdot||$ *denotes the norm operator of a vector, for example,* $||\beta|| = \sqrt{\beta'\beta}$.

The proof of Theorem 12.3 follows from straightforward algebra and, thus, the detail is omitted for brevity. The results given in Theorem 12.3 shed light on the computation of the marginal likelihood $m_{sgt}(D_o)$ via the SSM for the skewed generalized t-link model. Furthermore, the construction of the augmented power posterior leads to a convenient implementation of MCMC sampling from $p_b(\beta, \delta, \nu_1, \mathbf{w}, \lambda, \mathbf{z}|D_o)$. The detailed description of the MCMC sampling algorithm is discussed in the next subsection. Let $\{(\beta_{kj}, \delta_{kj}, \nu_{1,kj}, \mathbf{w}_{kj}, \lambda_{kj}, \mathbf{z}_{kj}), j = 1, 2, \ldots, J\}$ denote an MCMC sample from $p_{b_k}(\beta, \delta, \nu_1, \mathbf{w}, \lambda, \mathbf{z}|D_o)$ for $k = 0, 1, \ldots, K-1$. Using the SSM, an MC estimate of $\log m_{sgt}(D_o)$ is given by

$$\log \hat{m}_{sgt}(D_o) = -n\log(2\sqrt{c_0}) + \sum_{k=1}^{K} \log\Big[\frac{1}{J}\sum_{j=1}^{J}\prod_{i=1}^{n}\exp\Big\{-\frac{(b_k - b_{k-1})\lambda_{i,k-1,j}}{2} \times \Big([w_{i,k-1,j} - \mathbf{x}_i'\beta_{k-1,j} - \delta_{k-1,j}\{z_{i,k-1,j} - E(z)\}]^2 - c_0 w_{i,k-1,j}^2\Big)\Big\}\Big].$$

In Section 12.3.4.2, we use $J = 50$ and $c_0 = 0.1$.

12.3.3.2 Sampling from the Power Posterior

To sample from the power posterior $p_b(\beta, \delta, \nu_1, \mathbf{w}, \lambda, \mathbf{z}|D_o)$ given in (12.3.17), we require sampling from the conditional distributions: (i) $[w_i|\beta, \delta, \nu_1, \lambda_i, z_i, D_o]$; (ii) $[z_i|\beta, \delta, \nu_1, w_i, \lambda_i, D_o]$; (iii) $[\beta, \delta|\nu_1, \mathbf{w}, \lambda, \mathbf{z}, D_o]$; and (iv) $[\nu_1, \lambda|\beta, \delta, \mathbf{w}, \mathbf{z}, D_o]$. We briefly discuss how to sample from each of the above conditional posterior distributions. We discuss the case $G = \mathscr{E}$ only, as the MCMC sampling algorithms for other choices of G are similar. For (i),

$$w_i \mid \beta, \delta, \nu_1, \lambda_i, z_i, D_o \sim N\left(\frac{b[\mathbf{x}_i'\beta + \delta\{z_i - E(z)\}]}{b + c_0(1-b)}, \frac{1}{\lambda_i\{b + c_0(1-b)\}}\right) \times [1\{y_i = 0\}1\{w_i \le 0\} + 1\{y_i = 1\}1\{w_i > 0\}]$$

for $i = 1, 2, \ldots, n$ and for (ii)

$$z_i \mid \beta, \delta, \nu_1, w_i, \lambda_i, D_o \sim N\left(\frac{b\lambda_i\{w_i - \mathbf{x}_i'\beta + \delta E(z)\}\delta - 1}{b\lambda_i\delta^2}, \frac{1}{b\lambda_i\delta^2}\right) 1\{z_i > 0\}$$

for $i = 1,2,\ldots,n$. Since $[w_i|\beta,\delta,\nu_1,\lambda_i,z_i,D_o]$ and $[z_i|\beta,\delta,\nu_1,w_i,\lambda_i,D_o]$ are two truncated normals, sampling w_i and z_i is straightforward. For (iii), we apply the collapsed Gibbs technique of Liu (1994) via

$$[\beta,\delta|\nu_1,\mathbf{w},\lambda,\mathbf{z},D_o] = [\beta|\delta,\nu_1,\mathbf{w},\lambda,\mathbf{z},D_o][\delta|\nu_1,\mathbf{w},\lambda,\mathbf{z},D_o]. \tag{12.3.18}$$

That is, we sample δ after collapsing out β. Given δ, ν_1, $\mathbf{w}$, λ, $\mathbf{z}$, and D_o, we observe that

$$\beta \mid \nu_1,\mathbf{w},\lambda,\delta,\mathbf{z},D_o \sim N_{p+1}\left(\Sigma_\beta^{-1}B,\Sigma_\beta^{-1}\right),$$

where $\Sigma_\beta = b\sum_{i=1}^n \lambda_i\mathbf{x}_i\mathbf{x}_i' + (1/\tau_0)I_{p+1}$ and $B = b\sum_{i=1}^n \lambda_i[w_i - \delta\{z_i - E(z)\}]\mathbf{x}_i$. Thus, sampling β is straightforward. In (12.3.18), given ν_1, $\mathbf{w}$, λ, $\mathbf{z}$, and D_o, we have

$$\delta \mid \nu_1,\mathbf{w},\lambda,\mathbf{z},D_o \sim N\left(C^{-1}D, C^{-1}/b\right)1\{0<\delta<1\},$$

where

$$C = \sum_{i=1}^n \lambda_i\{z_i - E(z)\}^2 - b\Big[\sum_{i=1}^n \lambda_i\{z_i - E(z)\}\mathbf{x}_i\Big]'\Big\{b\sum_{i=1}^n \lambda_i\mathbf{x}_i\mathbf{x}_i' + (1/\tau_0)I_{p+1}\Big\}^{-1} \times \Big[\sum_{i=1}^n \lambda_i\{z_i - E(z)\}\mathbf{x}_i\Big]$$

and

$$D = \sum_{i=1}^n \lambda_i w_i\{z_i - E(z)\} - b\Big[\sum_{i=1}^n \lambda_i\{z_i - E(z)\}\mathbf{x}_i\Big]'\Big\{b\sum_{i=1}^n \lambda_i\mathbf{x}_i\mathbf{x}_i' + (1/\tau_0)I_{p+1}\Big\}^{-1} \times \Big(\sum_{i=1}^n \lambda_i w_i\mathbf{x}_i\Big).$$

Since $[\delta \mid \nu_1,\mathbf{w},\lambda,\mathbf{z},D_o]$ is truncated normal distribution, sampling δ is straightforward. For (iv), we again apply the collapsed Gibbs technique of Liu (1994) via

$$[\nu_1,\lambda|\beta,\delta,\mathbf{w},\mathbf{z},D_o] = [\lambda|\nu_1,\beta,\delta,\mathbf{w},\mathbf{z},D_o][\nu_1|\beta,\delta,\mathbf{w},\mathbf{z},D_o]. \tag{12.3.19}$$

Given ν_1, β, δ, w_i, z_i, and D_o, the λ_i are conditionally independent and

$$\lambda_i \mid \nu_1,\beta,\delta,w_i,z_i,D_o \sim \text{Gamma}\Big(\frac{\nu_1+1}{2}, \frac{1+(1-b)c_0w_i^2 + b[w_i - \mathbf{x}_i'\beta - \delta\{z_i - E(z)\}]^2}{2}\Big)$$

for $i = 1,2,\ldots,n$. In (12.3.19), the conditional posterior density for $[\nu_1|\beta,\delta,\mathbf{w},\mathbf{z},D_o]$ has the form

$$\begin{aligned}
&\pi(\nu_1|\beta,\delta,\mathbf{w},\mathbf{z},D_o)\\
&\propto \left\{\frac{\Gamma\left(\frac{\nu_1+1}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)}\right\}^n \prod_{i=1}^n \left(1+(1-b)c_0 w_i^2 + b\left[w_i - \mathbf{x}_i'\beta - \delta\{z_i - E(z)\}\right]^2\right)^{-\nu_1/2}\\
&\quad\times\ \nu_1^{\zeta_0-1}\ \exp(-\gamma_0\nu_1)1(\nu_1>1). \qquad (12.3.20)
\end{aligned}$$

Therefore, we sample λ_i from a gamma distribution and use the Metropolis-Hastings algorithm (Hastings, 1970) to sample ν_1 from (12.3.20).

Finally, we note that when $b=0$, the power posterior reduces to

$$\begin{aligned}
\pi_{b=0}(\beta,\delta,\nu_1,\mathbf{w},\lambda,\mathbf{z}|D_o) = \prod_{i=1}^n \Big\{ &[1\{y_i=0\}1\{w_i\le 0\} + 1\{y_i=1\}1\{w_i>0\}]\\
&\times \frac{1}{\sqrt{2\pi}}\lambda_i^{1/2}\exp\left(-\frac{c_0\lambda_i}{2}w_i^2\right)\pi(\lambda_i|\nu_1)g(z_i)\Big\}\pi(\beta)\pi(\delta)\pi(\nu_1).
\end{aligned}$$

In this special case, we do not need MCMC sampling. Specifically, we let $w_i^* = \sqrt{\lambda_i}w_i$ for $i=1,2,\ldots,n$ and $\mathbf{w}^* = (w_1^*, w_2^*, \ldots, w_n^*)'$. Then, the power posterior after the above transformation becomes

$$\begin{aligned}
\pi_{b=0}(\beta,\delta,\nu_1,\mathbf{w}^*,\lambda,\mathbf{z}|D_o) = \prod_{i=1}^n \Big\{ &[1\{y_i=0\}1\{w_i^*\le 0\} + 1\{y_i=1\}1\{w_i^*>0\}]\\
&\times \frac{1}{\sqrt{2\pi}}\exp\left[-\frac{c_0}{2}(w_i^*)^2\right]\pi(\lambda_i|\nu_1)g(z_i)\Big\}\pi(\beta)\pi(\delta)\pi(\nu_1).
\end{aligned}$$

Thus, we can directly generate a random sample from $\pi_{b=0}(\beta,\delta,\nu_1,\mathbf{w}^*,\lambda,\mathbf{z}|D_o)$.

12.3.4 A Case Study

12.3.4.1 The Data

We consider retrospective data from a cohort study of men treated with radical prostatectomy ($n=1273$). The data are a subset of the data published in D'Amico et al. (2002). The primary endpoint D'Amico et al. (2002) considered was the prostate specific antigen (PSA) recurrence time. In this dicussion we focus on another aspect of the data. We use a binary response y defined as an indicator for the event that the tumor has penetrated the prostate wall (CAP). We consider four prognostic factors as covariates, PSA, PPB, GLEAS and T2. Here, PPB is the percent positive prostate biopsies, GLEAS is the biospy Gleason score, and T2 is an indicator for the tumor being classified as T2 tumor by the 1992 AJCC (American Joint Commission on Cancer) classification. We consider two alternative codings for PSA, either on the original scale (PSA), or as natural logarithm (LOGPSA). For these covariates, the means and standard deviations were 2.114 and 0.603 for LOGPSA, 0.401 and 0.247 for PPB, and 5.997 and 1.123 for GLEAS. There were 492 patients who had clini-

cal T2 category disease. In addition, 328 patients had the disease extending into or having penetrated through the prostate capsule.

About 70% of the patients ($n = 866$) were randomly chosen as a testing cohort, the remaining 30% of patients ($n = 371$) formed a validation cohort. For the testing and validation cohorts we find 26.44% and 26.68%, respectively, with CAP$= 1$. We fit the two covariates models (PSA, PPB, GS, T2) and (LogPSA, PPB, GS, T2) with logistic and C log-log links to the testing data. The AIC values are 846.1 and 853.0 for the logit and C log-log links using covariates (PSA, PPB, GS, T2) and 837.5 and 841.3 for the logit and C log-log links using covariates (LogPSA, PPB, GS, T2), respectively. According to the AIC criterion, the logit link models outperform than the C log-log link models. Also, the covariates (LogPSA, PPB, GS, T2) fits the data better than the covariates (PSA, PPB, GS, T2). Does the symmetric logit link indeed fit the data better skewed links? Due to the nature of prostate cancer data, it is expected that a skewed link may be more desirable. The question is: Why is the fit under the asymmetric C log-log link model much worse than the symmetric logit model? We examine this issue carefully under Bayesian model comparison criteria in the subsequent subsections.

12.3.4.2 Model Fitting

For the testing data, we fit the logit and C log-log link models as well as the symmetric and skewed generalized t-link models under each of the covariates models (PSA, PPB, GS, T2) and (LogPSA, PPB, GS, T2). We computed DICs and the marginal likelihoods for all models. The results are given in Table 12.9. It shows that (i) the generalized t-links fit the data much better than the logit and C log-log links under both covariates models; (ii) the C log-log link fits the data worst; and (iii) the negatively skewed generalized t-link with $G = \mathscr{N}\mathscr{E}$ fits the data best. When $G = \Delta_{\{0\}}$, which corresponds to a symmetric link, the symmetric generalized t-link model outperforms the symmetric logit model. We also observe that both the marginal likelihood and DIC criteria consistently indicate the same best model. We note that the conclusions we obtained from the testing data is consistent with those obtained by Kim, Chen, and Dey (2008) using the whole data. We observe that C log-log link model fits the data poorly. As pointed out by Kim, Chen, and Dey (2008), the primary reasons for this are that (i) the C log-log link is positively skewed and (ii) the tails of the link are too light. We also note that under the logit and C log-log links, the DIC values are almost identical to the corresponding AIC values.

For all DIC's and marginal likelihoods shown in Table 12.9, we used the moderate priors for β and γ_0 specified in Section 12.3.2. Since the dimension of β does not change across all models considered in Table 12.9, we expect that the marginal likelihoods will not be too sensitive to the choice of τ_0 for $\pi(\beta)$. However, the hyperparameters (ζ_0, γ_0) for $\pi(\nu_1)$ are associated only with those generalized t-link models. Therefore, we conducted a sensitivity analysis on the specification of priors for both ν_1 and β. Instead of $\tau_0 = 10$, $\zeta_0 = 1$ and $\gamma_0 = 2$, we considered $\tau_0 = 20$, $\zeta_0 = 1$ and $\gamma_0 = 1$. Under this specification of hyperparameters, the log marginal

likelihoods are -437.6, -443.0, -436.6, -435.3, and -436.7 for the logit, C log-log, and generalized t-links with $G = \Delta_{\{0\}}$, $\mathscr{NE}$, and $\mathscr{E}$, respectively, under the covariates model (PSA, PPB, GS, T2); and -433.4, -436.9, -433.0, -432.5, and -432.9, for these five links, respectively, under the covariates model (LogPSA, PPB, GS, T2). Although the values of log marginal likelihoods are different than those in Table 12.9, the negatively skewed generalized t-link with $G = \mathscr{NE}$ remains to be the top model based on the marginal likelihoods criterion. We also considered other vague priors and the top models still remain the same. Finally, we mention that in all marginal likelihood computations, we used MCMC sample size of 50,000 and the Monte Carlo errors were less than 0.05.

TABLE 12.9. DIC values and log marginal likelihoods $\log m(D_0)$ of two covariates models under symmetric and skewed links for the testing data.

Link		Covariates Model			
		(PSA, PPB, GS, T2)		(LogPSA, PPB, GS, T2)	
F	G	DIC	$\log m(D_0)$	DIC	$\log m(D_0)$
Logit	—	846.1	-436.0	837.5	-431.8
C log-log	—	853.0	-441.3	841.4	-435.3
Generalized t	$\Delta_{\{0\}}$	843.7	-434.4	836.1	-431.1
(ν_1 random)	$\mathscr{NE}$	842.3	-433.5	835.4	-430.4
	$\mathscr{E}$	844.6	-434.3	836.8	-430.9

12.3.4.3 Predictive Validation

Let $D_{new} = (\mathbf{y}_{new}, X_{new})$ denote the validation data, where $\mathbf{y}_{new} = (y_{new,1}, y_{new,2}, \ldots, y_{new,n_{new}})'$, $X_{new} = (\mathbf{x}_{new,1}, \mathbf{x}_{new,2}, \ldots, \mathbf{x}_{new,n_{new}})'$, and $\mathbf{x}_{new,i} = (1, x_{new,i1}, \ldots, x_{new,ip})'$ is the $(p+1)$-dimensional vector of covariates for $i = 1, 2, \ldots, n_{new}$. We consider the predictive mean square error (PMSE). Let $z_{new,i}$ denote the future observation for the i^{th} subject. Then, the PMSE is defined as follows:

$$\text{PMSE} = \sum_{i=1}^{n_{new}} E[(z_{new,i} - y_{new,i})^2 | D_o]. \tag{12.3.21}$$

Since $E(z_{new,i}^2 | D_o) = E(z_{new,i} | D_o)$, (12.3.21) reduces to

$$\text{PMSE} = \sum_{i=1}^{n_{new}} \{y_{new,i} - E(z_{new,i} | D_0)\}^2 + \sum_{i=1}^{n_{new}} E(z_{new,i} | D_o)\{1 - E(z_{new,i} | D_o)\}.$$

For the validation data, we computed the PMSEs for the logit, C log-log, generalized t-links under two covariates models (PSA, PPB, GS, T2) and (LogPSA, PPB, GS, T2). The results are shown in Table 12.10. According to the PMSE criterion, the best model is the generalized t link with $G = \mathscr{NE}$ under the covariates model

(LogPSA, PPB, GS, T2). However, the difference in PMSEs between the skewed and symmetric generalized t links is smaller under the covariates model (LogPSA, PPB, GS, T2) than under the covariates model (PSA, PPB, GS, T2). The results based on PMSEs are very consistent with those from DICs or the marginal likelihoods, which validates the link and the covariates model selected by the Bayesian criteria based on the testing data.

TABLE 12.10. PMSEs of two covariates models under symmetric and skewed links for the validation data.

Link		Covariates Model (PSA, PPB, GS, T2)	Covariates Model (LogPSA, PPB, GS, T2)
F	G	PMSE	PMSE
Logit	—	115.6	114.8
C log-log	—	118.6	116.5
Generalized t	$\Delta_{\{0\}}$	114.8	114.3
(ν_1 random)	$\mathscr{NE}$	114.4	114.1
	$\mathscr{E}$	114.9	114.5

12.3.4.4 Posterior Estimates

Table 12.11 shows the posterior means, the posterior standard deviations and the 95% highest posterior density (HPD) intervals of the parameters for covariate models (PSA, PPB, GS, T2) and (LogPSA, PPB, GS, T2) under the symmetric generalized t-link and the skewed generalized t-link with $G = \mathscr{NE}$. Except for the intercept, the posterior estimates of all regression coefficients are positive, which implies that the probability of CAP=1 is an increasing function of PSA, PPB, GS, and T2. Also, all four covariates are highly significant under both links as all 95% HPD interval estimates do not contain 0. In addition, we notice that the posterior means of ν_1 are approximately 1.4 for all four models considered in Table 12.11. This result implies that the links with moderate heavy tails fit the data better, which further explains why the C log-log link fits the data poorly.

12.3.5 Discussion

In this section, we have demonstrated that the choice of the link-function is important in fitting binary response data and in particular, the direction of a skewed link plays an important role. Sections 12.3.4.2 and 12.3.4.3 empirically show that it is also important that the choice of links should be done in conjunction with the variable slection.

TABLE 12.11. Posterior estimates under symmetric and skewed generalized t-link models.

Covariates Model	Link	Variable	Posterior Mean	Posterior SD	95% HPD Interval
(PSA, PPB, GS, T2)	$G = \Delta_{\{0\}}$	Intercept	-0.997	0.194	(-1.366, -0.650)
		PSA	0.473	0.160	(0.189, 0.795)
		PPB	0.550	0.127	(0.328, 0.812)
		GS	0.301	0.109	(0.103, 0.521)
		T2	0.233	0.092	(0.066, 0.422)
		ν_1	1.352	0.244	(1.001, 1.772)
	$G = \mathcal{NE}$	Intercept	-1.136	0.209	(-1.548, -0.755)
		PSA	0.549	0.179	(0.230, 0.911)
		PPB	0.598	0.137	(0.341, 0.859)
		GS	0.320	0.114	(0.117, 0.560)
		T2	0.252	0.096	(0.073, 0.446)
		δ	0.669	0.247	(0.186, 0.999)
		ν_1	1.370	0.236	(1.001, 1.777)
(LogPSA, PPB, GS, T2)	$G = \Delta_{\{0\}}$	Intercept	-1.037	0.204	(-1.434, -0.674)
		LOGPSA	0.439	0.124	(0.214, 0.685)
		PPB	0.528	0.125	(0.297, 0.772)
		GS	0.310	0.110	(0.111, 0.535)
		T2	0.233	0.089	(0.065, 0.413)
		ν_1	1.375	0.248	(1.001, 1.794)
	$G = \mathcal{NE}$	Intercept	-1.190	0.227	(-1.643, -0.788)
		LOGPSA	0.500	0.137	(0.255, 0.777)
		PPB	0.584	0.138	(0.328, 0.853)
		GS	0.341	0.118	(0.122, 0.575)
		T2	0.265	0.099	(0.083, 0.467)
		δ	0.623	0.258	(0.141, 0.999)
		ν_1	1.362	0.229	(1.002, 1.760)

In Section 12.3.3, for the logistic probit, and complementary log-log regression models, we develop the SSM method based on the power posterior given in (12.3.10). The SSM can be extended by using the power posterior given by

$$q_b(\beta) = \{L(\beta|X,\mathbf{y})\pi(\beta)\}^b\{\pi^*(\beta)\}^{1-b}, \tag{12.3.22}$$

where $\pi^*(\beta)$ is a "working prior." One key assumption for this working prior is that it is known in closed form including the normalizing constant. Recently, Lefebvrea, Steele, and Vandal (2010) derived the expression of the J-divergence between the working prior and the posterior distribution and showed that the J-divergence is helpful for choosing the working prior that minimizes the error of the MC estimator of the marginal likelihood. According to Lefebvrea, Steele, and Vandal (2010), a simple way to choose a working prior is to use the large sample normal approximation to the posterior. Assume that a normal prior $N(\beta_0, \Sigma_0)$ is specified for $\pi(\beta)$. Then, using (12.3.2) and following Chen (1985), the joint posterior distribution (12.3.6) can be approximated by

$$\beta \sim N_{p+1}\left(\left[\widehat{\Sigma}^{-1}+\Sigma_0^{-1}\right]^{-1}(\widehat{\Sigma}^{-1}\hat{\beta}+\Sigma_0^{-1}\beta_0), \left[\widehat{\Sigma}^{-1}+\Sigma_0^{-1}\right]^{-1}\right), \qquad (12.3.23)$$

where $\hat{\beta}$ is the maximum likelihood estimate of β and

$$\widehat{\Sigma} = \left\{-\frac{\partial^2 \log L(\beta|X,\mathbf{y})}{\partial\beta\partial\beta'}\Big|_{\beta=\hat{\beta}}\right\}^{-1}.$$

Since a "good" working prior smooths the gap between $\pi^*(\beta)$ and the posterior $\pi(\beta|D_o)$, the use of (12.3.23) as the working prior can ease the computational burden (i.e., a much smaller J needed) and reduce the MC error of the SSM estimate of the marginal likelihood. In the same spirit, for the skewed generalized t-link model, an improved version of the working prior used in (12.3.16) can be developed. The SSM is quite general. This MC method can be used for computing marginal likelihoods not only for the binary regression models but also for many other Bayesian models such as Bayesian phylogenetic models discussed in Xie et al. (2009).

Acknowledgments: The authors wish to thank Dr. Anthony V. D'Amico of Brigham and Women's Hospital and Dana Farber Cancer Institute for providing the prostate cancer data. Dr. Chen's research was partially supported by NIH grants #GM 70335 and #CA 74015. Dr. Kim's research was supported by the Intramural Research Program of National Institutes of Health, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development. Dr. Kuo's research was partially supported by Connecticut Stem Cell Initiatives #06SCC04.

Chapter 13
Bayesian Geophysical, Spatial and Temporal Statistics

Spatio-temporal models give rise to many challenging research frontiers in Bayesian analysis. One simple reason is that the spatial and/or time series nature of the data implies complicated dependence structures that require modeling and lead to often challenging inference problems. The power of the Bayesian approach comes to bear especially when inference is desired on aspects of the model that are removed from the data by various levels in the hierarchical model. In this chapter we discuss two examples of such problems and also review the use of non-informative priors in spatial models.

13.1 Modeling Spatial Gradients on Response Surfaces

Sudipto Banerjee and Alan E. Gelfand

Spatial data are widely modelled using spatial processes that assume, for a study region D, a collection of random variables $\{W(\mathbf{s}) : \mathbf{s} \in D\}$, where $\mathbf{s}$ indexes the points in D. This set is viewed as a randomly realized surface over D which, in practice, is only observed at a finite set of locations in $\mathscr{S} = \{\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n\}$. For point referenced spatial data that are assumed (perhaps after suitable transformation) to be normally distributed, we employ a Gaussian spatial process to specify the joint distribution for an arbitrary number of and arbitrary choice of locations in D. This leads to spatial regression models of the form

$$Y(\mathbf{s}) = \mathbf{x}(\mathbf{s})'\tilde{\beta}(\mathbf{s}) + \varepsilon(\mathbf{s}), \tag{13.1.1}$$

where $\varepsilon(\mathbf{s}) \overset{iid}{\sim} N(0, \tau^2)$, $\mathbf{x}(\mathbf{s})$ is a $p \times 1$ vector and includes an intercept, and $\tilde{\beta}(\mathbf{s})$ follows a multivariate Gaussian spatial process model. The model in (13.1.1) is referred to as a spatially varying coefficient model (Gelfand et al., 2004). This locally linear form is very flexible and offers a likelihood-based alternative to geographi-

M.-H. Chen et al. (eds.), *Frontiers of Statistical Decision Making and Bayesian Analysis*, DOI 10.1007/978-1-4419-6944-6_13,

cally weighted regression (GWR, see, e.g., Fotheringham, Brunsdon, and Charlton, 2006). For each explanatory variable, it allows the possibility of a coefficient surface over the study region rather than restriction to a constant coefficient. Instead of specifying these surfaces parametrically (as in, e.g., Luo and Wahba, 2007) we model them as realizations from dependent spatial processes.

Writing $\tilde{\beta}(\mathbf{s}) = \beta + \mathbf{w}(\mathbf{s})$, we can rewrite the model in (13.1.1) as

$$Y(\mathbf{s}) = \mathbf{x}(\mathbf{s})'\beta + \mathbf{x}(\mathbf{s})'\mathbf{w}(\mathbf{s}) + \varepsilon(\mathbf{s}), \tag{13.1.2}$$

where β is interpreted as a global (spatially static) coefficient vector and $\mathbf{w}(\mathbf{s})$ is a mean $\mathbf{0}$ multivariate Gaussian process providing local variation around β. In particular, the spatially varying intercept in $\mathbf{w}(\mathbf{s})$ acts as a familiar spatial random effect which, along with $\varepsilon(\mathbf{s})$ would yield a customary residual at location $\mathbf{s}$. Gelfand et al. (2003) note that $\mathbf{x}(\mathbf{s})'\mathbf{w}(\mathbf{s}) + \varepsilon(\mathbf{s})$ in (13.1.2) could be viewed as a residual with the first piece interpreted as a spatial component and the second as a pure error component.

Statistical inference comprises estimating the response surface $\mathbf{x}(\mathbf{s})'\tilde{\beta}(\mathbf{s})$ or the components of $\mathbf{w}(\mathbf{s})$. Once such an interpolated surface has been obtained, investigation of rapid change on the surface may be of interest. Estimating mean surface gradients would be relevant in various applications including estimation of weather surfaces such as for temperature or precipitation, pollution surfaces such as for ozone or particulate matter, risk surfaces reflecting risk for a particular adverse health outcome. The mean surface provides global or first order description of the process guiding the data while gradient analysis provides local or second order analysis to enhance our understanding of the nature of the mean surface. Gradient analysis is post hoc; if a model choice criterion (e.g., Gelfand and Ghosh, 1998; Spiegelhalter et al., 2002) selects a model with mean surface as in (13.1.2), then gradients will be studied for this model.

In practice, spatial data analysis proceeds from response surface displays obtained from surface interpolators yielding contour and image plots. Surface representations and contouring methods range from tensor-product interpolators for gridded data to more elaborate adaptive control-lattice or tessellation based interpolators for scattered data. Mitas and Mitasova (1999) provide a review of several such methods available in GIS software (e.g., GRASS: `http://grass.itc.it/`). These methods are often fast and simple to implement and can reveal topographic features but are descriptive, lacking formal inference, not accounting for association and uncertainty in the data. For us, they play a complementary role, creating plots from the raw data in the pre-modeling stage and providing visual displays of estimated response or residual surfaces in the post-modeling stage. Such displays can certainly assist in locating points, or even zones, with high gradients.

This section is largely based upon the theoretical developments in Banerjee, Gelfand, and Sirmans (2003) and uses an application that has been comprehensively treated by Majumdar et al. (2006). We also look at a nonparametric extension of this work presented in Guindani and Gelfand (2006). We review a fully inferential framework to examine spatial gradients on the response surface. Since none

of the components of $\tilde{\beta}(\mathbf{s})$, hence of $\mathbf{x}(\mathbf{s})'\tilde{\beta}(\mathbf{s})$, are ever observed, inference about associated gradients is, arguably, most easily implemented within a Bayesian framework. Thus, entire posterior distributions can be proffered for the magnitude of any particular directional gradient. Furthermore, since, at a given location, for a given surface, we can identify the direction of maximal gradient and the magnitude of the gradient in that direction, posteriors for these unknowns can be obtained as well.

In Section 13.1.1 we offer a review of the theory behind directional derivative processes, while Section 13.1.3 discusses the modeling and inference in a Bayesian setting. Section 13.1.2 discusses different versions of mean square gradients with Section 13.1.4 looking at the nonparamteric version. Then, Section 13.1.5 presents an illustration and Section 13.1.6 concludes the manuscript with some discussion.

13.1.1 Directional Derivative Processes

Derivatives (more generally, linear functionals) of random fields have been discussed in Adler (1981), Mardia et al. (1996), and Banerjee, Gelfand, and Sirmans (2003). Let $W(\mathbf{s})$ be a real-valued stationary spatial process with covariance function $\text{Cov}\{W(\mathbf{s}_1), W(\mathbf{s}_2)\} = K(\mathbf{s}_1 - \mathbf{s}_2)$, where K is a positive definite function on $\Re^d$. Stationarity is not strictly required, but simplifies forms for the induced covariance function. The process $\left\{W(\mathbf{s}) : \mathbf{s} \in \Re^d\right\}$ is L_2 (or mean square) continuous at $\mathbf{s}_0$ if $\lim_{\mathbf{s}\to\mathbf{s}_0} E\left(|W(\mathbf{s}) - W(\mathbf{s}_0)|\right)^2 = 0$. Under stationarity, we have $E\left(|W(\mathbf{s}) - W(\mathbf{s}_0)|\right)^2 = 2(K(\mathbf{0}) - K(\mathbf{s} - \mathbf{s}_0))$, hence the process $W(\mathbf{s})$ is mean-square continuous at all sites $\mathbf{s}$ if K is continuous at $\mathbf{0}$.

The notion of a mean square differentiable process can be formalized using the analogous definition of total differentiability of a function in $\Re^d$ in a non-stochastic setting. To be precise, we say that $W(\mathbf{s})$ is mean square differentiable at $\mathbf{s}_0$ if it admits a first order linear expansion for any scalar h and any unit vector (direction) $\mathbf{u} \in \Re^d$,

$$W(\mathbf{s}_0 + h\mathbf{u}) = W(\mathbf{s}_0) + h\langle \nabla W(\mathbf{s}_0), \mathbf{u}\rangle + o(h) \tag{13.1.3}$$

in the L_2 sense as $h \to 0$, where $\nabla W(\mathbf{s}_0)$ is a $d \times 1$ vector called the gradient vector and $\langle \cdot, \cdot \rangle$ is the usual Euclidean inner-product on $\Re^d$. We write $\nabla' W(\mathbf{s})$ to denote the gradient vector as a $1 \times d$ row-vector.

Note that, unlike the non-stochastic setting, $W(\mathbf{s}_0)$ is a random realization at $\mathbf{s}_0$. Also, $\nabla W(\mathbf{s}_0)$ is not a function but a random d-dimensional vector. That is, for any unit vector $\mathbf{u}$, (13.1.3) is interpreted as

$$\lim_{h\to 0} E\left(\frac{W(\mathbf{s}_0 + h\mathbf{u}) - W(\mathbf{s}_0)}{h} - \langle \nabla W(\mathbf{s}_0), \mathbf{u}\rangle\right)^2 = 0. \tag{13.1.3$'$}$$

The linearity in (13.1.3) ensures that mean square differentiable processes are mean square continuous. A counter example when this condition does not hold is given in Banerjee and Gelfand (2003).

Spatial gradients can be developed from *finite difference processes*. For any *parent process* $W(\mathbf{s})$ and given direction $\mathbf{u}$ and any scale h we have

$$W_{\mathbf{u},h}(\mathbf{s}) = \frac{W(\mathbf{s}+h\mathbf{u}) - W(\mathbf{s})}{h}. \tag{13.1.4}$$

Clearly, for a fixed $\mathbf{u}$ and h, $W_{\mathbf{u},h}(\mathbf{s})$ is a well-defined process on $\Re^d$ – in fact, with $\delta = \mathbf{s} - \mathbf{s}'$, its covariance function is given by

$$C_{\mathbf{u}}^{(h)}(\mathbf{s},\mathbf{s}') = \frac{(2K(\delta) - K(\delta+h\mathbf{u}) - K(\delta-h\mathbf{u}))}{h^2} \tag{13.1.5}$$

whence $\mathrm{Var}\left(W_{\mathbf{u},h}(\mathbf{s})\right) = 2\left(K(\mathbf{0}) - K(h\mathbf{u})\right)/h^2$. The *directional derivative process* or *directional gradient process* is defined as $D_{\mathbf{u}}W(\mathbf{s}) = \lim_{h\to 0} W_{\mathbf{u},h}(\mathbf{s})$ when this mean-square limit exists. Indeed, when the parent process is mean-square differentiable, i.e. (13.1.3′) holds for every $\mathbf{s}_0$, then it immediately follows that, for each $\mathbf{u}$, $D_{\mathbf{u}}W(\mathbf{s})$ exists and equals $\langle \nabla W(\mathbf{s}), \mathbf{u}\rangle$ exists with equality again in the L_2 sense. In fact, under stationarity of the parent process, whenever the second-order partial and mixed derivatives of K exist and are continuous, $D_{\mathbf{u}}W(\mathbf{s})$ is a well-defined process whose covariance function is obtained from the limit of (13.1.5) as $h \to 0$, yielding

$$C_{\mathbf{u}}(\mathbf{s},\mathbf{s}') = -\mathbf{u}' H_K(\delta)\mathbf{u}, \tag{13.1.6}$$

where $H_K(\delta) = ((\partial^2 K(\delta)/\partial\delta_i\partial\delta_j))$ is the $d \times d$ Hessian matrix of $K(\delta)$.

More generally, collecting a finite set of m directions in $\Re^d$ into the $d \times m$ matrix $U = [\mathbf{u}_1, \ldots, \mathbf{u}_m]$, we can write the collection of directional derivatives as the $m \times 1$ vector, $D_U W(\mathbf{s}) = (D_{\mathbf{u}_1}W(\mathbf{s}), \ldots, D_{\mathbf{u}_m}W(\mathbf{s}))'$, so that $D_U W(\mathbf{s}) = U'\nabla W(\mathbf{s})$. In particular, setting $m = d$ and taking U as the $d \times d$ identity matrix (i.e., taking the canonical basis, $\{\mathbf{e}_1, \ldots \mathbf{e}_d\}$, as our directions), we have $D_I W(\mathbf{s}) = \nabla W(\mathbf{s})$ which gives a representation of $\nabla W(\mathbf{s})$ in terms of the *partial derivatives* of the components of $W(\mathbf{s})$. Explicitly, $\nabla W(\mathbf{s}) = \left(\frac{\partial W(\mathbf{s})}{\partial s_1}, \ldots, \frac{\partial W(\mathbf{s})}{\partial s_d}\right)'$, where $\mathbf{s} = \sum_{i=1}^d s_i \mathbf{e}_i$, so s_i's are the coordinates of $\mathbf{s}$ with respect to the canonical basis, and $D_U W(\mathbf{s}) = U' D_I W(\mathbf{s})$. Thus, the derivative process in a set of arbitrary directions is a linear transformation of the partial derivatives in the canonical directions. Inference about arbitrary directional derivatives can be built from this relationship and, in fact, only d directional derivatives are needed to learn about all directional derivatives. For instance, in say two dimensional space, we only need work with North and East directional derivatives processes in order to study directional derivatives in arbitrary directions.

Note that the linearity of $D_{\mathbf{u}}W(\mathbf{s})$ immediately reveals that $D_{-\mathbf{u}}W(\mathbf{s}) = -D_{\mathbf{u}}W(\mathbf{s})$. Furthermore, applying the Cauchy-Schwarz inequality to the directional derivative, for every unit vector $\mathbf{u}$, we obtain

$$|D_{\mathbf{u}}W(\mathbf{s})|^2 = |\langle \nabla W(\mathbf{s}), \mathbf{u}\rangle|^2 \le \|\nabla W(\mathbf{s})\|^2 = \sum_{i=1}^d D_{\mathbf{e}_i}^2 W(\mathbf{s}).$$

Hence, $\sum_{i=1}^{d} D_{\mathbf{e}_i}^2 W(\mathbf{s})$ is the maximum over all directions of $D_{\mathbf{u}} W(\mathbf{s})$. At location $\mathbf{s}$, it is achieved in the direction $\nabla W(\mathbf{s}) / \|\nabla W(\mathbf{s})\|$.

Formally, finite difference processes require less assumption for their existence. To compute differences we need not worry about formalizing a degree of smoothness for the realized spatial surface. However, issues of numerical stability can arise if h is too small. In practice, the nature of the data collection and the scientific questions of interest would often determine the choice of directional finite difference processes vs. directional derivative processes. In particular, in applications involving spatial data, scale is usually a critical question (e.g., in environmental, ecological or demographic settings). Infinitesimal local rates of change may be of less interest than finite differences at the scale of a map of inter-point distances. On the other hand, gradients are of fundamental importance in geometry and physics and researchers in the physical sciences (e.g., geophysics, meteorology, oceanography) often formulate relationships in terms of gradients. Data arising from such phenomena may require inference through derivative processes.

13.1.2 Mean Surface Gradients

From (13.1.2), we obtain the mean surface as

$$\mu(\mathbf{s}) = E[Y(\mathbf{s}) \,|\, \beta, \mathbf{w}(\mathbf{s})] = \mathbf{x}(\mathbf{s})' \tilde{\beta}(\mathbf{s}) = \mathbf{x}(\mathbf{s})' \beta + \mathbf{x}(\mathbf{s})' \mathbf{w}(\mathbf{s}),$$

where $\beta = (\beta_1, \ldots, \beta_p)'$ and $\mathbf{w}(\mathbf{s}) = (w_1(\mathbf{s}), \ldots, w_p(\mathbf{s}))'$. Consider the setting where $x_1(\mathbf{s}) \equiv 1$ and $x_l(\mathbf{s})$'s are each differentiable functions of $\mathbf{s}$ for $l = 2, \ldots, p$. This would immediately be the case in customary trend surface specifications. Also, in our application (Section 13.1.5), each $x_l(\mathbf{s}) = \|\mathbf{s} - \mathbf{s}_l^*\|$ is the Euclidean distance between $\mathbf{s}$ and a fixed location of interest, $\mathbf{s}_l^*$. We offer further discussion of this point below.

Applying the ∇ operator to the mean surface, and keeping in mind that $x_1(\mathbf{s}) \equiv 1$ is a constant, yields

$$\begin{aligned} \nabla \mu(\mathbf{s}) &= \sum_{l=2}^{p} \nabla x_l(\mathbf{s}) \beta_l + \sum_{l=1}^{p} \nabla \{x_l(\mathbf{s}) w_l(\mathbf{s})\} \\ &= \sum_{l=2}^{p} \nabla x_l(\mathbf{s}) \beta_l + \sum_{l=2}^{p} w_l(\mathbf{s}) \nabla x_l(\mathbf{s}) + \sum_{l=1}^{p} x_l(\mathbf{s}) \nabla w_l(\mathbf{s}). \end{aligned} \tag{13.1.7}$$

Following the discussion in Section 13.1.1, it is immediate from (13.1.2) that for any direction $\mathbf{u}$,

$$D_{\mathbf{u}} \mu(\mathbf{s}) = \mathbf{u}' \nabla \mu(\mathbf{s}) = \sum_{l=2}^{p} D_{\mathbf{u}} x_l(\mathbf{s}) \beta_l + \sum_{l=2}^{p} w_l(\mathbf{s}) D_{\mathbf{u}} x_l(\mathbf{s}) + \sum_{l=1}^{p} x_l(\mathbf{s}) D_{\mathbf{u}} w_l(\mathbf{s}).$$

Note that $D_{\mathbf{u}}\mu(\mathbf{s})$ is itself a stochastic process whose realizations will be determined by the set of parent processes $w_l(\mathbf{s})$'s as well as their corresponding gradient processes $\nabla w_l(\mathbf{s})$'s.

Evidently, smoothness of response surface realizations will be required to ensure the general existence of directional derivatives. While the smoothness of $x_l(\mathbf{s})$'s is determined from its assumed functional form, care is needed in addressing the existence of $\nabla w_l(\mathbf{s})$. Fortunately, smoothness of spatial process realizations can be secured through proper specification of the covariance function. Choice of covariance function to provide almost sure continuity of process realizations has been discussed by Kent (1989). For our purpose we will require mean square differentiability (Stein, 1999; Banerjee, Gelfand, and Sirmans, 2003). which is ensured by the existence of a second derivative at the origin for the covariance function, as (13.1.5) reveals. In particular the Matérn family of correlation functions is given by

$$\rho(\mathbf{s}_1,\mathbf{s}_2;\phi,\nu)=\frac{1}{2^{\nu-1}\Gamma(\nu)}(\|\mathbf{s}_1-\mathbf{s}_2\|\phi)^{\nu}\mathcal{K}_{\nu}(\|\mathbf{s}_1-\mathbf{s}_2\|;\phi);\ \ \phi>0,\ \nu>0,$$

where $\mathcal{K}_{\nu}(\cdot;\phi)$ is a modified Bessel function of the second type (e.g., Stein, 1999). The parameter ν controls the smoothness of the realized surface: $\nu\in[m,m+1)$ ensures that the process is m times mean square differentiable, but not $m+1$ times. So, we require $\nu\geq 1$.

It is worth pointing out that in many instances we will have certain predictors that do not change continuously over space. For instance, if $x_l(\mathbf{s})$ denoted crime-rate, it would be a tiled surface. Note that such variables are differentiable *almost everywhere* (a.e.) over the region, i.e., $D_{\mathbf{u}}x_l(\mathbf{s})$ exists and, in fact, equals 0 a.e. Another example could be "dummy variables" that are spatially referenced with respect to areal units or regions. These often appear in (13.1.2) $\sum_r \beta_r I(\mathbf{s}\in B_r)$ where B_r is the rth region and I is an indicator function. So, for such a regressor, the contribution to the gradient in (13.1.7) is 0. To clarify, certain types of regressors provide explanation for the mean surface but will appear only partially or not at all in the gradient analysis.

Finally, note that in some applications our outcome or response $Y(\mathbf{s})$ is on a different scale from the original quantity of interest. For instance, in our illustration in Section 13.1.5, $Y(\mathbf{s})$ is the log land value at location $\mathbf{s}$. So, if we consider $D_{\mathbf{u}}\mu(\mathbf{s})$ we are obtaining gradients for the mean log land value surface. Instead, we might wish to consider gradients for the mean on a transformed scale (e.g., the original scale), that is, $D_{\mathbf{u}}g(\mu(\mathbf{s}))$ where say $g(\cdot)=\exp(\cdot)$. Then, a simple application of the *chain rule* yields

$$D_{\mathbf{u}}g(\mu(\mathbf{s}))=\left\{\frac{d}{d\mu}g(\mu(\mathbf{s}))\right\}D_{\mathbf{u}}\mu(\mathbf{s}).$$

Note that $D_{\mathbf{u}}E(g(Y(\mathbf{s})))$ is not accessible here: a new model for $g(Y(\mathbf{s}))$ is required and different spatial processes would need to be introduced.

13.1.3 Posterior Inference for Gradients

13.1.3.1 Estimating Spatially Varying Regression Models

In order to provide the model specification in (13.1.2), we need to propose a multivariate spatial process model for $\mathbf{w}(\mathbf{s})$. Multivariate spatial processes, are completely characterized by their mean and a cross-covariance (matrix) function, $C_{\mathbf{w}}(\mathbf{s}_1,\mathbf{s}_2;\theta) = \text{Cov}\{\mathbf{w}(\mathbf{s}_1),\mathbf{w}(\mathbf{s}_2)\}$. Valid constructions using convolutions of kernels or correlation functions are possible (Ver Hoef and Barry, 1998; Gaspari and Cohn, 1999). An attractive, easily interpretable and flexible approach develops versions of the linear model of coregionalization (LMC) as in, e.g., Grzebyk and Wackernagel (1994), Wackernagel (2003), Schmidt and Gelfand (2003), Gelfand et al. (2004), and Reich and Fuentes (2007).

Specifically, we set $\mathbf{w}(\mathbf{s}) = A\mathbf{v}(\mathbf{s})$ where the components of $\mathbf{v}(\mathbf{s})$, $\{v_l(\mathbf{s}), l = 1,2,...p\}$ are independent spatial processes defined on D, with mean 0 and variance 1. Furthermore, we assume that $v_l(\mathbf{s})$ has an isotropic correlation function $\rho(\cdot,\phi_l)$, $l = 1,2,...p$. Thus, $\mathbf{v}(\mathbf{s})$ has a diagonal cross-covariance matrix, $C_{\mathbf{v}}(\|\mathbf{s}_1 - \mathbf{s}_2\|)$ with lth diagonal element as $\rho_l(\|\mathbf{s}_1 - \mathbf{s}_2\|;\phi_l)$. The resultant cross-covariance function, $C_{\mathbf{w}}(\mathbf{s}_1,\mathbf{s}_2) = \text{Cov}\{\mathbf{w}(\mathbf{s}_1),\mathbf{w}(\mathbf{s}_2)\}$, is a $p \times p$ matrix whose (i,j)th entry is given by $\sum_{k=1}^{p} a_{ik}a_{kj}\rho(\cdot,\phi_k)$. Note that $\mathbf{w}(\mathbf{s})$ is a multivariate isotropic process and $\text{var}\{\mathbf{w}(\mathbf{s})\} = AA'$. Therefore, without loss of generality, we may work with a lower triangular (Cholesky) form for $A = \{a_{ij}\}$. Parameters in the model include the global β vector, the lower triangular elements of A, the ϕ_l's and τ^2 which we collect into θ.

Now, with observations $\mathbf{Y} = (Y(\mathbf{s}_1),\ldots,Y(\mathbf{s}_n))'$ at locations $\mathbf{s}_i$, $i = 1,\ldots,n$, let $\mathbf{w} = (\mathbf{w}(\mathbf{s}_1),\ldots,\mathbf{w}(\mathbf{s}_n))'$ be the $np \times 1$ column vector of a coefficient process realization. The resulting covariance matrix for $\mathbf{w}$ is of the form:

$$\Sigma_{\mathbf{w}} = \tilde{A}\tilde{D}(\phi)\tilde{A}',$$

where $\tilde{A} = A \otimes I_{n\times n}$ and $\tilde{D}(\phi)$ is block diagonal with lth diagonal being the $n \times n$ matrix $R((\phi_l))$ whose (i,j)th element is given by $\rho_l(\|\mathbf{s}_i - \mathbf{s}_j\|;\phi_l)$ with $\delta_{ij} = \mathbf{s}_i - \mathbf{s}_j$. This corresponds to a Matérn correlation function with $\nu = 3/2$.

Marginalizing over the random effects $\mathbf{w}$ is helpful, leaving us to run the much lower dimensional MCMC algorithm for θ. Let $X' = (\mathbf{x}(\mathbf{s}_1),\ldots,\mathbf{x}(\mathbf{s}_n))$ be the $p \times n$ matrix and let $\tilde{X}$ be an $n \times np$ block diagonal matrix with its ith block entry given by $\mathbf{x}(\mathbf{s}_i)'$, $i = 1,\ldots,n$. Then the marginal likelihood is:

$$[\mathbf{Y} \mid \theta] \propto |\Sigma_Y(\theta)|^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2}(\mathbf{Y} - X\beta)\,\Sigma_{\mathbf{Y}}(\theta)^{-1}(\mathbf{Y} - X\beta)\right\}, \tag{13.1.8}$$

where $\Sigma_{\mathbf{Y}}(\theta) = \tilde{X}\Sigma_{\mathbf{w}}\tilde{X}' + \tau^2 I$ is $np \times np$.

Customary prior specifications include flat priors for the β_l's, inverse gamma $IG(a_{\tau^2}, b_{\tau^2})$ prior for τ^2, gamma $G(a_l, b_l)$ priors (mean = a_l/b_l) for each of the decay parameters ϕ_l, while those for the entries in the lower-triangular matrix A were induced from an inverted Wishart prior on AA' (also see Gelfand et al., 2004).

We note that the joint full conditional distribution for β is multivariate normal. For the rest of the parameters in θ, the full conditional distributions are non-standard, updated using a Metropolis-Hastings algorithm.

The posterior distribution of the spatial realization over the over observed locations is given by

$$[\mathbf{w} \mid \mathbf{Y}] = \int [\mathbf{w} \mid \theta, \mathbf{Y}][\theta \mid \mathbf{Y}], \tag{13.1.9}$$

where $[\mathbf{w} \mid \theta, \mathbf{y}]$ is the full-conditional distribution of $\mathbf{w}$ and easily seen to be multivariate normal. The integral in (13.1.9) can be computed using composition sampling. Once the posterior samples $\{\theta^{(t)}\}_{t=1}^{T}$ have been obtained (post-burn-in), we draw $\mathbf{w}^{(t)} \sim [\mathbf{w} \mid \theta^{(t)}, \mathbf{Y}]$ for $t = 1, \ldots, T$. The resulting set $\{\mathbf{w}^{(t)}\}_{t=1}^{T}$ constitutes the desired posterior sample. For an unobserved location, say $\mathbf{s}_0$, we can recover the posterior distribution of $\mathbf{w}(\mathbf{s}_0)$ as

$$[\mathbf{w}(\mathbf{s}_0) \mid \mathbf{Y}] = \int [\mathbf{w}(\mathbf{s}_0) \mid , \mathbf{w}, \theta][\mathbf{w} \mid \theta, \mathbf{Y}][\theta \mid \mathbf{Y}]. \tag{13.1.10}$$

Since $[\mathbf{w}(\mathbf{s}_0) \mid , \mathbf{w}, \theta]$ is again Gaussian, we compute (13.1.10) easily by sampling $\mathbf{w}(\mathbf{s}_0)^{(t)} \sim [\nabla \mathbf{w}(\mathbf{s}_0) \mid \mathbf{w}^{(t)}, \theta^{(t)}]$ for $t = 1, \ldots, T$.

13.1.3.2 Estimation of Mean Surface Gradients

Turning to gradients, we seek, for various locations and various directions, gradients for the $w_l(\mathbf{s})$'s although the primary interest rests upon gradients for the $\mu(\mathbf{s})$ surface – as in (13.1.7). Note that directional derivatives do not exist for the $Y(\mathbf{s})$ surface since the $\varepsilon(\mathbf{s})$ surface is not even continuous, let alone differentiable.

Let $\nabla' \mathbf{w}(\mathbf{s}) = (\nabla' w_1(\mathbf{s}), \ldots, \nabla' w_p(\mathbf{s}))$ denote the $1 \times pd$ vector comprising p blocks of the $1 \times d$ gradient vectors written as $\nabla' w_j(\mathbf{s})$'s. When $\mathbf{w}(\mathbf{s})$ is stationary, the $p(d+1) \times 1$ vector $\mathbf{z}(\mathbf{s}) = (\mathbf{w}(\mathbf{s})', \nabla' \mathbf{w}(\mathbf{s}))'$ is a stationary multivariate Gaussian process with cross-covariance matrix $C_{\mathbf{z}}(\delta)$

$$\begin{pmatrix} \mathrm{Cov}\{\mathbf{w}(\mathbf{s}), \mathbf{w}(\mathbf{s}+\delta)\} & \mathrm{Cov}\{\mathbf{w}(\mathbf{s}), \nabla \mathbf{w}(\mathbf{s}+\delta)\} \\ \mathrm{Cov}\{\nabla \mathbf{w}(\mathbf{s}), \mathbf{w}(\mathbf{s}+\delta)\} & \mathrm{Cov}\{\nabla \mathbf{w}(\mathbf{s}), \nabla \mathbf{w}(\mathbf{s}+\delta)\} \end{pmatrix} = \begin{pmatrix} C_{\mathbf{w}}(\delta) & -\nabla' C_{\mathbf{w}}(\delta) \\ \nabla C_{\mathbf{w}}(\delta) & -H_{C_{\mathbf{w}}}(\delta) \end{pmatrix}.$$

Here $C_{\mathbf{w}}(\delta)$ is the $p \times p$ cross-covariance matrix of $\mathbf{w}(\mathbf{s})$, $\nabla' C_{\mathbf{w}}(\delta)$ is the $p \times pd$ block matrix whose (i,j)th block is the $1 \times d$ matrix obtained by applying ∇' to the (i,j)th element of $C_{\mathbf{w}}(\delta)$, $\nabla C_{\mathbf{w}}(\delta)$ is its transpose, and $H_{C_{\mathbf{w}}}(\delta)$ is the $pd \times pd$ block matrix whose (i,j)th block is the $d \times d$ Hessian of the (i,j)th element of $C_{\mathbf{w}}(\delta)$. The above expression for $C_{\mathbf{z}}(\delta)$ is easily derived from the cross-covariance expressions for the corresponding finite difference process and letting $h \to 0$.

The cross-covariance matrix ensures the validity of the joint distribution $p(\mathbf{w}, \nabla \mathbf{w}(\mathbf{s}) \mid \theta)$. In fact, in our setting it is Gaussian. This is convenient for implementing predictive inference on not only the gradient of each element of $\mathbf{w}(\mathbf{s})$ at arbitrary points, say $\mathbf{s}_0$, but also for functions thereof, including the direction of the maximal gradient $(\nabla w_l(\mathbf{s}_0)/||\nabla w_l(\mathbf{s}_0)||)$ and the size of the maximal gradient $(||\nabla w_l(\mathbf{s}_0)||)$

for $l = 1, \ldots, p$. All such inference can proceed in posterior predictive fashion by computing

$$[\nabla \mathbf{w}(\mathbf{s}_0) \,|\, \mathbf{Y}] = \int [\nabla \mathbf{w}(\mathbf{s}_0) \,|\, \mathbf{w}, \theta][\mathbf{w} \,|\, \theta, \mathbf{Y}][\theta \,|\, \mathbf{Y}]. \tag{13.1.11}$$

Again, composition sampling easily delivers samples from $[\nabla \mathbf{w}(\mathbf{s}_0) \,|\, \mathbf{Y}]$. Since $[\nabla \mathbf{w}(\mathbf{s}_0) \,|\, , \mathbf{w}, \theta]$ is again Gaussian, we compute (13.1.11) easily by sampling

$$\nabla \mathbf{w}(\mathbf{s}_0)^{(t)} \sim [\nabla \mathbf{w}(\mathbf{s}_0) \,|\, \mathbf{w}^{(t)}, \theta^{(t)} \text{ for } t = 1, \ldots, T.$$

Finally, once we have obtained posterior samples of $\nabla \mathbf{w}(\mathbf{s}_0)^{(t)}$, we can directly obtain the posterior samples of $\nabla \mu(\mathbf{s})$, and hence of $D_{\mathbf{u}} \mu(\mathbf{s})$ for any direction $\mathbf{u}$, using (13.1.7).

13.1.4 Gradients under Spatial Dirichlet Processes

Here, we digress to briefly discuss a nonparametric extension of the foregoing development discussed in Guindani and Gelfand (2006). In some cases, both the Gaussian and stationarity assumptions will be viewed as inappropriate. Recently, Gelfand, Kottas, and MacEachern (2005) have proposed a spatial Dirichlet process (SDP) mixture model which adopts the distribution of a stochastic process as its base measure. This is assumed to be stationary and Gaussian; nevertheless the resulting process is nonstationary and the joint finite dimensional distributions are not normal. The use of the SDP specification to model the distribution of the spatial component in a spatial random effect model leads to a fully Bayesian semi-parametric approach that, for fitting purposes, relies on well-known results and algorithms developed for Dirichlet process (DP) mixing.

We first consider conditions under which the random surfaces sampled from a SDP are smooth. As might be expected, such conditions are related to the behavior of the base spatial process. We also consider the gradient processes associated with those random surfaces, obtaining induced distribution theory. In particular, we show that the directional finite difference and derivative processes are themselves samples from a SDP, whose base measure is the distribution of the corresponding gradient for the original base stochastic process.

A frequent approach for specifying random distributions is the Dirichlet process (DP) (Ferguson, 1973, 1974). In particular, given the space Θ (equipped with a σ-field $\mathscr{B}$), let $DP(\nu G_0)$ denote the DP, where $\nu > 0$ is a scalar (precision parameter) and G_0 a specified base distribution defined on $(\Theta, \mathscr{B})$. A random distribution function on $(\Theta, \mathscr{B})$ arising from $DP(\nu G_0)$ is almost surely discrete and admits the representation $\sum_{j=1}^{\infty} p_j \delta_{\theta_j^*}$, where δ_z denotes a point mass at z, $p_1 = q_1$, $p_j = q_j \prod_{r=1}^{j-1}(1 - q_r)$, $j = 2, 3, \ldots$, with $\{q_r, r = 1,2,...\}$ i.i.d. from Beta$(1, \nu)$ and independently $\{\theta_j^*, j = 1,2,...\}$ i.i.d. from G_0 (Sethuraman, 1994). In this notation

θ_j^* is assumed to be scalar or perhaps vector valued, the latter case leading to a multivariate DP.

To model $\mathbf{W}_D = \{W(\mathbf{s}) : s \in D\}$, following Gelfand, Kottas, and MacEachern (2005), one can conceptually extend θ_j^* to a realization of a random field by replacing it with $\theta_{j,D}^* = \{\theta_j^*(\mathbf{s}) : \mathbf{s} \in D\}$. For instance, G_0 might be a stationary GP with each $\theta_{j,D}^*$ being a realization from G_0, i.e., a surface over D. The resulting random distribution, G, for $\mathbf{W}_D$ is denoted by $\sum_{j=1}^{\infty} p_j \delta_{\theta_{j,D}^*}$ and the construction will be referred to as a spatial Dirichlet process model. Henceforth, $W_D|G \equiv \{W(\mathbf{s}), \mathbf{s} \in D|G\}$ denotes a field whose distribution is a given realization of G.

We say that $\mathbf{W}_D$ is mean square continuous at a point $\mathbf{s}_0$ if $E[W(\mathbf{s})^2|G] < \infty$ and $E[(W(\mathbf{s}) - W(\mathbf{s}_0))^2|G] \to 0$ as $||\mathbf{s} - \mathbf{s}_0|| \to 0$ with probability one. Again, to investigate the mean square continuity of samples from a SDP, we can limit the study to random surfaces drawn according to a G in S. One point is critical here. If we marginalize with respect to the unknown G, mean square continuity of $\mathbf{W}_D$ follows easily from mean square continuity of the base process, since $E[(W(\mathbf{s}) - W(\mathbf{s}_0))^2] = E_{G_0}[(\theta_1^*(\mathbf{s}) - \theta_1^*(\mathbf{s}_0))^2]$. However, we are interested in $\mathbf{W}_D|G$, the realized surface. Since $E[(W(\mathbf{s}) - W(\mathbf{s}_0))^2] = E\left\{E\left[(W(\mathbf{s}) - W(\mathbf{s}_0))^2|G\right]\right\}$, we might expect that mean square continuity of the base measure implies mean square continuity of $W_D|G$. But, since L_2 convergence does not imply a.s. convergence, for a given G, mean square continuity of the base measure is not enough to claim mean square continuity of the samples $\mathbf{W}_D$ from G. This is not totally unexpected, since any G is a discrete probability measure with probability one, and therefore we expect its smoothness properties to depend on the smoothness of the $\theta_{j,D}^*$, $j = 1, 2, \ldots$ which define the support of the realized G. In fact, we can show that if G_0 is a separable process a.s. continuous on a compact $K \subset D$, then, any random field $\mathbf{W}_D$ sampled from a SDP is mean square continuous on K (see Guindani and Gelfand, 2006).

We turn to mean square differentiability of a process arising from a SDP. For any given G, unit vector $\mathbf{u}$ and scalar $h > 0$, we consider the finite differences $W_{\mathbf{u},h}(\mathbf{s})$ and, as in Section 13.1.1, define the directional derivative $D_{\mathbf{u}}W(s)$ as the L_2 limit of the finite difference process with respect to G, i.e.

$$\lim_{h \to 0} E\left[\left(W_{\mathbf{u},h}(\mathbf{s}) - D_{\mathbf{u}}W(\mathbf{s})\right)^2|G\right] = 0, \tag{13.1.12}$$

if the limit exists. If $D_{\mathbf{u}}W(\mathbf{s})$ exists for all $\mathbf{s} \in D$, then we will denote the directional derivative process by $D_{\mathbf{u}}\mathbf{W}_D$. In particular, as above, if $D_{\mathbf{u}}W(\mathbf{s})$ is a linear function of $\mathbf{u}$, we say that $\mathbf{W}_D|G$ is mean square differentiable.

Again, if we marginalize with respect to the unknown G, the differentiability of $\mathbf{W}_D$ follows immediately from that of the base measure. However, given G, mean square differentiability of the base measure is not enough to conclude about mean square differentiability of $\mathbf{W}_D|G$. In fact, the latter relies on the analytical properties of the surfaces specifying the realized support of G, similar to our mean square continuity result above.

Some associated distribution theory. Let $\mathbf{W}_D$ be a random field sampled from a $SDP(\nu G_0)$ and $W_{\mathbf{u},h}(\mathbf{s})$ be the associated directional finite difference process. Then it is easy to prove that $W_{\mathbf{u},h}(\mathbf{s})$ also is a sample from a SDP with same precision parameter ν and with base measure the distribution of the finite difference process $\theta^*_{\mathbf{u},h}(\mathbf{s})$, say $G_0^{\mathbf{u},h}$. Therefore, the necessary distribution theory for the directional finite difference process is obtained from the general theory of the SDP.

Now consider the directional derivative process $D_{\mathbf{u}}W(\mathbf{s})$ and suppose that G_0 admits a directional derivative process for each $\mathbf{u}$. Let $G'_{0,\mathbf{u}}$ denote the distribution of the process $D_{\mathbf{u}}\theta^*(\mathbf{s})$. Then, we can show that $D_{\mathbf{u}}W(\mathbf{s})$ is a sample from a SDP with smooth parameter ν and base measure $G'_{0,\mathbf{u}}$. In symbols, $D_{\mathbf{u}}W(\mathbf{s})|G'_{\mathbf{u}} \sim G'_{\mathbf{u}}$ and $G'_{\mathbf{u}} \sim SDP(\nu G'_{0,\mathbf{u}})$.

In particular, if G_0 is mean square differentiable, then $D_{\mathbf{u}}W(\mathbf{s}) = \mathbf{u}'\nabla_W(\mathbf{s})$, where $\nabla_W(\mathbf{s})$ is a vector valued process, whose distribution is a realization from a SDP, defined for all Borel sets A as

$$P(\nabla_W(\mathbf{s}) \in A) = \sum_{j=1}^{\infty} p_j \, \delta_{\nabla_{\theta^*_j(\mathbf{s})}}(A),$$

according to Sethuraman's representation. Here, $\nabla_{\theta^*_j(\mathbf{s})} = (D_{\mathbf{e}_1}\theta^*(\mathbf{s}), \ldots, D_{\mathbf{e}_d}\theta^*(\mathbf{s}))$ is the vector of directional derivatives of G_0 with respect to an orthonormal basis set of directions $(\mathbf{e}_1, \ldots, \mathbf{e}_d)$. As in Section 13.1.1, if the base measure is mean square differentiable, it is possible to study the behavior of $D_{\mathbf{u}}W(\mathbf{s})|G'_{\mathbf{u}}$ in arbitrary directions by means of an orthonormal basis.

13.1.5 Illustration

We illustrate the above modeling in the context of urban land value gradients. There is a considerable theory and literature in urban land economics which focuses on the structure of land values with particular emphasis on gradients. See Majumdar et al. (2006) for a more comprehensive treatment of this application.

Olcott's Land Values Blue Book of Chicago, which has been published annually since the early 1900s, provides an oft-analyzed spatiotemporal data set of urban land values. We illustrate with the Olcott data for the year 1990. For each year, we take $p = 4$ in the model in (13.1.2), introducing distance-based regressors in addition to an intercept. We define $x_2(\mathbf{s})$ to be the distance from the Central Business District (CBD), $x_3(\mathbf{s})$ for distance from Midway airport, and $x_3(\mathbf{s})$ to be the distance from a Secondary Employment Center (SEC). Employment centers were identified using data obtained from the Northern Illinois Planning Commission (NIPC). NIPC reports the level of employment per one-half mile by one-half miles square (a quarter section) for the Chicago PMSA. Areas within the Chicago area were identified as employment centers if the total employment within a one mile radius of a quarter section was greater than 16,000. Within the employment sub-centers identified, we arbitrarily selected one of these employment sub-centers for our analysis.

Our models were fitted using a Markov chain Monte Carlo algorithm. We assumed flat priors for the β_l's (the global regression parameters) and an inverse-gamma $IG(2,1)$ prior for the variance τ^2. For the parameters in the four dimensional spatial-process, we assumed weakly informative gamma priors for the four correlation decay parameters (scaled to have a mean range of about half of the maximum distance), while an inverted-Wishart $IW(4, 0.01I_4)$ prior (with four degrees of freedom, and diagonal means $= 100$) for AA' induces the prior on A. Convergence was diagnosed by monitoring mixing of two parallel chains with over-dispersed starting values. With Gaussian likelihoods and weakly informative priors, convergence was diagnosed within 2000 iterations, with a further 1000 iterates retained from each chain as posterior samples.

In the interest of simple model comparison, we also fitted a version of (13.1.2) with only a spatially varying intercept (i.e., a usual spatial random effect), that is with constant coefficients for the distance-based regressors. We used the posterior predictive model selection criterion of Gelfand and Ghosh (1998). Letting G denote the goodness of fit term, P the penalty term and D the criterion value, for 1990, for the simpler model $G = .189$, $P = .511$ and $D = .700$ while for the spatially varying coefficient model, $G = .157$, $P = .521$, and $D = .678$. So, the full model is preferred and hence, below we present posterior inference under this model including the gradient analysis.

TABLE 13.1. Posterior estimates of model parameters for Olcott data: 1990.

Parameters	Percentiles 1990 50% (2.50%, 97.50%)
β_0	5.675 (5.346, 5.813)
β_1	-1.947 (-2.117,-1.806)
β_2	1.371 (1.106, 1.508)
β_3	1.241 (1.174, 1.533)
$\phi_0 \times 10^3$	0.539 (0.273, 0.869)
$\phi_1 \times 10^3$	0.451 (0.112, 0.792)
$\phi_2 \times 10^3$	0.626 (0.238, 0.900)
$\phi_3 \times 10^3$	0.466 (0.183, 0.846)
τ^2	0.779 (0.533, 1.317)
T_{00}	0.413 (0.271, 0.513)
T_{11}	0.814 (0.606, 0.991)
T_{22}	1.321 (0.574, 1.896)
T_{33}	0.973 (0.763, 1.089)
$T_{01}/\sqrt{(T_{00} * T_{11})}$	-0.412 (-0.495,-0.163)
$T_{02}/\sqrt{(T_{00} * T_{22})}$	-0.126 (-0.173, 0.151)
$T_{03}/\sqrt{(T_{00} * T_{33})}$	-0.176 (-0.364, 0.004)
$T_{12}/\sqrt{(T_{11} * T_{22})}$	0.047 (-0.242, 0.121)
$T_{13}/\sqrt{(T_{11} * T_{33})}$	0.025 (-0.190, 0.152)
$T_{23}/\sqrt{(T_{22} * T_{33})}$	-0.360 (-0.529, 0.005)

The posterior inference for 1990 are summarized in Table 13.1. The four global regression parameters reveal a negative impact on land value for the distance from

the CBD (land-value decreases for locations distant from the CBD), while an opposite effect (land value increasing with distance) is seen for the other two distances. The result related to Midway is not surprising when one considers that the sample is dominated by residential land values. Close proximity to areas of congestions and significant noise, such as Midway, leads to lower land values. At first, the positive effect on land values at greater distances from SEC seems counterintuitive until one considers the location of this employment sub-center. Its location is to the southeast of the CBD and approximately 1.6 km to the west of the shore of Lake Michigan. The positive parameter could reflect the proximity to the CBD, but as the analysis will show, it is more likely attributed to a lake effect found in this region.

Also shown are estimates of the correlation decay parameters, the measurement error τ^2 and the spatial variance-covariance parameters as appearing in the matrix $AA' = T$ (with the covariances converted to correlations). The relatively large contributions of T_{00} through T_{33} toward the variability justifies a rather rich spatial model for the data. Also, the tendency toward negative correlation between the intercept and the regression parameters (with some being significantly so) is expected.

Figure 13.1 displays the coefficient process surface for 1990. This is an image plot with overlaid contour lines indicating the levels. The rather rich distribution of contours seems to justify the use of the spatially varying coefficients. Figure 13.2 displays the posterior mean of the mean surface for 1990. In general, we find from the mean land value price surface that the surface's maximum value corresponds to the location of the CBD and that land prices fall as one moves further from the lake in both time periods. However, the results do not support the idea of a constant gradient over the entire urban area in either time period. In fact, the gradient's magnitude varies not only with location but also depends on the direction from the center of the city for which it is evaluated. The gradient is also found to be steepest close to the CBD and then flattens in all directions as distance increases.

The mean surface in Figure 13.2 suggests several directions to examine. The putative location of the CBD is at the intersection of the 447902.14038 Easting and 4636874.42216367 Northing near Lake Michigan and is at the conjunction of the four rays in the figure. The four rays denote the four directions in which we travel to understand gradient behavior. These directions arevindicated in the figures as NLk (Northern Lake vicinity), NW (Northwest of the CBD), SW (Southwest of the CBD), and SLk (Southern Lake vicinity). In Table 13.2 we examine several points at different distances (indicated as 0.5, 1.5, 3.0, and 6.0) from the CBD in kilometer units. For each ray, two fundamental directional gradients are evaluated: (i) as one moves *away* from the CBD *along* the ray, and (ii) the direction *normal* or orthogonal to the ray. Note that moving *towards* the CBD *along* a ray would be the negative of that evaluated in (i).

A closer look at Table 13.2 for the directional gradients evaluated at points along the different rays reveals that a strong gradient exists in all directions from the CBD; the gradient is steepest close to the CBD and tends to diminish as distance increases. Indeed, if we imagine an arc passing through the 0.5km points, the directional gradients are significantly negative and quite similar in magnitude. This slope, however, steadily decreases in significance when the arc is extended to pass through the 1.5,

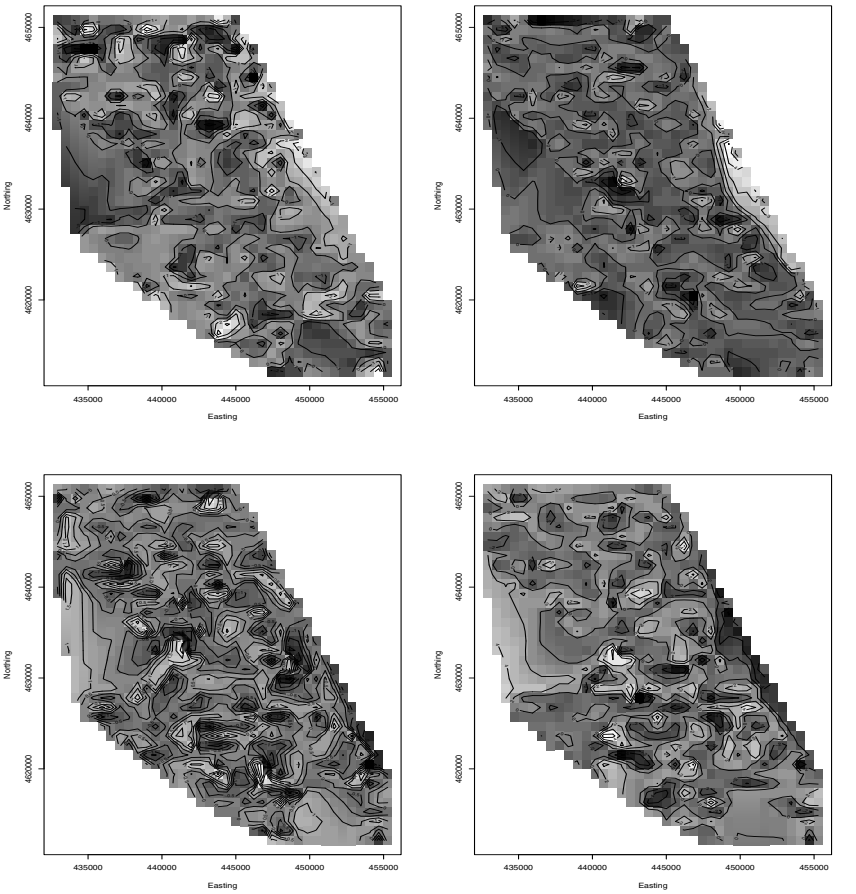


FIGURE 13.1. The coefficient process surfaces for 1990. Clockwise from top left: the intercept process $\beta_0(\mathbf{s})$, the CBD distance coefficient process $\beta_1(\mathbf{s})$, the Area1 distance coefficient process $\beta_2(\mathbf{s})$ and the Midway distance coefficient process $\beta_3(\mathbf{s})$.

3.0, and 6.0 km points. For example, significant gradients along the NW and SW rays are seen only until the 1.5 km points. No significant gradients are seen in the orthogonal directions to the NW and SW rays.

To further illuminate the inferential ability of our approach, two locations anticipated to reveal differential gradient behavior were chosen. These are a Secondary Employment Center (SEC) and a Secondary Population Center (SPC). Employment and population centers are important parts of the urban geography. The agglomeration economies associated with employment centers, as well as the clustering of individuals to benefit from neighborhood amenities, should lead to interesting land-value gradient patterns. The SEC is located at 450040.902 Easting,

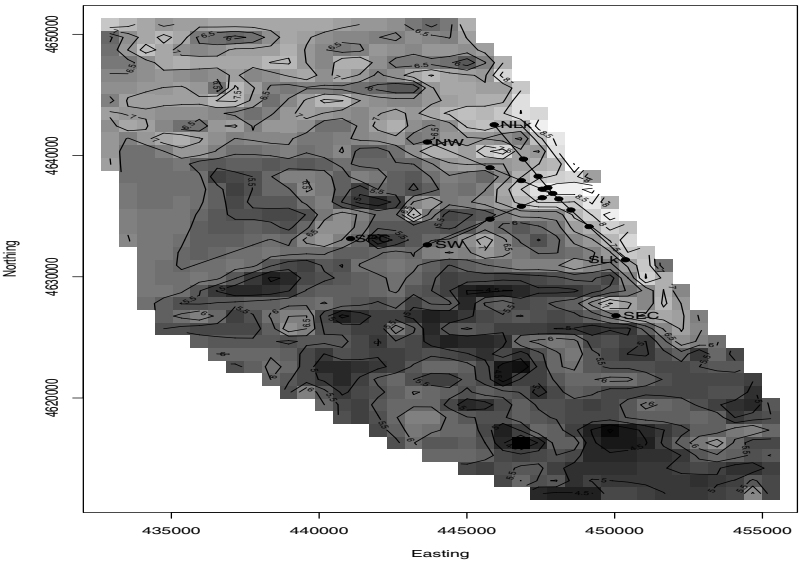


FIGURE 13.2. Posterior mean surface and locations for directional gradient analysis in the 1990 Olcott data.

TABLE 13.2. Directional derivative gradients at different directions related to 1990 data.

Points	Away from CBD Percentiles 50 2.5 97.50	Orthogonal to direction away from CBD Percentiles 50 2.5 97.50
NW0.5	-2.457 (-4.166,-0.776)	0.255 (-1.296, 1.670)
NW1.5	-2.165 (-3.818,-0.094)	0.110 (-1.386, 1.452)
NW3.0	-1.366 (-3.179, 0.440)	0.135 (-1.387, 1.539)
NW6.0	-1.221 (-3.008, 0.292)	0.208 (-1.284, 1.510)
SW.5	-2.307 (-3.851,-0.796)	-0.183 (-1.584, 1.166)
SW1.5	-2.078 (-3.536,-0.381)	-0.081 (-1.592, 1.663)
SW3.0	-0.492 (-1.939, 0.836)	1.033 (-0.492, 2.311)
SW6.0	-1.217 (-2.836, 0.063)	0.502 (-1.106, 1.996)
NLk.5	-2.212 (-3.974,-0.587)	0.324 (-1.305, 2.073)
NLk1.5	0.088 (-1.906, 1.603)	-1.333 (-3.046, 0.560)
NLk3.0	0.034 (-1.415, 1.280)	-0.967 (-2.738, 0.802)
NLk6.0	-1.044 (-2.371, 0.723)	-1.057 (-2.502, 0.302)
SLk.5	-2.493 (-4.197,-0.868)	0.036 (-1.631, 1.653)
SLk1.5	-0.170 (-1.951, 1.025)	-2.248 (-4.212,-0.524)
SLk3.0	-0.054 (-1.467, 1.657)	-2.586 (-3.902,-0.633)
SLk6.0	-0.150 (-1.632, 1.332)	-2.084 (-3.665,-0.754)

4626786.843 Northing and was used in part of the analysis above. The SPC is located at 441062.526 Easting, 4633147.015 Northing. Both are labelled in Figure 13.2. For each point, we obtained posterior distributions of the angle of the maximal gradient relative to the line from the point to the CBD as well as the posterior distribution of the difference between the maximal gradient and the gradient in the direction away from the CBD. These plots are shown in Figures 13.3 and 13.4.

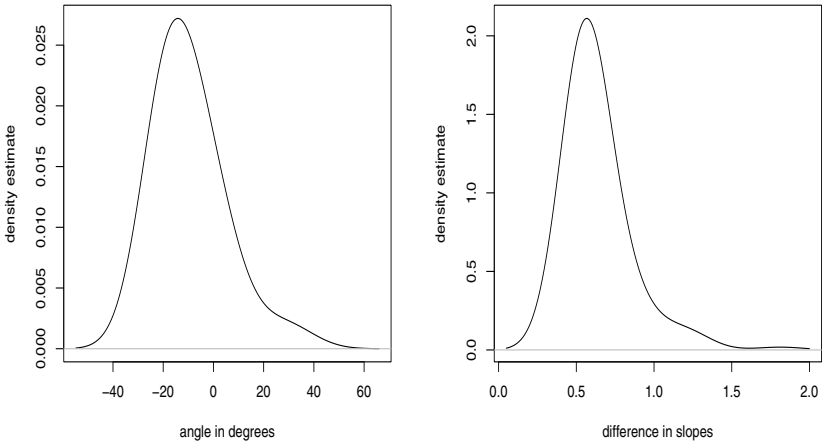


FIGURE 13.3. Density of angle between ray of maximum gradient and CBD in degrees and the absolute difference between values of maximal gradient and directional gradient from CBD for SPC in the Olcott data. See text for details.

Since these two points are rather far from the CBD, simple contour analysis or other descriptive methods are inadequate for formal evaluation of the above geometric quantities. However, using our sampling-based methods we find that SPC has maximal gradient direction not significantly different from the direction away from the CBD (contains 0). On the other hand, SEC, being in the southeast along the lake has a much more significant difference between the maximal gradient direction and the direction from the CBD. In support, Table 13.2 showed that the gradient quickly becomes flat along a ray moving south along the lake away from the CBD, while the gradient orthogonal to this direction stays significant. Thus, the gradient is larger in a westerly direction (90°), i.e., perpendicular to the lake. However, Figure 13.4 shows that the direction of maximal gradient is expected to be roughly southwest (45°) from the lake.

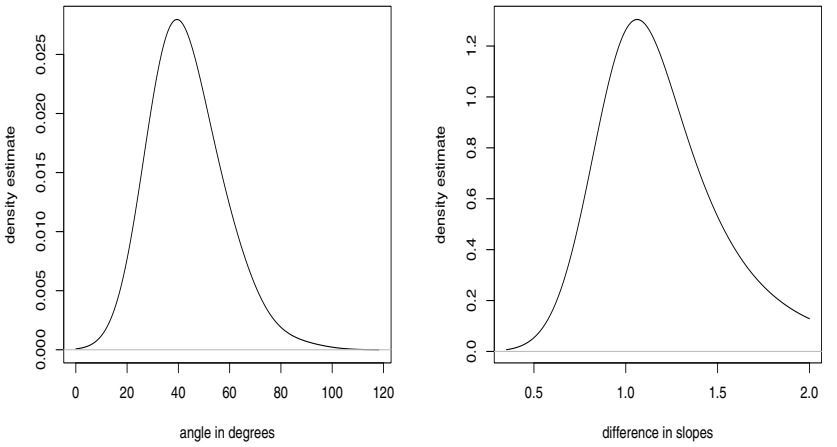


FIGURE 13.4. Density of angle between ray of maximum gradient and CBD in degrees and the absolute difference between values of maximal gradient and directional gradient from CBD for SEC. See text for details.

13.1.6 Concluding Remarks

Based upon our earlier work in Banerjee, Gelfand, and Sirmans (2003) and Majumdar et al. (2006), we have discussed a general theoretical approach, including full distribution theory, to examine gradients associated with spatially varying coefficient models. The theory is built through dependent spatial coefficient and intercept surfaces and yields a resulting mean surface. Gradients at arbitrary locations in arbitrary directions can be studied for each of these surfaces. Furthermore, the distribution of quantities that are functions of gradients, such as the direction of maximum gradient and the magnitude of the maximum gradient, can also be obtained. Our Bayesian framework enables this full range of inference. We illustrated with a study of land value gradients but this framework encompasses other potential applications where gradient analysis of the mean surface would be of interest, e.g., weather and pollution data modeling.

While our focus here have been on directional rates of change at points, local assessments of spatial surfaces are not restricted to points, but are often desired for curves and boundaries. For instance, environmental scientists are interested in ascertaining whether natural boundaries (e.g., mountains, forest edges, etc.) represent a zone of rapid change in weather and ecologists are interested in determining curves that delineate differing zones of species abundance. The above objectives require the notion of gradients and, in particular, assigning gradients to curves (*curvilinear gradients*) in order to identify curves that track a path through the region where the surface is rapidly changing. Such boundaries are commonly referred to as difference

boundaries or *wombling boundaries*, named after Womble (1951), who discussed their importance in understanding scientific phenomena (also see Fagan, Fortin, and Soykan, 2003). Recently, Banerjee and Gelfand (2006) formulated a Bayesian inferential framework for point-referenced curvilinear gradients or boundary analysis, a conceptually harder problem due to the lack of definitive candidate boundaries. Spatial process models help in estimating not only response surfaces, but residual surfaces after covariate and systematic trends have been accounted. Depending upon the scientific application, boundary analysis may be desirable on either. While Banerjee and Gelfand (2006) discussed only mean surfaces with spatially varying intercepts, their framework can be extended, fairly straightforwardly, to spatially varying regression models such as (13.1.1).

Acknowledgments: This work was carried out in the "Geostatistics" working group at the Statistics and Applied Mathematical Sciences Institute (SAMSI) in Research Triangle Park as a part of its "Program on Space-time Analysis for Environmental Mapping, Epidemiology and Climate Change". The work of the authors was partially supported by SAMSI and by National Institutes of Health (NIH) grant 1-R01-CA95995.

13.2 Non-Gaussian Hierarchical Generalized Linear Geostatistical Model Selection

Xia Wang, Dipak K. Dey, and Sudipto Banerjee

With the emergence of Geographical Information Systems (GIS), scientists and policy makers encounter spatially referenced data sets in a wide array of scientific disciplines. Very often, such data will be referenced over a fixed set of locations or points referenced by a coordinate system (latitude-longitude, Easting-Northing etc.). These are called *point referenced* or *geostatistical data*. Statistical theory and methods to model and analyze such data depend upon these configurations. The last decade has seen enormous developments in such modeling; see, for example, the books by Cressie (1993), Chilés and Delfiner (1999), Stein (1999), Møller (2003), Schabenberger and Gotway (2004), Banerjee, Carlin, and Gelfand (2004) and references therein for a variety of methods and applications.

The key ingredient in modeling geostatistical datasets is a *spatial process*. Spatial process models are widely used for inference in applied areas such as meteorology, environmental monitoring, ecological systems, forestry, econometrics and public health. Such models presume, for a region of study D, a collection of random variables $Y(\mathbf{s})$ where $\mathbf{s}$ indexes the locations in D. The set $\{Y(\mathbf{s}) : \mathbf{s} \in D\}$ can be viewed as a randomly realized surface over the region. In practice, this surface is only observed at a finite set of locations say $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n$. Inferential interest typically resides in estimation of the parameters of the spatial process model as well as in spatial interpolation or prediction (kriging) of the process over the entire domain.

Most of the existing methods on spatial process models have focused upon modeling continuous outcomes that can be reasonably treated as partial realizations of a Gaussian process or some transformation thereof. Heagerty and Lele (1998) considered a composite likelihood approach to binary spatial regression. A classic paper by Diggle, Tawn, and Moyeed (1998) discussed the use of spatial process models for non-Gaussian data within the framework of generalized linear models. They proposed incorporating the spatial process in the mean of the link function. Unlike in spatial process models for Gaussian outcomes, where proximity of locations would imply high spatial associations between the observations themselves, now we would expect the means to be more spatially associated for proximate locations. With non-Gaussian links, this association between the means does not translate to association between the outcomes. As an example, recently Finley, Banerjee, and McRoberts (2008) explored logistic regression models with spatially varying intercepts to classify forest attributes. The outcome variable was a binary process $Y(\mathbf{s}) = 1$ or $Y(\mathbf{s}) = 0$. Spatial association was induced by the logistic link on the *probability* $P(Y(\mathbf{s}) = 1)$. A probit link would of course lead to a similar interpretation. In related work, Fahrmeir and Lang (2001) and Kneib and Fahrmeir (2006) considered semiparametric regression with splines and Markov random fields to model spatial effects.

A key question that has hitherto received scant attention is the appropriateness of the link function in such models. In the latent variable approach, a link function on probability $P(Y(\mathbf{s})) = 1$, in fact, discretizes an unobservable continuous process $u(\mathbf{s})$, which is in turn specified from a distribution arising out of the chosen link function. The outcomes of $Y(\mathbf{s}) = 1$ or $Y(\mathbf{s}) = 0$ is actually induced by setting a threshold point within the distribution. For example, a probit link function assumes a normally distributed hidden continuous process and $Y(\mathbf{s}) = 1$ when $u(\mathbf{s}) > 0, Y(\mathbf{s}) = 0$ when $u(\mathbf{s}) \leq 0$, where $u(\mathbf{s})$ is the latent process and without loss of generality 0 is the threshold value.

The underlying process inducing the binary response $Y(\mathbf{s})$ may not always be symmetric as suggested by the logistic or probit links. In particular, Chen, Dey, and Shao (1999) suggested that when the underlying process has a very skewed distribution or, when the probability of a given binary response approaches 0 at a different rate than it approaches 1, the symmetric links, such as the logit or probit are inappropriate. With the high correlation among the spatial data points, overdispersion is inevitable. This overdispersion can be modeled by either the link function or the covariance structure within the data. The classical logit link assumes that the latent variable follows a symmetric logit distribution. Any other variation in the data might be absorbed into the assumed variance structure. This, however, may not be appropriate as the latent variable itself may follow a very skewed distribution. If the assumed covariance cannot incorporate these variances, wrongly assuming a logit model will likely result in an inferior fit for the data.

Thus, an essential question is now to allow flexibility in deciding the underlying process of $Y(\mathbf{s})$ when modeling spatial data. A very rich family which provides us with this flexibility, as discussed in detail in Section 13.2.2, is the generalized extreme value (GEV) distribution (Wang and Dey, 2009). With a free shape param-

eter, this family embeds a wide range of skewed processes and gives us an especially helpful tools to model extreme events discretized as binary response.

In the following discussion, we explore, both from a theoretical and computational standpoint, the possibility of more choices of the link functions, including the commonly used probit link, the complementary log-log link (Cloglog), and the new generalized extreme value (GEV) link. We further study how the different link models affect inference.

13.2.1 A Review on the Generalized Linear Geostatistical Model

Trans-Gaussian kriging (Cressie, 1993) has been one practical way to cope the non-Gaussian problem in geostatistical applications. Diggle, Tawn, and Moyeed (1998) originally proposed the generalized linear geostatistics model (GLGM), whose philosophy they claimed is "analogous to the embedding of the Gaussian linear model for mutually independent data within the framework of the generalized linear model ..." (Diggle, Tawn, and Moyeed, 1998). Diggle and Ribeiro Jr. (2007) also indicated that the root of the GLGM lies in the generalized linear model (GLM) (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989).

Under the framework of the GLM, continuous and discrete data are treated with a unified methodology for regression analysis. The generalized linear mixed model (GLMM), an important extension to the GLM, introduces unobservable random effects into the linear predictor. In particular, it specifies the systematic component in the GLM specification as $\eta_i = \sum \beta_i x_{ij} + \mathrm{w}_i$, where $\mathbf{w} = (\mathrm{w}_1, \ldots, \mathrm{w}_n)$ is from a zero-mean multivariate distribution. The GLMM not only provides a way to model the over-dispersion in the data, it can also easily incorporate various dependence structures among observations based on different practical contexts, such as its application in longitudinal studies. Details on this line of models and examples can be further found in Breslow and Clayton (1993).

Diggle, Tawn, and Moyeed (1998) presented the GLGM as a natural extension of the mixed model by assuming the random effects $\mathbf{w}(\mathbf{s}) = (\mathrm{w}(\mathbf{s}_1), \ldots, \mathrm{w}(\mathbf{s}_n))$ as the underlying Gaussian signal process at each of the sample locations $\mathbf{s}_i$, $i = 1, \ldots, n$. Particularly, the GLGM model is specified as follows:

(a) $\mathbf{w}(\mathbf{s}) = (\mathrm{w}(\mathbf{s}_1), \ldots, \mathrm{w}(\mathbf{s}_n))'$ follows an n-dimensional Gaussian distribution with $E[\mathbf{w}(\mathbf{s})] = 0$ and covariance matrix $\Sigma = \mathrm{cov}[\mathrm{w}(\mathbf{s}_i), \mathrm{w}(\mathbf{s}_j)]$. The covariance matrix Σ is a function of σ^2 and ϕ, where σ^2 represents the spatial process variance and ϕ is the range parameter that reflects the rate at which spatial association decreases as one considers observations farther apart. For example, $\Sigma = \sigma^2 H(\phi)$, where H is a correlation matrix with $H_{ij} = \rho(\mathbf{s}_i - \mathbf{s}_j; \phi)$ and ρ is a valid isotropic correlation function on R^2 indexed by a parameter (or parameters) ϕ. See Banerjee, Carlin, and Gelfand (2004, Chapter 2) for various covariance structures;

(b) The response $y(\mathbf{s}_i)$'s are conditionally independent given the $\mathrm{w}(\mathbf{s}_i)$'s. The conditional density of $\mathbf{Y}(\mathbf{s})\,|\,\mathbf{w}(\mathbf{s})$ is specified by the values of the conditional expectations $\mu(\mathbf{s}_i) = E\{y(\mathbf{s}_i)|\mathrm{w}(\mathbf{s}_i)\}$ as $f\{\mathbf{Y}(\mathbf{s})|\mathbf{w}(\mathbf{s})\} = \prod_{i=1}^{n} f\{y(\mathbf{s}_i);\mu(\mathbf{s}_i)\}$, where $i = 1,\ldots,n$; (c) $F^{-1}\{\mu(\mathbf{s}_i)\} = \mathbf{x}_i(\mathbf{s}_i)'\beta + \mathrm{w}(\mathbf{s}_i)$, for some known link function F^{-1}, explanatory variables $\mathbf{x}_i(\mathbf{s}_i)$ and parameters β.

When $y(\mathbf{s}_i)$ is a Bernoulli random variable with $p(\mathbf{s}_i) = P\{y(\mathbf{s}_i) = 1\}$ (note $p(\mathbf{s}_i) = \mu(\mathbf{s}_i) = E\{y(\mathbf{s}_i)\,|\,\mathrm{w}(\mathbf{s}_i)\}$), the GLGM is

$$F^{-1}(p(\mathbf{s}_i)) = \mathbf{x}(\mathbf{s}_i)'\beta + \mathrm{w}(\mathbf{s}_i). \tag{13.2.1}$$

The function $F^{-1}(p(\mathbf{s}_i)) = \log\{p(\mathbf{s}_i)/(1-p(\mathbf{s}_i))\}$ gives the logit link, which is a symmetric link for Bernoulli model. The logit link has been the default model used in most of current research on GLGMs for the binary response data. Alternative link functions have been suggested even starting from the discussion in Diggle, Tawn, and Moyeed (1998, p. 339). In the development page of the R software for the GLGM binary model, a possible development of another symmetric link, the probit link model, is also mentioned. The probit link model is achieved by setting $F^{-1}\{p(\mathbf{s}_i)\} = \Phi^{-1}\{p(\mathbf{s}_i)\}$, where Φ^{-1} is the inverse cumulative distribution function of $N(0,1)$ distribution. The asymmetric Cloglog link is specified as $F^{-1}\{p(\mathbf{s}_i)\} = -\log\{-\log(1-p(\mathbf{s}_i))\}$. With the variation in the data, a more flexible link function may be necessary in a GLGM model. We consider in this section the GEV link as $F^{-1}\{p(\mathbf{s}_i)\} = GEV^{-1}(1-p(\mathbf{s}_i);\mu=0,\sigma=1,\xi)$, where $GEV^{-1}(\cdot)$ is the inverse cumulative distribution function at $1-p(\mathbf{s}_i)$ of a GEV distribution with a location parameter $\mu = 0$, scale parameter $\sigma = 1$, and shape parameter ξ to be decided in the model. The GEV link has been recently introduced by Wang and Dey (2009). It includes the complementary log-log as a special case and it approximates a symmetric link with certain parameter values. We discuss this family of link functions in greater detail in next section.

13.2.2 Generalized Extreme Value Link Model

The generalized extreme value distribution is a family of continuous probability distributions developed within the extreme value theory to combine the Gumbel, Fréchet and Weibull families. It has a cumulative distribution function as follows:

$$G(x) = \exp\left[-\left\{1+\xi\frac{(x-\mu)}{\sigma}\right\}_{+}^{-\frac{1}{\xi}}\right], \tag{13.2.2}$$

where $\mu \in R$ is the location parameter, $\sigma \in R^{+}$ is the scale parameter, $\xi \in R$ is the shape parameter and $x_{+} = \max(x,0)$. A more detailed discussion on the extreme value distributions can be found in Coles (2001) and Smith (2003). Its advantage as a link function arises from the fact that the shape parameter ξ in model (13.2.2) purely controls the tail behavior of the distribution. When $\xi \to 0$, it gives the Gumbel

distribution with $G(x) = \exp[-\exp\{-(x-\mu)/\sigma\}]$, which is the least positively skewed distribution in the GEV class when ξ is non-negative.

In (13.2.1), assume the link function derived from the GEV distribution. Then

$$p(\mathbf{s}_i) = P\{y(\mathbf{s}_i) = 1\} = 1 - \exp\left\{(1 - \xi \mathbf{x}(\mathbf{s}_i)'\beta)_+^{-\frac{1}{\xi}}\right\} = 1 - GEV(-\mathbf{x}(\mathbf{s}_i)'\beta; \xi), \tag{13.2.3}$$

where $GEV(x;\xi)$ represents the cumulative probability at x for the GEV distribution with $\mu = 0$, $\sigma = 1$, and an unknown shape parameter ξ.

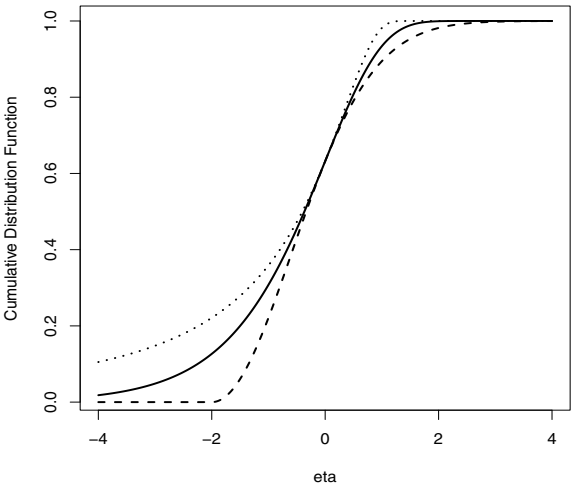


FIGURE 13.5. Cumulative distribution function plots of the GEV link with $\xi = -0.5, 0, 0.5$ from the right to the left.

Wang and Dey (2009) have shown that the GEV link model specified in (13.2.3) is negatively skewed for $\xi < \log 2 - 1$, and positively skewed for $\xi > \log 2 - 1$. The skewness varies with different shape parameters and a much wider range of skewness can be fitted compared to the commonly used Cloglog link. Figure 13.5 shows the response curves with ξ equal to -0.5, 0, and 0.5. The solid line is the response curve corresponding to the Cloglog link for $\xi \to 0$. The link is negatively skewed with ξ equal to -0.5 (dashed line) and positively skewed with ξ equal to 0.5 (dotted line). Wang and Dey (2009) further show that the GEV links provide much more flexible and improved skewed link regression models than the existing skewed links for independent binary response data, especially when there exists extreme difference between the number of 0's and 1's for which the response curve would be extremely skewed.

A symmetric link can be approximated by the class of the GEV links as a special case. For example, by matching the first 3 moments, the standard normal distribution can be well approximated by the GEV distribution with $\mu \approx -0.35579, \sigma \approx 0.99903$, and $\xi \approx -0.27760$. Figure 13.6 shows the quantile plots between the GEV model and the probit model. The plot is approximately a straight line between 0.02 and 0.98 quantiles. The discrepancy lies mainly in the tail area.

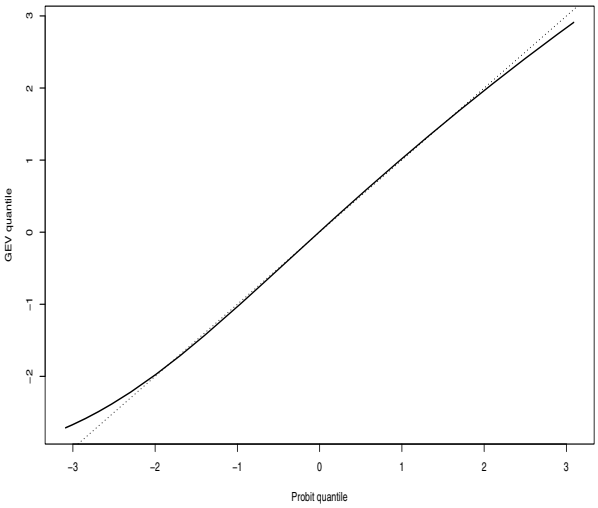


FIGURE 13.6. Plot of GEV quantiles with $\mu \approx -0.35579, \sigma \approx 0.99903$, and $\xi \approx -0.27760$ against probit quantiles for probabilities between 0.001 and 0.999. The solid line is the quantile plot, and the dotted line is the 45° reference line.

13.2.3 Prior and Posterior Distributions for the GLGM Model under Different Links

Let $D_{obs} = (n, \mathbf{y}, \mathbf{x})$ denote the observed data. Then from (13.2.1), the likelihood function for the GLGM model is given by

$$\begin{aligned} & L(\beta, \sigma^2, \phi, \mathbf{w}, \xi | D_{obs}, \mathbf{w}) \\ &= \prod_{i=1}^{n} [g^{-1}(\mathbf{x}(\mathbf{s}_i)'\beta + \mathbf{w}(\mathbf{s}_i))]^{y(\mathbf{s}_i)} [1 - g^{-1}(\mathbf{x}(\mathbf{s}_i)'\beta + \mathbf{w}(\mathbf{s}_i))]^{1-y(\mathbf{s}_i)}, \end{aligned}$$

where g^{-1} is the inverse of the link function.

The priors for $\beta, \mathbf{w}, \sigma^2$, and ϕ are specified hierarchically as follows. Normal priors with large variances are assumed for the linear regression parameters, that is, $\beta_j \sim N(0, c_0)$, where $j = 1, \ldots, p$ and $N(0, c_0)$ is a normal distribution with mean 0 and variance c_0. The prior for the spatial random effect $\mathbf{w}$ is $p(\mathbf{w} \mid \sigma^2, \phi) \sim MVN(\mathbf{0}, \Sigma)$, where $MVN(\mathbf{0}, \Sigma)$ is a multivariate normal distribution with a mean vector $\mathbf{0}$ and a variance-covariance matrix Σ. We assume the covariance structure of Σ is exponential, that is, the covariance function $\mathrm{C}(d_{ij}) = \sigma^2 \exp(-d_{ij}/\phi)$ when $i \neq j$, and $\mathrm{C}(d_{ij}) = \sigma^2$ when $i = j$ with d_{ij} denoting the distance between locations $\mathbf{s}_i$ and $\mathbf{s}_j$, $i, j = 1, \cdots, n$. We further assume that the signal variance σ^2 follows an inverse-gamma distribution given ϕ, that is $p(\sigma^2 | \phi) \sim IG(a = a_0, b = b_0)$ and $p(\phi) = p(\phi \mid d_0) \propto \exp(-d_0\phi)$. The GEV link has an additional shape parameter ξ. We assume that $p(\xi)$ is proper and further assume that $p(\xi) \propto N(0, \sigma_\xi^2)$, to allow a large flexibility for ξ.

For logit, probit and Cloglog links, the joint posterior distribution of $(\beta, \mathbf{w}, \sigma^2, \phi)$ based on D_{obs} is given by

$$\begin{aligned} p(\beta, \sigma^2, \phi, \mathbf{w} | D_{obs}) \propto & \prod_{i=1}^{n} [g^{-1}\{\mathbf{x}_i\beta + \mathbf{w}(\mathbf{s}_i)\}]^{y(\mathbf{s}_i)} [1 - g^{-1}\{\mathbf{x}_i\beta + \mathbf{w}(\mathbf{s}_i)\}]^{\{1-y(\mathbf{s}_i)\}} \\ & \times \exp\{-\beta'\beta/(2c_0)\} |\Sigma|^{-\frac{1}{2}} \exp(-\mathbf{w}'\Sigma^{-1}\mathbf{w}/2)(\sigma^2)^{-(a_0+1)} \\ & \times \exp\left\{-1/(\sigma^2 b_0)\right\} \exp\left(-d_0\phi\right). \end{aligned}$$

For the GEV link, the joint posterior distribution of $(\beta, \mathbf{w}, \sigma^2, \phi, \xi)$ based on the observed data D_{obs} is given by

$$\begin{aligned} & p(\beta, \sigma^2, \phi, \xi, \mathbf{w} | D_{obs}) \\ \propto & \prod_{i=1}^{n} (1 - GEV[-\{\mathbf{x}(\mathbf{s}_i)'\beta + \mathbf{w}(\mathbf{s}_i)\}; \xi])^{y(\mathbf{s}_i)} (GEV[-\{\mathbf{x}_i\beta + \mathbf{w}(\mathbf{s}_i)\}; \xi])^{\{1-y(\mathbf{s}_i)\}} \\ & \times \exp\{-\beta'\beta/(2c_0)\} |\Sigma|^{-\frac{1}{2}} \exp\left(-\mathbf{w}'\Sigma^{-1}\mathbf{w}/2\right) (\sigma^2)^{-(a_0+1)} \exp\left\{-1/(\sigma^2 b_0)\right\} \\ & \times \exp\left(-d_0\phi\right) \exp\left(-\xi^2/\sigma_\xi^2\right). \end{aligned}$$

In Section 13.2.4 we show the flexibility of the GEV model when the true underlying regression model is known with a simulated data. We also compare different link models using a real data set in Section 13.2.5. In these studies, we choose $c_0 = 1000$, $a_0 = 2$, $b_0 = 1$, $d_0 = 5$ and $\sigma_\xi^2 = 100$.

13.2.4 A Simulated Data Example

In this example we consider a complementary log-log regression simulation. Our primary aims are to (a) show the possible bias in the mean response estimates with link mis-specification and (b) test the flexibility of the GEV link models. We perform

a comprehensive Bayesian analysis for a given simulated dataset and evaluate the model comparison using the Deviance Information Criterion (DIC) measure.

First, we generate a realization of the Gaussian process $\mathbf{w} = (\mathrm{w}(\mathbf{s}_1), \ldots, \mathrm{w}(\mathbf{s}_n))'$ at equally spaced 15 grids within a unit square with $n = 225$, $\mu = 0$, $\sigma^2 = 1$, and exponential correlation function with $\phi = 0.1$ (See Figure 13.7). Then, we independently generate $x(\mathbf{s}_i) \sim N(0,1)$, $i = 1, 2, \ldots, n = 225$. A simulated dataset with 225 independent Bernoulli response variables, $y(\mathbf{s}_i)$'s, is then drawn with $p(\mathbf{s}_i) = 1 - \exp[-\exp\{\mathbf{x}(\mathbf{s}_i)'\beta + \mathbf{w}_i(\mathbf{s}_i)\}]$, where $\mathbf{x}(\mathbf{s}_i)'\beta = \beta_1 + x_i\beta_2$, $\beta_1 = 0$ and $\beta_2 = 4$. The simulated dataset is in fact from a GLGM model with the Cloglog link function. The resulting dataset contains 94 0's and 131 1's for the response variable in simulation. We fit the datasets with the symmetric logit, Cloglog and GEV links.

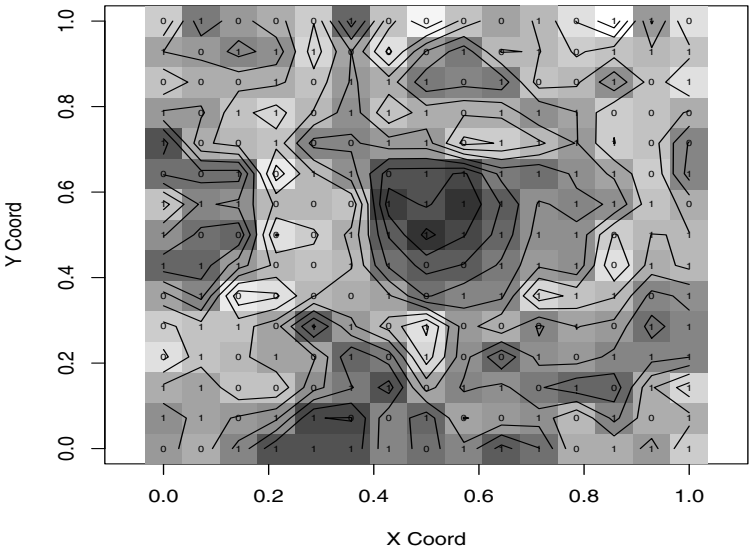


FIGURE 13.7. A simulation of the binomial GLGM model with complementary log-log link. The numbers are the values of Bernoulli random variable corresponding to locations at the center of each grid square. The gray scale represents the value of the underlying Gaussian process at each location.

Table 13.3 presents the posterior estimates from the simulated data. The simulated GEV model yields estimates of the regression coefficient $\beta = (\beta_1, \beta_2)$ that are almost identical to the true complementary log-log regression model. However, the symmetric logit model presents several problems. In particular, the mean estimates of β are significantly different from the true value of β with the true values of β

not being included in the 95% interval. The variations in the estimates of β and the variance parameter are relatively larger compared to those in the GEV model.

TABLE 13.3. Posterior estimates under different link models for the simulated data.

Parameter	Model	Mean	Standard Deviation	95% Interval
$\beta_1 = 0$	Logit	1.441	0.337	**(0.833, 2.161)**
	Cloglog	0.391	0.205	(-0.027, 0.771)
	GEV	0.135	0.311	(-0.535, 0.641)
$\beta_2 = 4$	Logit	6.724	1.050	**(5.021, 9.120)**
	Cloglog	5.101	0.694	(3.879, 6.588)
	GEV	4.852	0.892	(3.604, 6.932)
$\sigma^2 = 1$	Logit	1.122	0.129	(0.891, 1.404)
	Cloglog	1.311	0.121	(0.923, 1.394)
	GEV	0.880	0.094	(0.710, 1.083)
$\xi = 0$	GEV	0.139	0.248	(-0.234, 0.708)

The DIC value for the GEV model (99.47) is very close to that of the Cloglog model (98.01) and both the GEV model and the Cloglog model provide better fit than the logit model (DIC=101.20) based on the DIC measure. This result is consistent with the fact that the Cloglog model is a special case of the GEV model and the simulated data is in fact based on a GEV regression model with $\xi = 0$.

13.2.5 Analysis of Celastrus Orbiculatus Data

The data were collected from 603 locations in Connecticut with presence or absence of species Celastrus Orbiculatus, along with some environmental predictors. The outcome variable $Y(\mathbf{s})$ is a presence-absence binary indicator (0 for absence) for Celastrus Orbiculatus at location $\mathbf{s}$. Figure 13.8 shows the map of the outcome variable at 603 locations, where the solid dot represents the presence of the species and the circle represents the absence of the species at a location. Out of the 603 observations, there are 251 observations with $Y(\mathbf{s})$ equal to 1 and 352 observations equal to 0. There are three predictors with multiple nominal categories: habitat class (representing the current state of the habitat) of four different types, land use and land cover (LULC) types (Land use/cover history of the location; e.g., always forest, formerly pasture now forest, etc.) at five levels, a year 1970 category number (LULC at one point in the past: 1970; e.g., forest, pasture, residential, etc.) with six levels. In addition we have an ordinal covariate, canopy closure percentage (percent of the sky that is blocked by "canopy" of leaves of trees), a binary predictor for heavily managed points (0 if "no"; "heavy management" implies active landscaping or lawn mowing) and a continuous variable measuring the distance from the forest edge in the logarithm scale. A location under mature forest would have close to 100% canopy closure while a forest edge would have closer to 25% with four levels in increasing order.

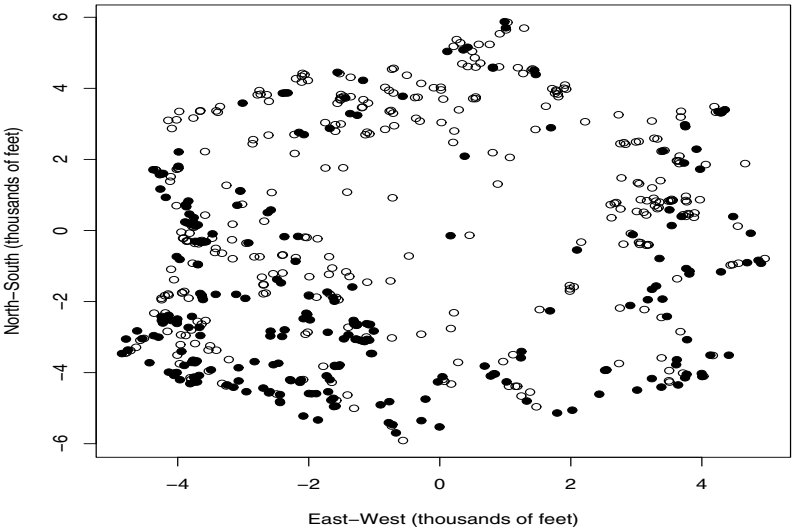


FIGURE 13.8. Celastrus Orbiculatus presence or absence Data. The solid dot represents the presence of the species and the circle represents the absence of the species at a location.

The initial runs using the built-in function `binom.krige.bayes` in the **geoRglm** package (Christensen and Ribeiro Jr., 2002) in **R** have shown strong spatial correlation within the data. We then carry out Bayesian analysis as specified in Section 13.2.3. Table 13.5 shows the 0.025, 0.5, and 0.975 quantile of the parameter estimation and the DIC measure. Based on the DIC measure, the GEV link and the Cloglog models give a better fit than the symmetric probit and logit link models. The shape parameter ξ is significantly positive. To have a rough look at how the fitted values compared with the observed pattern in the response variable, we plot surfaces of the presence of Celastrus Orbiculatus versus the fitted probabilities from the four spatial regression models (Figure 13.9).

TABLE 13.4. Cross-classification of Celastrus Orbiculatus presence and the habitat classes (HC).

	HC=1	HC=2	HC=3	HC=4
Not present ($\mathbf{Y}(\mathbf{s}) = 0$)	48	147	10	147
Present ($\mathbf{Y}(\mathbf{s}) = 1$)	82	71	3	95

We observe some difference in the parameter estimation between the GEV model and the other three models, not only in the magnitude but in the statistical significance. Table 13.4 is the cross-classification of the Habitat Class and the Presence. In the regression model, HabitatClass=1 is used as the reference. The table shows

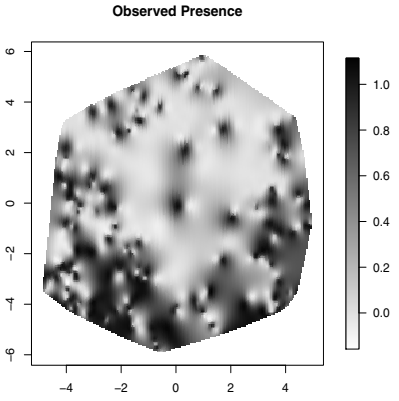
that the probability of observing the species at the Habitat Class 1, which is 0.63, is higher than those at other three habitat classes, which are 0.33, 0.23, and 0.39 for habitat classes 2, 3, and 4, respectively. The p-value for the Chi-square test of independence between presence and habitat class is less than 0.0001. Thus, the data show that there is a significant correlation between the presence of the species and the habitat class. In the logit, probit and even the Cloglog model, this association is mostly detected as insignificant, while the GEV link model estimates the effect as all negatively significant. Similar results are observed for LULCChange=2, cat1970=3, and cat1970=5, where the significance of the estimated parameters differs between the GEV model and the other three models.

TABLE 13.5. Model comparison under logit, probit, Cloglog, and GEV links for the GLGM models.

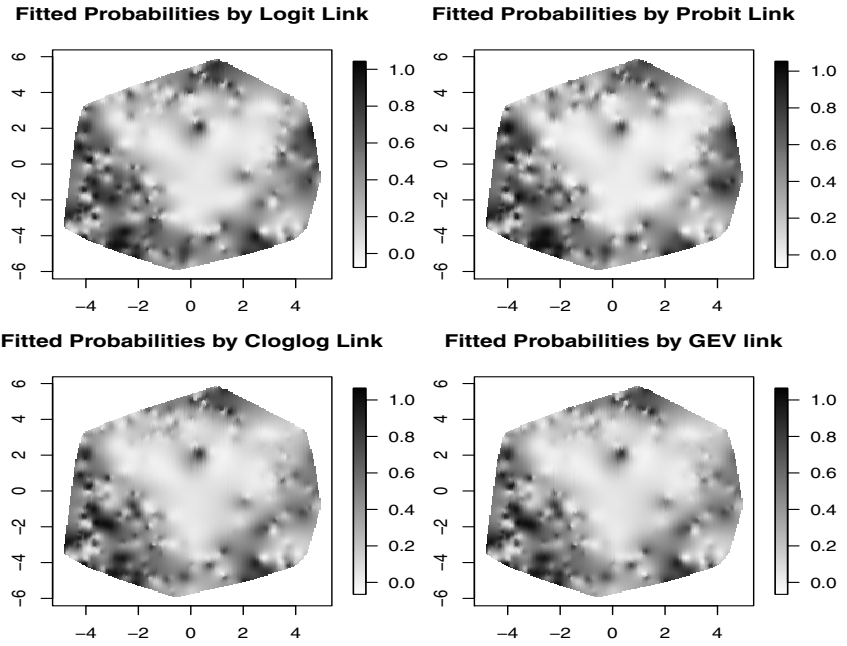
	Logit			Probit			Cloglog			GEV		
Variables	0.50	0.025	0.975	0.50	0.025	0.975	0.50	0.025	0.975	0.50	0.025	0.975
Intercept	-2.38	-3.44	-1.53	-1.43	-1.88	-0.84	-1.84	-2.43	-1.19	-6.13	-7.08	-5.35
HabitatClass												
2	-0.60	-1.40	0.12	-0.30	-0.66	0.13	-0.32	-0.79	0.05	-2.28	-2.88	-1.56
3	-0.58	-2.10	0.87	-0.35	-1.29	0.53	-0.71	-1.93	0.55	-8.01	-10.60	-4.78
4	-0.37	-1.01	0.37	-0.34	-0.64	-0.04	-0.19	-0.53	0.30	-0.85	-1.63	-0.22
LULCChange												
2	0.84	-0.07	2.06	0.65	-0.12	1.32	0.66	-0.35	1.31	3.46	2.20	4.90
3	1.50	0.67	2.43	0.95	0.41	1.48	1.00	0.22	1.52	3.18	1.78	3.98
4	1.71	1.02	2.57	1.05	0.48	1.50	1.30	0.67	1.70	3.92	3.19	5.02
5	2.85	1.79	4.17	1.99	1.18	2.60	1.96	1.24	2.71	4.46	3.19	5.78
cat1970												
2	-15.37	-41.32	-0.22	-7.26	-14.24	-1.69	-38.24	-57.21	-29.44	-32.67	-56.86	-17.27
3	-0.28	-1.18	0.59	0.03	-0.42	0.45	-0.35	-1.02	0.14	-2.15	-4.72	-0.14
4	0.99	0.00	1.81	0.44	0.10	0.85	0.16	-0.32	0.68	0.90	0.14	1.62
5	-0.53	-1.31	0.13	-0.34	-0.73	0.07	-0.87	-1.28	-0.48	-2.61	-4.34	-1.22
6	1.30	0.60	2.20	0.91	0.44	1.25	0.65	0.15	1.13	1.17	0.13	2.26
CanopyClosure	0.51	0.17	0.94	0.36	0.12	0.47	0.31	0.11	0.55	0.97	0.68	1.21
HeavilyManagedPts	-2.52	-4.21	-1.30	-1.29	-1.88	-0.64	-2.45	-3.80	-1.30	-10.77	-18.30	-6.65
LogEdgeDistance	-0.83	-1.11	-0.55	-0.53	-0.68	-0.37	-0.57	-0.77	-0.38	-0.47	-0.73	-0.24
σ^2	1.03	0.71	1.35	0.80	0.57	1.61	0.91	0.62	1.40	1.89	1.30	2.27
ϕ	0.13	0.09	0.16	0.11	0.08	0.22	0.14	0.09	0.22	0.09	0.07	0.11
ξ	-	-	-	-	-	-	-	-	-	2.18	1.58	2.46
DIC	679.89	—	—	676.05	—	—	675.841	—	—	673.51	—	—
p	19.56	—	—	26.81	—	—	19.10	—	—	7.74	—	—

Besides the difference in the mean estimates of the parameter, we also notice that there are some non-trivial differences in the estimation for the variance parameter σ^2 and the range parameter ϕ. The estimation of higher σ^2 and lower ϕ in the GEV link model implies higher spatial variance and a lower spatial range (hence weaker spatial association) compared to the estimation by other link models. For example, the effect range at the posterior median for the GEV model is 0.27, while it is 0.39 for the logit link.

To test the predictive power of the four different link models, we randomly divide the data into training and hold-out parts, with 10% locations (which gives 68 observations) as the hold-out data points. For prediction at the holdout locations, the goal is to generate samples from the conditional distribution of $(\mathbf{w}(\mathbf{s}), \mathbf{w}(\mathbf{s}^\star))$ given $\mathbf{Y}$, where $\mathbf{w}(\mathbf{s})$ and $\mathbf{w}(\mathbf{s}^\star)$ are the spatial random variables at the training and



(a) Observed response



(b) Fitted probability

FIGURE 13.9. The surfaces of the observed presence and the fitted probability of Celastrus Orbiculatus.

hold-out locations, respectively. As suggested in Diggle and Ribeiro Jr. (2007), this can be achieved by drawing a random sample from the multivariate Gaussian distribution $[\mathbf{w}(\mathbf{s}^\star)|\mathbf{Y},\beta,\mathbf{w}(\mathbf{s}),\sigma^2,\phi]$ where $\beta,\mathbf{w}(\mathbf{s}),\sigma^2$, and ϕ are the values generated from previous MCMC samples based on the training data. Since $\mathbf{w}(\mathbf{s}^\star)|\mathbf{w}(\mathbf{s})$ is independent of $\mathbf{Y}$ and β, it is then equivalent to sample from the following multivariate Gaussian distribution

$$[\mathbf{w}(\mathbf{s}^*)|\mathbf{w}(\mathbf{s})] \sim MVN\left(\Sigma_{21}\Sigma_{11}^{-1}\mathbf{w}(\mathbf{s}), \Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right), \tag{13.2.4}$$

where $\Sigma_{11}=\text{Var}(\mathbf{w}(\mathbf{s}))$, $\Sigma_{12}=\Sigma_{21}=\text{Cov}(\mathbf{w}(\mathbf{s}),\mathbf{w}(\mathbf{s}^\star))$, and $\Sigma_{22}=\text{Var}(\mathbf{w}(\mathbf{s}^\star))$. The deviance ($\hat{D}$) for the hold-out data part is then calculated as

$$\hat{D}=\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}D(\beta^{(i)},\sigma^{2(i)},\phi^{(i)},\mathbf{w}(\mathbf{s}^\star)^{(i,j)}), \tag{13.2.5}$$

where $\beta^{(i)},\sigma^{2(i)}$, $\phi^{(i)}$ are the i^{th} MCMC sampling from the training data and $\mathbf{w}(\mathbf{s}^\star)^{(i,j)}$ is the j^{th} simulated spatial random variables $\mathbf{w}(\mathbf{s}^\star)$ generated at the hold-out locations by (13.2.4) using the parameters $\sigma^{2(i)}$, $\phi^{(i)}$ in the i^{th} MCMC sampling for the training data. We set $M=1000$ and $N=25000$.

Table 13.6 shows the result. Although the DIC measure for the training data still indicates that the GEV model gives the best fit, the GEV model has the largest predictive deviance for the hold-out data. By closely looking at the data, we notice that large deviance is caused by the almost opposite fitted probability compared to the observed response at some locations. For example, Celastrus Orbiculatus were not observed at three locations but the GEV model predicted the probability of presence is very close to 1. The other three models also have a high fitted probability in these three locations but the fitted value is not very close to 1. The GEV model does not show advantage in predictive inference in the Celastrus Orbiculatus data.

TABLE 13.6. Model comparison under different links for the GLGM model: posterior deviance.

	Logit	Probit	Cloglog	GEV
DIC (training)	629.40	624.92	608.31	604.60
Posterior Deviance (holdout)	81.96	104.99	101.69	124.25

13.2.6 Discussion

This section introduces a new flexible skewed link model for analyzing spatial binary response data with covariates. The GEV link model provides great flexibility in fitting skewness in the response curve. We show that bias in the mean response estimates may result from the misspecification of the link functions in the GLGM

model. We investigate the application of the flexible GEV link models in spatial data scenario. Though we have seen that the choice of link models affects not only the estimation of the linear parameter but also that of the spatial parameters, it needs further study on how these two components interact in statistical inference.

As discussed in Reich, Hodges, and Zadnik (2006), introducing additional parameters, like the shape parameter ξ in the link function discussed here, may raise identifiability concerns. It needs further investigation on the conditions for identifying the parameters in the GEV model.

Acknowledgments: The authors wish to thank Jenica Allen and Dr. John A. Silander, Jr. for providing the data. Dr. Banerjee's work was partially sponsored by NSF-DMS-0706870 grant.

13.3 Objective Bayesian Analysis for Gaussian Random Fields

Victor De Oliveira

Gaussian random fields are useful mathematical tools for modeling spatially varying phenomena and are often the default model of choice (possibly after a transformation of the data). Their use has become standard in diverse areas of science such as epidemiology, geography, geology, and hydrology, where collection and analysis of spatial data are increasingly common tasks.

The Bayesian approach for the analysis of geostatistical data using Gaussian random fields was pioneered by Kitanidis (1986), Le and Zidek (1992) and Handcock and Stein (1993), while later developments include De Oliveira, Kedem, and Short (1997) and Ecker and Gelfand (1997). These works specify prior distributions for the model parameters using a combination of intuition and ad-hoc methods, but no systematic study was undertaken. But specification of prior distributions for these models is a somewhat challenging task. First, it is difficult to carry out subjective elicitation of these prior distributions, either because of lack of prior information or difficulties in interpreting some of the parameters. For instance, the so-called 'range parameters' have units of distance, so any sensible prior for this parameter must take into account the dimensions of the region under study. Second, it is often the case that little or no information is available about parameters controlling the 'smoothness' of the random field. Finally, naive specification of the prior distribution may give rise to improper posterior distributions.

Berger, De Oliveira, and Sansó (2001) provided an extensive discussion on theoretical issues involved in the Bayesian analysis of Gaussian random fields, and initiated the work on objective (default) Bayesian methods for the analysis of these models. They studied in detail the case when the correlation function depends on a single parameter, this being a 'range parameter'. For this model Berger, De Oliveira, and Sansó (2001) derived Jeffreys and reference prior distributions, the main properties of the resulting posterior distributions, and uncovered several interesting and un-

usual behaviors. First, it was shown that previously used naive priors yield improper posteriors, and the same holds for the commonly often prescribed independence Jeffreys prior when the random field has a constant term in the mean function. This is mainly due to an unusual behavior of the integrated likelihood of correlation parameters. Second, it was found that the up to then 'standard' reference prior algorithm, proposed by Berger and Bernardo (1992a) and obtained by asymptotic marginalization, agrees with the independence Jeffreys prior so it also yields an improper posterior. Finally, they derived a reference prior obtained by exact marginalization and showed that it always yields a proper posterior distribution. An additional desirable property of this reference prior is that, in spite of being improper, it can be used for correlation function selection among models with the same mean structure. Extensions of this initial work and applications to related models include Paulo (2005), De Oliveira (2007, 2010), Ferreira and De Oliveira (2007), De Oliveira and Song (2008) and Ren, Sun, and He (2009).

This section provides a review of the main results obtained in the last decade on objective (default) Bayesian methods for the analysis of spatial data using Gaussian random fields. The section ends with a discussion on the success (or lack of) these models and some relevant open problems.

13.3.1 Gaussian Random Field Models

Consider a phenomenon that varies spatially *continuously* over a region $\mathscr{D} \subset \mathbb{R}^l$, with $l \in \mathbb{N}$. It is assumed that this variation is modeled by a Gaussian random field $\{Y(\mathbf{s}) : \mathbf{s} \in \mathscr{D}\}$ with

$$E\{Y(\mathbf{s})\} = \sum_{j=1}^{p} \beta_j f_j(\mathbf{s}) \quad \text{and} \quad \text{cov}\{Y(\mathbf{s}), Y(\mathbf{u})\} = \sigma_y^2 K(\mathbf{s}, \mathbf{u}),$$

where $f_1(\mathbf{s}), \ldots, f_p(\mathbf{s})$ are known location-dependent covariates, $\beta = (\beta_1, \ldots, \beta_p)' \in \mathbb{R}^p$ are unknown regression parameters, $\sigma_y^2 = \text{var}\{Y(\mathbf{s})\} > 0$ unknown and $K(\mathbf{s}, \mathbf{u})$ is a correlation function in $\mathbb{R}^l$. For this class of models spatial association is specified marginally through correlation functions, where the most commonly used models are isotropic correlation functions, meaning that $K(\mathbf{s}, \mathbf{u}) = K(d)$, with $d = ||\mathbf{s} - \mathbf{u}||$ (the Euclidean distance between $\mathbf{s}$ and $\mathbf{u}$). Numerous isotropic correlation functions have been proposed in the literature (see Cressie (1993) for listings of the most common models). We consider here as an illustrative example the Matérn family of isotropic correlation functions given by

$$K(d) = \frac{1}{2^{\theta_2 - 1}\Gamma(\theta_2)} \left(\frac{d}{\theta_1}\right)^{\theta_2} \mathscr{K}_{\theta_2}\left(\frac{d}{\theta_1}\right); \quad \theta_1 > 0,\ \theta_2 > 0, \tag{13.3.1}$$

where $\Gamma(\cdot)$ is the gamma function and $\mathscr{K}_{\theta_2}(\cdot)$ is the modified Bessel function of the second type and order θ_2. For this family (and many others), θ_1 controls (mainly)

how fast the correlation decays with distance (the so-called *range* parameter), which has units of distance, and θ_2 controls geometric properties of the random field, such as mean square differentiability (the so-called *smoothness* or *roughness* parameter), which is unitless.

The data $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ consist of possibly noisy measurements of the random field taken at known sampling locations $\mathbf{s}_1, \ldots, \mathbf{s}_n \in \mathscr{D}$, where

$$Y_i = Y(\mathbf{s}_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$

$\{\varepsilon_i\}_{i=1}^n \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma_\varepsilon^2)$ represent measurement errors independently distributed of the random field $Y(\cdot)$, and $\sigma_\varepsilon^2 \geq 0$ (the so-called *nugget* parameter). In this case we have

$$E\{Y_i\} = \sum_{j=1}^{p} \beta_j f_j(\mathbf{s}) \quad \text{and} \quad \mathrm{cov}\{Y_i, Y_j\} = \sigma_y^2 K(\mathbf{s}_i, \mathbf{s}_j) + \sigma_\varepsilon^2 \mathbf{1}(\mathbf{s}_i = \mathbf{s}_j).$$

In the discussions that follow β and σ_y^2 are always assumed unknown, but some of the covariance parameters θ_1, θ_2 and σ_ε^2 would be assumed known. We will have, possibly after a reparametrization of the covariance parameters, that

$$E\{\mathbf{Y}\} = X\beta \quad \text{and} \quad \mathrm{var}\{\mathbf{Y}\} = \sigma^2 \Sigma_\theta,$$

where $(X)_{ij} = f_j(\mathbf{s}_i)$ is a known $n \times p$ design matrix of rank p and Σ_θ is an $n \times n$ positive definite matrix for any $\theta \in \Theta \subset \mathbb{R}^q$, $q \in \mathbb{N}$. What σ^2, θ and Σ_θ are in any particular case will depend on the parametrization that is used. We review some univariate ($q = 1$) cases in Sections 13.3.3–13.3.6, and consider some multivariate ($q > 1$) cases in Section 13.3.7. Then the likelihood function of the model parameters $\eta = (\beta, \sigma^2, \theta) \in \Omega = \mathbb{R}^p \times (0, \infty) \times \Theta$ based on the observed data $\mathbf{y}$ is given by

$$L(\eta; \mathbf{y}) \propto (\sigma^2)^{-\frac{n}{2}} |\Sigma_\theta|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2\sigma^2} (\mathbf{y} - X\beta)' \Sigma_\theta^{-1} (\mathbf{y} - X\beta) \right\}. \tag{13.3.2}$$

13.3.2 Integrated Likelihoods

Similarly to what is often done for Bayesian analysis of ordinary linear models, a sensible class of (improper) prior distributions for η is given by the family

$$\pi(\eta) \propto \frac{\pi(\theta)}{(\sigma^2)^a}, \qquad \eta \in \Omega, \tag{13.3.3}$$

where $a \in \mathbb{R}$ is a hyper-parameter and $\pi(\theta)$ is the 'marginal' prior of θ with support Θ. Priors of this form, with $a = 1$, were first proposed by Kitanidis (1986) and Handcock and Stein (1993), but with little guidance on how to select $\pi(\theta)$. The relevance of this class of priors will become apparent since several Jeffreys and

reference priors belong to this class. An obvious choice, used by may authors, is to set $a = 1$ and $\pi(\theta) = \mathbf{1}_\Theta(\theta)$, but the use of this prior yields improper posteriors in some cases.

From Bayes theorem follows that the posterior distribution of η is proper if and only if $0 < \int_\Omega L(\eta;\mathbf{y})\pi(\eta)d\eta < \infty$. A standard calculation with the above likelihood and prior shows that

$$\int_{\mathbb{R}^p \times (0,\infty)} L(\eta;\mathbf{y})\pi(\eta)d\beta d\sigma^2 = L^I(\theta;\mathbf{y})\pi(\theta),$$

with

$$L^I(\theta;\mathbf{y}) \propto |\Sigma_\theta^{-1}|^{\frac{1}{2}}|X'\Sigma_\theta^{-1}X|^{-\frac{1}{2}}(S^2_\theta)^{-(\frac{n-p}{2}+a-1)},$$

where $S^2_\theta = (\mathbf{y}-X\hat{\beta}_\theta)'\Sigma_\theta^{-1}(\mathbf{y}-X\hat{\beta}_\theta)$, $\hat{\beta}_\theta = (X'\Sigma_\theta^{-1}X)^{-1}X'\Sigma_\theta^{-1}\mathbf{y}$, and $L^I(\theta;\mathbf{y})$ is called the *integrated likelihood* of θ. Then, the posterior distribution of η is proper if and only if

$$0 < \int_\Theta L^I(\theta;\mathbf{y})\pi(\theta)d\theta < \infty,$$

so to determine propriety of posterior distributions based on priors (13.3.3) is necessary to determine the behavior of the integrated likelihood $L^I(\theta;\mathbf{y})$ and 'marginal' prior $\pi(\theta)$ on Θ.

13.3.3 Reference Priors

The reference prior algorithm is (arguably) the most successful general method to derive default priors. The basic idea is to find a prior that conveys minimal information about the quantity of interest (in an entropy distance sense) relative to that provided by the data. Use of such prior would then make the data dominant for posterior inference. The computation of reference priors depends, in general, on the entertained model and a classification of the parameters according to their inferential importance; see Bernardo (1979) and Berger and Bernardo (1992a) for the motivation and initial developments of the algorithm, and Bernardo (2005) and Berger, Bernardo, and Sun (2009) for recent theoretical developments. A reference prior for the model with general isotropic correlation function having all three parameters θ_1, θ_2 and σ^2_ε unknown can be computed (see Section 13.3.7), but its behavior and properties of the resulting posterior are so far unknown.

In this section reference priors for the parameters of model (13.3.2) are reviewed in some cases in which θ is univariate, which would be denoted by ϑ. It will be assumed that (σ^2, ϑ) is the parameter of interest and β is the nuisance parameter. For that we use the two-step reference prior algorithm that uses exact marginalization, as described in Berger, De Oliveira, and Sansó (2001). The first step is to factor the joint prior distribution as $\pi^R(\eta) = \pi^R(\beta \mid \sigma^2, \vartheta)\pi^R(\sigma^2, \vartheta)$ and use $\pi^R(\beta \mid \sigma^2, \vartheta) \propto 1$, since this is the conditional Jeffreys-rule (or reference) prior for

β for model (13.3.2) when (σ^2, ϑ) is known. Second, $\pi^R(\sigma^2, \vartheta)$ is computed using the Jeffreys-rule algorithm based on the 'marginal model' provided by the integrated likelihood of (σ^2, ϑ)

$$L^I(\sigma^2, \vartheta; \mathbf{z}) = \int_{\mathbb{R}^p} L(\beta, \sigma^2, \vartheta; \mathbf{z}) \pi^R(\beta \mid \sigma^2, \vartheta) d\beta$$
$$\propto (\sigma^2)^{-\frac{n-p}{2}} |\Sigma_\vartheta|^{-\frac{1}{2}} |X' \Sigma_\vartheta^{-1} X|^{-\frac{1}{2}} \exp\left\{ -\frac{S^2_\vartheta}{2\sigma^2} \right\}. \quad (13.3.4)$$

Theorem 13.1 (Berger, De Oliveira, and Sansó, 2001, Theorem 2). *For the model with sampling distribution (13.3.2) and $\theta = \vartheta$ univariate the reference prior distribution, $\pi^R(\beta, \sigma^2, \vartheta)$, is of the form (13.3.3) with*

$$a = 1 \quad \textit{and} \quad \pi^R(\vartheta) \propto \left\{ \text{tr}[W_\vartheta^2] - \frac{1}{n-p} (\text{tr}[W_\vartheta])^2 \right\}^{\frac{1}{2}}, \quad (13.3.5)$$

where $W_\vartheta = (\frac{\partial}{\partial \theta} \Sigma_\vartheta) Q_\vartheta$ and $Q_\vartheta = \Sigma_\vartheta^{-1} - \Sigma_\vartheta^{-1} X (X' \Sigma_\vartheta^{-1} X)^{-1} X' \Sigma_\vartheta^{-1}$; $\frac{\partial}{\partial \vartheta} \Sigma_\vartheta$ denotes the matrix obtained by differentiating Σ_ϑ element-wise.

Remark 13.1. A bit of calculation shows that the right hand side of (13.3.5) is proportional to the sample standard deviation of the *positive* eigenvalues of W_ϑ. In some special cases (such as Case B below) it is possible to find explicit expressions for these eigenvalues, which would provide great analytical and computational simplifications. Unfortunately such explicit expressions are not in general available.

Remark 13.2. Prior (13.3.5) is quite general since the nature and interpretation of ϑ is immaterial for its computation. On the other hand, determining the behavior and properties of the resulting posterior would very much depend on what the actual parameter ϑ represents, so these tasks need to be undertaken on a case-by-case basis.

We now review some special cases that been previously studied in some detail.

Case A (Berger, De Oliveira, and Sansó, 2001). Consider the covariance parametrization where $\sigma^2 = \sigma_y^2$ and $\vartheta = \theta_1$ are unknown, and θ_2 and $\sigma_\varepsilon^2 = 0$ are known (the data contain no measurement error), so $(\Sigma_\vartheta)_{ij} = K(\mathbf{s}_i, \mathbf{s}_j)$. In this case, $\Sigma_\vartheta \to I_n$ as $\vartheta \to 0$ and $\Sigma_\vartheta \to \mathbf{1}\mathbf{1}'$ (a singular matrix) as $\vartheta \to \infty$, where $\mathbf{1}$ is the vector of ones. As a result of the latter the behavior of the reference prior and integrated likelihood depend on whether or not $\mathbf{1}$ is a column of X.[1] We assume throughout this article, unless stated otherwise, that this condition holds as this is the case in most practical problems.

In this case (as well as in Case B considered later) the integrated likelihood of ϑ may have quite unusual behaviors: depending on the value of the hyper-parameter a and whether or not $\mathbf{1}$ is a column of X, $L^I(\vartheta; \mathbf{y})$ may converge to zero, a positive

[1] More generally, it depends on whether or not $\mathbf{1}$ belongs to $\mathscr{C}(X)$, the column space of X, but little is gained by considering this extra generality.

constant or infinity, as ϑ goes to infinity. In particular, when $a = 1$ and $\mathbf{1}$ is a column of X, $L^I(\vartheta;\mathbf{y})$ is bounded between two positive constants regardless of the sample size. As a consequence, the use of some naive priors results in improper posteriors.

Proposition 13.1 (Berger, De Oliveira, and Sansó, 2001, Theorem 1). *Consider model (13.3.2) where Case A holds. Under some assumptions met by most spatial correlation functions, any of the following naive priors yield an improper posterior for $(\beta,\sigma^2,\vartheta)$: (i) $\pi(\beta,\sigma^2,\vartheta) \propto 1$; (ii) prior (13.3.3) with $a = 1$ and $\pi(\vartheta)$ not integrable either at 0 or ∞ (a special case being $\pi(\vartheta) = 1$); and (iii) prior (13.3.3) with $a \in \mathbb{R}$ and $\pi(\vartheta)$ not integrable at 0 (a special case being $\pi(\vartheta) = \vartheta^{-1}$).*

Proposition 13.2 (Berger, De Oliveira, and Sansó, 2001, Theorem 4). *Consider model (13.3.2) where Case A holds. Under some assumptions met by most spatial correlation functions we have: (i) the reference prior (13.3.5) yields a proper posterior for $(\beta,\sigma^2,\vartheta)$; and (ii) $\pi^R(\vartheta)$ in (13.3.5) is integrable on $(0,\infty)$.*

Remark 13.3. In spite of the myriad of available correlation functions available in the literature, a unified analysis is possible by using Taylor series expansion of the correlation function. Using such an expansion in Case A shows that, under mild conditions, Σ_ϑ can be written as

$$\Sigma_\vartheta = \mathbf{1}\mathbf{1}' + v(\vartheta)D + w(\vartheta)D^* + R(\vartheta),$$

where $v(\vartheta) > 0$, $w(\vartheta)$ and $R(\vartheta)$ are differentiable, D symmetric and nonsingular and D^* not depending on ϑ, and as $\vartheta \to \infty$, $v(\vartheta) = o(1)$, $w(\vartheta) = o(v(\vartheta))$, $w'(\vartheta) = o(v'(\vartheta))$, $||R(\vartheta)||_\infty = o(w(\vartheta))$, and $||R'(\vartheta)||_\infty = o(w'(\vartheta))$. For instance, for the Matérn family in (13.3.1) with $\theta_2 > 1$ and non-integer, $v(\vartheta) = 1/\vartheta^2$ and $w(\vartheta) = 1/\vartheta^{2\theta_2}$. As it turns out, the behavior at infinity of $L^I(\vartheta;\mathbf{y})$ and $\pi^R(\vartheta)$ are determined by the functions $v(\vartheta)$ and $w(\vartheta)$.

Remark 13.4. From results in Berger, De Oliveira, and Sansó (2001) follows that the reference marginal posterior of ϑ is heavy-tailed, but how much so depends on the correlation function that is used. For instance, for members of the Matérn family (13.3.1) with $\theta_2 > 1$ and non-integer, $\pi^R(\vartheta \mid \mathbf{y})$ has a first moment if and only if $\theta_2 > 3/2$. Existence of other moments will similarly depend on the value of θ_2.

Case B (De Oliveira, 2007). Consider the covariance parametrization where $\sigma^2 = \sigma^2_\varepsilon$ and $\vartheta = \sigma^2_y/\sigma^2_\varepsilon$ (the so-called signal-to-noise ratio) are unknown, and θ_1 and θ_2 are known, so $\Sigma_\vartheta = I_n + \vartheta H$, with $(H)_{ij} = K(\mathbf{s}_i,\mathbf{s}_j)$ known. As in Case A, the integrated likelihood of ϑ has the unusual behavior mentioned above, and hence any of the priors listed in Proposition 13.1 also yield an improper posterior for $(\beta,\sigma^2,\vartheta)$. But unlike Case A, the behavior of the reference prior and resulting posterior do not depend on whether or not $\mathbf{1}$ is a column of X.

For the following results, let L be a full-rank $n \times (n-p)$ matrix satisfying $L'X = 0$ and $L'L = I_{n-p}$, and for a square matrix A, let $\lambda_1(A) \geq \lambda_2(A) \geq \ldots \geq \lambda_n(A)$ be the ordered eigenvalues of A.

Proposition 13.3 (De Oliveira, 2007, Theorem 1). *Consider model (13.3.2) where Case B holds. Then the reference prior of $(\beta,\sigma^2,\vartheta)$ is of the form (13.3.3)*

$$a = 1 \text{ and } \pi^R(\vartheta) \propto \left(\sum_{j=1}^{n-p} \left\{ \frac{\lambda_j(L'HL)}{1+\vartheta\lambda_j(L'HL)} \right\}^2 - \frac{1}{n-p} \left[\sum_{j=1}^{n-p} \left\{ \frac{\lambda_j(L'HL)}{1+\vartheta\lambda_j(L'HL)} \right\} \right]^2 \right)^{\frac{1}{2}}. \tag{13.3.6}$$

The following result provides the main properties of the above reference prior and those of its corresponding reference posterior.

Proposition 13.4 (De Oliveira, 2007, Proposition 2, Corollary 1). *Suppose that* $\{\lambda_j(L'HL)\}_{j=1}^{n-p}$ *are not all equal. Then, (i) the 'marginal' reference prior of* ϑ *given in (13.3.6) is a continuous function on* $[0,\infty)$*; (ii) The reference prior yields a proper posterior for* $(\beta,\sigma^2,\vartheta)$*; (iii)* $\pi^R(\vartheta)$ *is strictly decreasing on* $[0,\infty)$*; (iv)* $\pi^R(\vartheta) = O\left(\vartheta^{-2}\right)$ *as* $\vartheta \to \infty$*, so it is integrable on* $(0,\infty)$*, but* $\pi^R(\vartheta \mid \mathbf{y})$ *does not have moments of any order* $k \geq 1$*; and (v) the marginal reference posterior* $\pi^R(\sigma^2 \mid \mathbf{y})$ *has a finite moment of order* $k \geq 1$ *if* $n \geq p + 2k + 1$.

13.3.4 Jeffreys Priors

Probably the most commonly used default prior is the Jeffreys-rule prior, which is given by $\pi(\eta) \propto (\det[I(\eta)])^{\frac{1}{2}}$, where $I(\eta)$ is the Fisher information matrix with (i,j) entry

$$[I(\eta)]_{ij} = E\left\{ \left(\frac{\partial}{\partial\eta_i} \log(L(\eta;\mathbf{Y})) \right) \left(\frac{\partial}{\partial\eta_j} \log(L(\eta;\mathbf{Y})) \right) \;\middle|\; \eta \right\}.$$

As for reference priors, Jeffreys-rule priors for the model with general isotropic correlation function having all three parameters θ_1, θ_2 and σ_ε^2 unknown can be computed, but its behavior and properties are unknown. In this section we also consider model (13.3.2) in some cases in which $\theta = \vartheta$ is univariate.

The Jeffreys-rule prior has several attractive features, such as invariance to one-to-one reparametrizations and restrictions of the parameter space, but it also has some not so attractive features. One of these is the poor frequentist properties that have been noticed in multi-parameter models. This section derives two versions of Jeffreys prior, the Jeffreys-rule prior and the independence Jeffreys prior, where the latter (intended to ameliorate the aforementioned unattractive feature) is obtained by assuming that β and (σ^2,ϑ) are 'independent' a priori and computing each marginal prior using Jeffreys-rule when the other parameter is assumed known.

Theorem 13.2 (Berger, De Oliveira, and Sansó, 2001, Theorem 5). *For the model with sampling distribution (13.3.2) and* $\theta = \vartheta$ *univariate the independence Jeffreys prior and the Jeffreys-rule prior, to be denoted by* $\pi^{J1}(\beta,\sigma^2,\vartheta)$ *and* $\pi^{J2}(\beta,\sigma^2,\vartheta)$*, respectively, are of the form (13.3.3) with, respectively,*

$$a = 1 \quad \text{and} \quad \pi^{J1}(\vartheta) \propto \left\{ \text{tr}[U_\vartheta^2] - \frac{1}{n}(\text{tr}[U_\vartheta])^2 \right\}^{\frac{1}{2}}, \tag{13.3.7}$$

and

$$a = 1 + \frac{p}{2} \quad \text{and} \quad \pi^{J2}(\vartheta) \propto |X'\Sigma_\vartheta^{-1}X|^{\frac{1}{2}}\pi^{J1}(\vartheta), \tag{13.3.8}$$

where $U_\vartheta = (\frac{\partial}{\partial\vartheta}\Sigma_\vartheta)\Sigma_\vartheta^{-1}$.

Case A. As for the reference prior, the behavior of Jeffreys priors and posteriors in this model depend on whether or not $\mathbf{1}$ is a column of X, and we continue assuming this condition holds unless stated otherwise.

Proposition 13.5 (Berger, De Oliveira, and Sansó, 2001, Theorems 6 and 7). *Consider model (13.3.2) where Case A holds. Under some assumptions met by most spatial correlation functions we have (i) the independence Jeffreys prior yields an improper posterior for* $(\beta, \sigma^2, \vartheta)$ *(although it yields a proper posterior when* $\mathbf{1}$ *is not a column of X); (ii)* $\pi^{J1}(\vartheta)$ *in (13.3.7) is not integrable on* $(0,\infty)$*; (iii) The Jeffreys-rule prior yields a proper posterior for* $(\beta, \sigma^2, \vartheta)$ *(regardless of whether or not* $\mathbf{1}$ *is a column of X); and (iv)* $\pi^{J2}(\vartheta)$ *in (13.3.8) is integrable on* $(0,\infty)$.

This case provides one of the few known examples in which posterior impropriety results when using the independence Jeffreys prior. Although the Jeffreys-rule prior does result in a proper posterior, its use is not advisable due to the poor frequentist properties of Bayesian inferences based on this prior, and the lack of validity of the use of Bayes factors for correlation function selection; see Section 13.3.6 for further comments.

Case B.

Proposition 13.6 (De Oliveira, 2007, Theorem 2). *Consider model (13.3.2) where Case B holds. Then the independence Jeffreys prior and Jeffreys-rule prior of* $(\beta, \sigma^2, \vartheta)$ *are of the form (13.3.3) with, respectively,*

$$a = 1 \text{ and } \pi^{J1}(\vartheta) \propto \Big(\sum_{i=1}^{n} \Big\{ \frac{\lambda_i(H)}{1+\vartheta\lambda_i(H)} \Big\}^2 - \frac{1}{n}\Big[\sum_{i=1}^{n} \Big\{ \frac{\lambda_i(H)}{1+\vartheta\lambda_i(H)} \Big\} \Big]^2 \Big)^{\frac{1}{2}},$$

and

$$a = 1 + \frac{p}{2} \text{ and } \pi^{J2}(\vartheta) \propto \Big[\frac{\prod_{j=1}^{n-p}\{1+\vartheta\lambda_j(L'HL)\}}{\prod_{i=1}^{n}\{1+\vartheta\lambda_i(H)\}} \Big]^{\frac{1}{2}} \pi^{J1}(\vartheta).$$

Proposition 13.7 (De Oliveira, 2007, Proposition 3). *Suppose that* $\{\lambda_i(H)\}_{i=1}^n$ *are not all equal. Then, the marginal independence Jeffreys and Jeffreys-rule priors of* ϑ *given above are continuous functions on* $[0,\infty)$ *satisfying: (i)* $\pi^{J1}(\vartheta)$ *and* $\pi^{J2}(\vartheta)$ *are strictly decreasing on* $[0,\infty)$*; and (ii)* $\pi^{J1}(\vartheta) = O\big(\vartheta^{-2}\big)$ *and* $\pi^{J2}(\vartheta) = O\big(\vartheta^{-(2+\frac{p}{2})}\big)$ *as* $\vartheta \to \infty$.

Corollary 13.1 (De Oliveira, 2007, Corollary 2). *(i) The marginal independence Jeffreys prior* $\pi^{J1}(\vartheta)$ *and joint independence Jeffreys posterior* $\pi^{J1}(\eta \mid \mathbf{y})$ *are both proper. (ii) The marginal independence Jeffreys posterior* $\pi^{J1}(\vartheta \mid \mathbf{y})$ *does not have moments of any order* $k \geq 1$. *(iii) The marginal independence Jeffreys posterior* $\pi^{J1}(\sigma^2 \mid \mathbf{y})$ *has a finite moment of order* $k \geq 1$ *if* $n \geq p + 2k + 1$.

Corollary 13.2 (De Oliveira, 2007, Corollary 3). *(i) The marginal Jeffreys-rule prior* $\pi^{J2}(\vartheta)$ *and joint Jeffreys-rule posterior* $\pi^{J2}(\eta \mid \mathbf{y})$ *are both proper. (ii) The marginal Jeffreys-rule posterior* $\pi^{J2}(\vartheta \mid \mathbf{y})$ *has a finite moment of order* k *if* $k < 1 + \frac{p}{2}$. *(iii) The marginal Jeffreys-rule posterior* $\pi^{J2}(\sigma^2 \mid \mathbf{y})$ *has a finite moment of order* k *if* $n \geq p + 2k + 1$.

Remark 13.5. It is worth noting the quite different behaviors of the independence Jeffreys prior in the previous two cases: it yields an improper posterior for the model parameters in Case A, while it yields a proper posterior in Case B.

13.3.5 Other Spatial Models

The methods described above to compute default priors can also be applied to other Gaussian spatial models called Gaussian Markov random fields. In these the region $\mathscr{D}$ is partitioned into subregions (or sites as they are also called) indexed by integers $1, 2, \ldots, n$, linked together according to a neighborhood system, $\{N_i : i = 1, \ldots, n\}$, where N_i denotes the collection of subregions that are neighbors of subregion i. An emblematic example commonly used in applications is the neighborhood system defined in terms of geographic adjacency

$$N_i = \{j : \text{subregions } i \text{ and } j \text{ share a boundary}\}, \quad i = 1, \ldots, n,$$

but other examples are also possible that include neighborhood systems defined based on distance from the centroids of subregions or similarity of an auxiliary variable; see Cressie (1993) and Rue and Held (2005) for treatments of these models.

For each subregion it is observed the variable of interest, Y_i, which is often (but not always) an aggregate or average over the subregion i, and a set of p explanatory variables, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$. For this class of models spatial association is specified conditionally through the set of full conditional distributions, which amounts to specifying the precision matrix of the data (rather than the covariance matrix). Here we describe the so-called conditional autoregressive (CAR) model whose full conditional distributions satisfy a form of autoregression given by

$$(Y_i \mid \mathbf{Y}_{(i)}) \sim \mathrm{N}\Big(\mathbf{x}_i'\beta + \sum_{j=1}^{n} c_{ij}(\theta)(Y_j - \mathbf{x}_j'\beta), \sigma_i^2\Big), \quad i = 1, \ldots, n, \tag{13.3.9}$$

where $\mathbf{Y}_{(i)} = \{Y_j : j \neq i\}$, $\beta = (\beta_1, \ldots, \beta_p)' \in \mathbb{R}^p$ are unknown regression parameters, and $\sigma_i^2 > 0$ and $c_{ij}(\theta) \geq 0$ are covariance parameters, with $c_{ii}(\theta) = 0$ for all i. A commonly used model results by assuming $\sigma_1^2 = \ldots = \sigma_n^2 = \sigma^2$ and $c_{ij}(\theta) = \vartheta w_{ij}$, where ϑ is an unknown scalar and $W = (w_{ij})$ is a known "weight" ("neighborhood") matrix that is nonnegative, symmetric and satisfies that $w_{ij} > 0$ if and only if sites i and j are neighbors. For the set of full conditional distributions (13.3.9) to determine a well defined joint distribution for $\mathbf{Y}$, the spatial parameter ϑ must belong to $(\lambda_n^{-1}(W), \lambda_1^{-1}(W))$ (where $\lambda_n(W) < 0 < \lambda_1(W)$). As before, let

$\eta = (\beta, \sigma^2, \vartheta) \in \Omega = \mathbb{R}^p \times (0,\infty) \times (\lambda_n^{-1}(W), \lambda_1^{-1}(W))$ denote the model parameters. In this case $\mathbf{Y}$ follows model (13.3.2) where

$$X = (\mathbf{x}_1 \cdots \mathbf{x}_n)' \quad \text{and} \quad \Sigma_\vartheta = (I_n - \vartheta W)^{-1},$$

so Theorems 2 and 5 in Berger, De Oliveira, and Sansó (2001) also hold for this model. Other (slightly) more general models can be reduced to the above one by an appropriate scaling of the data; see De Oliveira (2010) for details. In what follows we state the Jeffreys priors and their properties as well as properties of the resulting posteriors.

Theorem 13.3 (De Oliveira, 2010, Theorem 1). *Consider the CAR model determined by (13.3.9). Then the independence Jeffreys prior and the Jeffreys-rule prior of η, to be denoted by $\pi^{J1}(\eta)$ and $\pi^{J2}(\eta)$, are of the form (13.3.3) with, respectively,*

$$a = 1 \quad \text{and} \quad \pi^{J1}(\vartheta) \propto \left\{ \sum_{i=1}^n \left(\frac{\lambda_i(W)}{1 - \vartheta\lambda_i(W)} \right)^2 - \frac{1}{n} \left[\sum_{i=1}^n \frac{\lambda_i(W)}{1 - \vartheta\lambda_i(W)} \right]^2 \right\}^{\frac{1}{2}},$$

and

$$a = 1 + \frac{p}{2} \quad \text{and} \quad \pi^{J2}(\vartheta) \propto \left(\prod_{j=1}^p (1 - \vartheta\lambda_j(X_o' W X_o)) \right)^{\frac{1}{2}} \pi^{J1}(\vartheta),$$

where $X_o = XVT^{\frac{1}{2}}$, with V orthogonal and T diagonal being the components of the spectral decomposition of $X'X$ $(= VTV')$.

In what follows let $\mathscr{C}(X)$ denote the column space of X, and $\mathbf{u}_1$ and $\mathbf{u}_n$ be the normalized eigenvectors of W corresponding to, respectively, $\lambda_1(W)$ and $\lambda_n(W)$.

Proposition 13.8 (De Oliveira, 2010, Corollaries 1 and 2). *Consider the CAR model determined by (13.3.9). (i) The marginal independence Jeffreys prior $\pi^{J1}(\vartheta)$ is unbounded and not integrable. (ii) The joint independence Jeffreys posterior $\pi^{J1}(\eta \mid \mathbf{y})$ is proper when neither $\mathbf{u}_1$ nor $\mathbf{u}_n$ are in $\mathscr{C}(X)$, while it is improper when either $\mathbf{u}_1$ or $\mathbf{u}_n$ are in $\mathscr{C}(X)$. (iii) The marginal Jeffreys-rule prior $\pi^{J2}(\vartheta)$ is unbounded. Also, it is integrable when both $\mathbf{u}_1$ and $\mathbf{u}_n$ are in $\mathscr{C}(X)$, while it is not integrable when either $\mathbf{u}_1$ or $\mathbf{u}_n$ is not in $\mathscr{C}(X)$. (iv) The joint Jeffreys-rule posterior $\pi^{J2}(\eta \mid \mathbf{y})$ is always proper.*

Remark 13.6. Although propriety of the joint independence Jeffreys posterior is not always guaranteed, it is unlikely to encounter situations where either $\mathbf{u}_1$ or $\mathbf{u}_n$ belong to $\mathscr{C}(X)$, so $\pi^{J1}(\eta \mid \mathbf{y})$ would likely be proper in practice.

As mentioned before, a reference prior distribution for η can also be computed, which is given by (13.3.5), but a more explicit expression and knowledge about properties of the resulting posterior are so far lacking. On the other hand Ferreira and De Oliveira (2007) derived, for a different Gaussian Markov random field model with constant mean, explicit expressions for Jeffreys and reference priors as well as the properties of the corresponding posterior distributions. Similar analysis are also

possible for other spatial and non-spatial Gaussian models; see van der Linde (2000) and De Oliveira and Song (2008) for examples.

13.3.6 Further Properties

In this section some additional properties of inferences based on reference and Jeffreys priors are reviewed.

Frequentist properties. Frequentist properties are often proposed as a way to evaluate and compare Bayesian inferences based on default priors. The most common of these properties is the frequentist coverage of equal-tailed $100(1-\alpha)\%$ Bayesian credible intervals of the parameters of interest, in this case (σ^2, ϑ). Some limited simulation experiments reported in Berger, De Oliveira, and Sansó (2001), Paulo (2005) and De Oliveira (2007) suggest that frequentist properties of Bayesian inferences based on the reference prior are moderately good, and similar to Bayesian inferences based on the independence Jeffreys prior, when the latter yields a proper posterior. On the other hand, frequentist properties of Bayesian inferences based on the Jeffreys-rule prior tend to be inferior than the above two, and inadequate in situations where the mean function is not constant or the spatial association is strong. These frequentist properties were also found to hold for other classes of spatial models, where Bayesian inferences based on these default priors were superior than those based on maximum likelihood; see De Oliveira (2010) and De Oliveira and Song (2008) for examples.

Correlation function selection. A myriad of spatial correlation function have been proposed in the literature. In practice little or no subject-based information is available to guide this choice, and the selection of correlation function is often arbitrary, so methods for model selection in this context are important.

One of the good properties of Bayesian inference based on the reference prior is that Bayes factors can be used for correlation function selection between models that share a common mean function. Typically, model selection based on Bayes factors is precluded when using improper priors, but an important exception was studied by Berger, Pericchi, and Varshavsky (1998). When the models that are compared have the same mean structure, the model considered here fits this special situation when (a) priors of all models are of the form (13.3.3) with $a = 1$; and (b) the 'marginals' $\pi(\vartheta)$ of all models are proper.

The reference prior in Case A satisfies (a) and (b), as long as the mean function includes a constant term (i.e. when $\mathbf{1}$ is a column of X), while the reference prior in Case B always satisfies (a) and (b). In addition, the independence Jeffreys prior in Case B also satisfies (a) and (b). In all these cases Bayes factors computed from these priors can be used for correlation function selection. On the other hand, neither the independence Jeffreys prior in Case A nor the Jeffreys-rule prior in Cases A and B enjoy this desirable property, so Bayes factors computed from these priors can not be used for correlation function selection.

Sensitivity to design. The reference and Jeffreys priors described before depend on several features of the selected design, such as sampling size, sampling locations and regression matrix. Nevertheless, numerical explorations in De Oliveira (2007, 2010) suggest that sensitivity of the priors to these features is in general mild, except in one instance: the Jeffreys-rule prior displays substantial sensitivity to changes in the regression matrix X.

13.3.7 Multi-Parameter Cases

Working with models where $\theta = (\theta_1, \ldots, \theta_1)$ is considered univariate requires fixing some of its components, which may be a serious limitation when modeling some datasets. Extending the framework considered in Berger, De Oliveira, and Sansó (2001), Paulo (2005) considered the general case of θ multivariate. He derived expressions for the reference prior of η, when (σ^2, θ) is the parameter of interest and β is the nuisance parameter, the Jeffreys-rule prior and the independence Jeffreys prior, when β and (σ^2, θ) are assumed 'independent' a priori.

Proposition 13.9 (Paulo, 2005, Propositions 2.1 and 2.2). *Consider the model with sampling distribution (13.3.2). (i) The reference prior of η is of the form (13.3.3) with $a = 1$ and $\pi^R(\theta) \propto (\det[I^R(\theta)])^{\frac{1}{2}}$, where $I^R(\theta)$ is the $(q+1) \times (q+1)$ matrix*

$$I^R(\theta) = \begin{pmatrix} n-p & \operatorname{tr}[W_\theta^{(1)}] & \operatorname{tr}[W_\theta^{(2)}] & \cdots & \operatorname{tr}[W_\theta^{(q)}] \\ & \operatorname{tr}[(W_\theta^{(1)})^2] & \operatorname{tr}[W_\theta^{(1)} W_\theta^{(2)}] & \cdots & \operatorname{tr}[W_\theta^{(1)} W_\theta^{(q)}] \\ & & \ddots & \cdots & \vdots \\ & \text{symmetric} & & & \operatorname{tr}[(W_\theta^{(q)})^2] \end{pmatrix}, \tag{13.3.10}$$

with $W_\theta^{(j)} = (\frac{\partial}{\partial \theta_j} \Sigma_\theta) Q_\theta$ and $Q_\theta = \Sigma_\theta^{-1} - \Sigma_\theta^{-1} X (X' \Sigma_\theta^{-1} X)^{-1} X' \Sigma_\theta^{-1}$, $j = 1, \ldots, q$. (ii) The independence Jeffreys and Jeffreys-rule priors are of the form (13.3.3) with, respectively

$$a = 1 \quad \text{and} \quad \pi^{J1}(\theta) \propto (\det[I^J(\theta)])^{\frac{1}{2}},$$

and

$$a = 1 + \frac{p}{2} \quad \text{and} \quad \pi^{J2}(\theta) \propto |X' \Sigma_\theta^{-1} X|^{\frac{1}{2}} \pi^{J1}(\theta),$$

where $I^J(\theta)$ is similar to the matrix in (13.3.10), but with the 1-1 entry equal to n and the other entries obtained by replacing Q_θ with Σ_θ^{-1}.

As in the univariate case, the above formulas are quite general, but determining the behavior and properties of $L^I(\theta; \mathbf{y})$ and $\pi(\theta)$ in this general setting does not appear possible at the moment. Analytical progress is possible though in some cases that we now review.

13.3.7.1 The Separable Correlation Case

An important class of multi-parameter correlation functions that is often used in the design and analysis of computer experiments is the so-called *separable*. Before describing it, we introduce some needed notation. Suppose that $l \geq 2$ and that spatial locations are written as $\mathbf{s} = (\mathbf{s}_1, \ldots, \mathbf{s}_r)$, with $r \leq l$, $\mathbf{s}_k \subset \mathbb{R}^{l_k}$ and $\sum_{k=1}^{r} l_k = l$. Also, let $S = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ be the sampling design. Paulo (2005) studied the properties of the above default priors and their corresponding posterior distributions under the following assumptions:

Assumption A1. Separability of the correlation function:

$$K(\mathbf{s}, \mathbf{u}) = \prod_{k=1}^{r} K_k(\mathbf{s}_k, \mathbf{u}_k; \theta_k),$$

where $K_k(\mathbf{s}_k, \mathbf{u}_k; \theta_k)$ is an isotropic correlation function in $\mathbb{R}^{l_k}$ and $\theta_k > 0$ is a range parameter, $k = 1, \ldots, r$.

Assumption A2. Cartesian product of the sampling design:

$$S = S_1 \times S_2 \times \cdots \times S_r,$$

where $S_k = \{\mathbf{s}_{1,k}, \ldots, \mathbf{s}_{n_k,k}\} \subset \mathbb{R}^{l_k}$, $\#S_k = n_k$ and $\#S = \prod_{k=1}^{r} n_k \ \ (= n)$.

Assumption A3. $p = 1$ (so there is only one regression parameter) and

$$X = X_1 \otimes X_2 \otimes \cdots \otimes X_r,$$

where X_k has dimension $n_k \times 1$, $k = 1, \ldots, r$.

The above assumptions allow both the regression matrix and the covariance matrix of the data to be written as Kronecker products of simpler matrices, which in turn induces a form of partial separability[2] of $L^J(\theta; \mathbf{y})$. In addition, $\pi^R(\theta)$, $\pi^{J1}(\theta)$ and $\pi^{J2}(\theta)$ given above can be bounded above by a product of r univariate functions, with the kth function depending only on θ_k. The behavior at infinity of these r univariate functions is determined under some added assumptions (A4–A8 in Paulo (2005)) which together with the first three assumptions determine the central result for this class of models.

Theorem 13.4 (Paulo, 2005, Theorem 3.6). *Consider the model with sampling distribution (13.3.2) and prior (13.3.3) with $\pi(\theta)$ equal to $\pi^R(\theta)$, $\pi^{J1}(\theta)$ or $\pi^{J2}(\theta)$ given above. Then under assumptions $\mathscr{A}1$–$\mathscr{A}8$ the posterior distribution of θ is proper provided $a > 1/2$. It then follows that the reference and both Jeffreys priors always yield proper posterior distributions.*

Remark 13.7. The above result indicates sharp differences in posterior propriety behavior between this multivariate case and the univariate cases discussed in Sections

[2] This means that this function can be written as a product of r univariate functions, with the kth factor depending only on θ_k, times a multivariate function (say $f(\theta)$).

13.3.3 and 13.3.4, which are mainly due to the differences in the behavior of the integrated likelihood at infinity. Consider the practically common situation of a random field with constant mean and using prior (13.3.3) with $a = 1$. For the univariate case, $L^I(\theta;\mathbf{y})$ is bounded away from zero everywhere, regardless of the sample size, so $\pi(\theta) = 1$ yields an improper posterior. On the other hand, for the above multivariate model it can be shown from the results in Paulo (2005) that $L^I(\theta;\mathbf{y})$ goes to zero as any $\theta_k \to \infty$, $k = 1,\ldots,r$. Also, depending on l and for n_k large enough, $\pi(\theta) = 1$ yields an proper posterior distribution. Paulo (2005) claimed that the univariate cases discussed in Sections 13.3.3 and 13.3.4 are somewhat exceptional, and that in general "good" posterior behavior is to be expected for multivariate cases. But this is not quite so, as will be seen in the next section.

13.3.7.2 An Isotropic Multivariate Case

Recently Ren, Sun, and He (2009) generalized the results in Berger, De Oliveira, and Sansó (2001) and De Oliveira (2007) by considering models with isotropic correlation functions having both the range and nugget parameters unknown (in a sense combining Cases A and B considered before). Using a notation slightly different to that of Section 13.3.1, consider the model where $\sigma^2 = \sigma_y^2$, θ_1 and $\theta_2 = \sigma_\varepsilon^2/\sigma_y^2$ (the noise-to-signal ratio) are unknown, and the smoothness/roughness parameter is known, so $\Sigma_\theta = H_{\theta_1} + \theta_2 I_n$, with $(H_{\theta_1})_{ij} = K(\mathbf{s}_i,\mathbf{s}_j)$. For this case Ren, Sun, and He (2009) computed the Jeffreys-rule prior, several versions of independence Jeffreys priors and two reference priors obtained by exact marginalization, the second obtained by exact marginalization over both β and σ^2. In addition, they also computed several versions of reference priors obtained by asymptotic marginalization and found each of these agree with some version of independence Jeffreys prior (six different default priors in total). As before, all these different priors are of the form (13.3.3). They determined whether each of these priors yields a proper or improper posterior distribution and found that, under some technical conditions met by most correlation functions, the Jeffreys-rule prior and the reference prior obtained by exact marginalization are the only ones that always yield a proper posterior distribution. The following summarizes their key results.

Theorem 13.5 (Ren, Sun, and He, 2009). *Consider the model with sampling distribution (13.3.2) for the case described above. (i) The reference prior of* $(\beta,\sigma^2,\theta_1,\theta_2)$ *assuming* $(\sigma^2,\theta_1,\theta_2)$ *is the parameter of interest and* β *is the nuisance parameter is of the form (13.3.3), with* $a = 1$ *and* $\pi^R(\theta_1,\theta_2) \propto (\det[I^R(\theta_1,\theta_2)])^{\frac{1}{2}}$, *where* $I^R(\theta_1,\theta_2)$ *is given by (13.3.10) with* $q = 2$, $\frac{\partial}{\partial\theta_1}\Sigma_\theta = \frac{d}{d\theta_1}H_{\theta_1}$ *and* $\frac{\partial}{\partial\theta_2}\Sigma_\theta = I_n$. *(ii) The independence Jeffreys and Jeffreys-rule priors are of the form (13.3.3) with, respectively,* $a = 1$ *and* $\pi^{J1}(\theta_1,\theta_2) \propto (\det[I^J(\theta_1,\theta_2)])^{\frac{1}{2}}$, *and* $a = 1 + \frac{p}{2}$ *and* $\pi^{J2}(\theta_1,\theta_2) \propto |X'\Sigma_\theta^{-1}X|^{\frac{1}{2}}\pi^{J1}(\theta_1,\theta_2)$, *where* $I^J(\theta_1,\theta_2)$ *is similar to the matrix in (13.3.10), but with the 1-1 entry equal to n and the other entries obtained by replacing* Q_θ *with* Σ_θ^{-1}. *(iii) Under some assumptions met by most spatial correlation functions, the Jeffreys-rule prior always yields a proper posterior. On the*

other hand, the independence Jeffreys prior yields a proper posterior if and only if $\mathbf{1} \notin \mathscr{C}(X)$. *(iv) Under some assumptions met by most spatial correlation functions the reference prior always yields a proper posterior. Also, the marginal reference prior* $\pi^R(\theta_1, \theta_2)$ *is proper when* $\mathbf{1} \in \mathscr{C}(X)$.

Remark 13.8. It is worth noting that, in terms of propriety of posterior distributions, the results obtained for this multivariate isotropic case are all in agreement with those obtained for the univariate case by Berger, De Oliveira, and Sansó (2001), but disagree somewhat with those in Theorem 13.5 for the multivariate separable case. These differences can be attributed mainly to the different behaviors of $L^I(\theta; \mathbf{y})$ at infinity, which in turn are due to the different structures of the correlation function, isotropic or separable.

13.3.8 Discussion and Some Open Problems

Objective Bayesian methods for Gaussian random fields are still in the infancy stage, partly due to the limited current understanding of the likelihood behavior in these models. Also, the application of objective Bayesian methods in real dataset has been almost absent so far, so their practical value is still to be seen. The latter is mainly due to computational complexity since evaluation of reference priors, such as those in (13.3.5) and (13.3.10), is computationally quite expensive.

A point worth noting is that prediction/interpolation is often the main goal in the analysis of spatial datasets. The parameters, on the other hand, have only secondary importance and their classification (needed in the reference prior algorithm) is done more for operational rather than real reasons. A brief proposal on how to deal with situation where prediction is the main goal was given by Berger and Bernardo (1992a), and Kuboki (1998) proposed a more concrete approach. How to apply these methods in the current context and whether the same or different priors will result is still to be seen. Some relevant open problems in this area are: (i) Efficient approximation and computation of Jeffreys and reference priors; (ii) Understanding the behavior of Jeffreys and reference priors for 'smoothness/roughness' parameters; (iii) Understanding the behavior of Jeffreys and reference priors for the parameters of general isotropic correlation functions having all parameters (range, smoothness and nugget) unknown; and (iv) Development of objective Bayesian methods for complex hierarchical models that use Gaussian random fields as building blocks.

Acknowledgments: I warmly thank Jim Berger for having introduced default priors to me, and wish him the best for his 60th anniversary. This work was partially supported by the US National Science Foundation Grant DMS-0719508.

Chapter 14
Posterior Simulation and Monte Carlo Methods

A beauty of the Bayesian approach is the principled nature of inference. There is a gold standard of how to proceed, and the basic principle is easily explained. However, the actual implementation often gives rise to many challenges. One of the challenges that remains an important research frontier of Bayesian inference is the problem of numerically evaluating the desired posterior summaries. In this chapter we review some related specific research problems.

14.1 Importance Sampling Methods for Bayesian Discrimination between Embedded Models

Jean-Michel Marin and Christian P. Robert

The contribution of Jim Berger to the better understanding of Bayesian testing is fundamental and wide-ranging, from establishing the fundamental difficulties with p-values in Berger and Sellke (1987) to formalizing the intrinsic Bayes factors in Berger and Pericchi (1996a), to solving the difficulty with improper priors in Berger, Pericchi, and Varshavsky (1998), and beyond! While our contribution in this area is obviously much more limited, we aim at presenting here the most standard approaches to the approximation of Bayes factors.

The Bayes factor indeed is a fundamental procedure that stands at the core of the Bayesian theory of testing hypotheses, at least in the approach advocated by both Jeffreys (1961) and by Jaynes (2003). Note that Robert, Chopin, and Rousseau (2009) provide a reassessment of the crucial role of Jeffreys (1961) in setting a formal framework for Bayesian testing as well as for regular inference. Given an hypothesis $H_0 : \theta \in \Theta_0$ on the parameter $\theta \in \Theta$ of a statistical model, with observation y and density $f(y|\theta)$, under a compatible prior of the form

$$\pi(\Theta_0)\pi_0(\theta) + \pi(\Theta_0^c)\pi_1(\theta),$$

M.-H. Chen et al. (eds.), *Frontiers of Statistical Decision Making and Bayesian Analysis*, DOI 10.1007/978-1-4419-6944-6_14,

the *Bayes factor* is defined as the posterior odds to prior odds ratio, namely

$$B_{01}(y) = \frac{\pi(\Theta_0|y)}{\pi(\Theta_0^c|y)} \Big/ \frac{\pi(\Theta_0)}{\pi(\Theta_0^c)} = \int_{\Theta_0} f(y|\theta)\pi_0(\theta)\mathrm{d}\theta \Big/ \int_{\Theta_0^c} f(y|\theta)\pi_1(\theta)\mathrm{d}\theta.$$

Model choice can be considered from a similar perspective, since, under the Bayesian paradigm (see, e.g., Robert, 2001), the comparison of models

$$\mathfrak{M}_i : y \sim f_i(y|\theta_i), \quad \theta_i \sim \pi_i(\theta_i), \quad \theta_i \in \Theta_i, \quad i \in \mathfrak{I},$$

where the family $\mathfrak{I}$ can be finite or infinite, leads to posterior probabilities of the models under comparison such that

$$\mathbb{P}(\mathfrak{M} = \mathfrak{M}_i|y) \propto p_i \int_{\Theta_i} f_i(y|\theta_i)\pi_i(\theta_i)\mathrm{d}\theta_i,$$

where $p_i = \mathbb{P}(\mathfrak{M} = \mathfrak{M}_i)$ is the prior probability of model $\mathfrak{M}_i$.

In this short survey, we consider some of the most common Monte Carlo solutions used to approximate a generic Bayes factor or its fundamental component, the *evidence*

$$m_i = \int_{\Theta_i} \pi_i(\theta_i) f_i(y|\theta_i)\,\mathrm{d}\theta_i,$$

aka the marginal likelihood. Longer entries can be found in Carlin and Chib (1995), Chen, Shao, and Ibrahim (2000), Robert and Casella (2004), or Friel and Pettitt (2008). Note that we only briefly mention here trans-dimensional methods issued from the revolutionary paper of Green (1995), since our goal is to demonstrate that within-model simulation methods allow for the computation of Bayes factors and thus avoids the additional complexity involved in trans-dimensional methods. While amenable to an importance sampling technique of sorts, the alternative approach of nested sampling (Skilling, 2006) is discussed in Chopin and Robert (2007) and Robert and Wraith (2009).

14.1.1 The Pima Indian Benchmark Model

In order to compare the performances of all methods presented in this survey, we chose to evaluate the corresponding estimates of the Bayes factor in the setting of a single variable selection for a probit model and to repeat the estimation in a Monte Carlo experiment to empirically assess the variability of those estimates.

We recall that a probit model can be represented as a natural latent variable model in that, if we consider a sample $z_1, \ldots, z_n$ of n independent latent variables associated with a standard regression model, i.e. such that $z_i|\theta \sim N\left(\mathbf{x}_i^{\mathrm{T}}\theta, 1\right)$, where the $\mathbf{x}_i$'s are p-dimensional covariates and θ is the vector of regression coefficients, then $y_1, \ldots, y_n$ such that

$$y_i = \mathbb{I}_{z_i>0}$$

is a probit sample. Indeed, given θ, the y_i's are independent Bernoulli random variables with $\mathbb{P}(y_i = 1|\theta) = \Phi\left(\mathbf{x}_i^{\mathrm{T}}\theta\right)$, where Φ is the standard normal cumulative distribution function.

The choice of a reference prior distribution for the probit model is open to debate, but the connection with the latent regression model induced Marin and Robert (2007) to suggest a g-prior model, $\theta \sim N\left(0_p, n(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\right)$, with n as the g factor and $\mathbf{X}$ as the regressor matrix. The corresponding posterior distribution is then associated with the density

$$\pi(\theta|\mathbf{y},\mathbf{X}) \propto \prod_{i=1}^{n} \left\{1 - \Phi\left(\mathbf{x}_i^{\mathrm{T}}\theta\right)\right\}^{1-y_i} \Phi\left(\mathbf{x}_i^{\mathrm{T}}\theta\right)^{y_i} \times \exp\left\{-\theta^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})\theta/2n\right\}, \quad (14.1.1)$$

where $\mathbf{y} = (y_1, \ldots, y_n)$. In the completed model, i.e. when including the latent variables $\mathbf{z} = (z_1, \ldots, z_n)$ into the model, the y_i's are deterministic functions of the z_i's and the so-called completed likelihood is

$$f(\mathbf{y},\mathbf{z}|\theta) = (2\pi)^{-n/2} \exp\left(-\sum_{i=1}^{n} \left(z_i - \mathbf{x}_i^{\mathrm{T}}\theta\right)^2/2\right) \prod_{i=1}^{n} \left(\mathbb{I}_{y_i=0}\mathbb{I}_{z_i\leq 0} + \mathbb{I}_{y_i=1}\mathbb{I}_{z_i>0}\right).$$

The derived conditional distributions

$$z_i|y_i,\theta \sim \begin{cases} N_+\left(\mathbf{x}_i^{\mathrm{T}}\theta, 1, 0\right) \text{ if } & y_i = 1, \\ N_-\left(\mathbf{x}_i^{\mathrm{T}}\theta, 1, 0\right) \text{ if } & y_i = 0, \end{cases} \quad (14.1.2)$$

are of interest for constructing a Gibbs sampler on the completed model, where $N_+\left(\mathbf{x}_i^{\mathrm{T}}\theta, 1, 0\right)$ denotes the Gaussian distribution with mean $\mathbf{x}_i^{\mathrm{T}}\theta$ and variance 1 that is left-truncated at 0, while $N_-\left(x_i^{\mathrm{T}}\theta, 1, 0\right)$ denotes the symmetrical normal distribution that is right-truncated at 0. The corresponding full conditional on the parameters is given by

$$\theta|\mathbf{y},\mathbf{z} \sim N\left(\frac{n}{n+1}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{z}, \frac{n}{n+1}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\right). \quad (14.1.3)$$

Indeed, since direct simulation from the posterior distribution of θ is intractable, Albert and Chib (1993) suggest implementing a Gibbs sampler based on the above set of full conditionals. More precisely, given the current value of θ, one cycle of the Gibbs algorithm produces a new value for $\mathbf{z}$ as simulated from the conditional distribution (14.1.2), which, when substituted into (14.1.3), produces a new value for θ. Although it does not impact the long-term properties of the sampler, the starting value of θ may be taken as the maximum likelihood estimate to avoid burning steps in the Gibbs sampler.

Given this probit model, the dataset we consider covers a population of women who were at least 21 years old, of Pima Indian heritage and living near Phoenix, Arizona. These women were tested for diabetes according to World Health Organization (WHO) criteria. The data were collected by the US National Institute of Diabetes and Digestive and Kidney Diseases, and is available with the basic R package (R Development Core Team, 2008). This dataset, used as a benchmark for super-

vised learning methods, contains information about 332 women with the following variables:

– `glu`: plasma glucose concentration in an oral glucose tolerance test;
– `bp`: diastolic blood pressure (mm Hg);
– `ped`: diabetes pedigree function;
– `type`: Yes or No, for diabetic according to WHO criteria.

For this dataset, the goal is to explain the diabetes variable `type` by using the explanatory variables `glu`, `bp` and `ped`. The following table is an illustration of a classical (maximum likelihood) analysis of this dataset, obtained using the R glm() function with the probit link:

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1347  -0.9217  -0.6963   0.9959   2.3235
Coefficients:
     Estimate Std. Error z value Pr(>|z|)
glu  0.012616   0.002406   5.244 1.57e-07 ***
bp  -0.029050   0.004094  -7.096 1.28e-12 ***
ped  0.350301   0.208806   1.678   0.0934 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*'
                0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family
 taken to be 1)

    Null deviance: 460.25  on 332  degrees of freedom
Residual deviance: 386.73  on 329  degrees of freedom
AIC: 392.73
Number of Fisher Scoring iterations: 4
```

This analysis sheds some doubt on the relevance of the covariate `ped` in the model and we can reproduce the study from a Bayesian perspective, computing the Bayes factor B_{01} opposing the probit model only based on the covariates `glu` and `bp` (model 0) to the probit model based on the covariates `glu`, `bp`, and `ped` (model 1). This is equivalent to testing the hypothesis $H_0 : \theta_3 = 0$ since the models are nested, where θ_3 is the parameter of the probit model associated with covariate `ped`. (Note that there is no intercept in either model.) If we denote by $\mathbf{X}_0$ the 332×2 matrix containing the values of `glu` and `bp` for the 332 individuals and by $\mathbf{X}_1$ the 332×3 matrix containing the values of the covariates `glu`, `bp`, and `ped`, the Bayes factor B_{01} is given by

$$B_{01} = (2\pi)^{1/2} n^{1/2} \frac{|(\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0)|^{-1/2}}{|(\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1)|^{-1/2}}$$
$$\times \frac{\int_{\mathbb{R}^2} \prod_{i=1}^n \{1-\Phi((\mathbf{X}_0)_{i,\cdot}\theta)\}^{1-y_i} \Phi((\mathbf{X}_0)_{i,\cdot}\theta)^{y_i} \exp\left\{-\theta^{\mathrm{T}}(\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0)\theta/2n\right\} \mathrm{d}\theta}{\int_{\mathbb{R}^3} \prod_{i=1}^n \{1-\Phi(\mathbf{X}_1)_{i,\cdot}\theta)\}^{1-y_i} \Phi(\mathbf{X}_1)_{i,\cdot}\theta)^{y_i} \exp\left\{-\theta^{\mathrm{T}}(\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1)\theta/2n\right\} \mathrm{d}\theta}$$
$$= \frac{\mathbb{E}_{N_2(0_2, n(\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0)^{-1})} \left[\prod_{i=1}^n \{1-\Phi((\mathbf{X}_0)_{i,\cdot}\theta)\}^{1-y_i} \Phi((\mathbf{X}_0)_{i,\cdot}\theta)^{y_i}\right]}{\mathbb{E}_{N_3(0_3, n(\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1)^{-1})} \left[\prod_{i=1}^n \{1-\Phi((\mathbf{X}_1)_{i,\cdot}\theta)\}^{1-y_i} \Phi((\mathbf{X}_1)_{i,\cdot}\theta)^{y_i}\right]} \tag{14.1.4}$$

using the shortcut notation that $A_{i,\cdot}$ is the i^{th} line of the matrix A.

14.1.2 The Basic Monte Carlo Solution

As already shown above, when testing for a null hypothesis (or a model) $H_0 : \theta \in \Theta_0$ against the alternative hypothesis (or the alternative model) $H_1 : \theta \in \Theta_1$, the Bayes factor is defined by

$$B_{01}(y) = \int_{\Theta_0} f(y|\theta_0)\pi_0(\theta_0)\mathrm{d}\theta_0 \Big/ \int_{\Theta_1} f(y|\theta_1)\pi_1(\theta_1)\mathrm{d}\theta_1.$$

We assume in this survey that the prior distributions under both the null and the alternative hypotheses are proper, as, typically, they should be. (In the case of common nuisance parameters, a common improper prior measure can be used on those, see Berger, Pericchi, and Varshavsky (1998), and Marin and Robert (2007). This obviously complicates the computational aspect, as some methods like crude Monte Carlo cannot be used at all, while others are more prone to suffer from infinite variance.) In that setting, the most elementary approximation to $B_{01}(y)$ consists in using a ratio of two standard Monte Carlo approximations based on simulations from the corresponding priors. Indeed, for $i = 0, 1$:

$$\int_{\Theta_i} f(y|\theta)\pi_i(\theta)\mathrm{d}\theta = \mathbb{E}_{\pi_i}[f(y|\theta)].$$

If $\theta_{0,1}, \ldots, \theta_{0,n_0}$ and $\theta_{1,1}, \ldots, \theta_{1,n_1}$ are two independent samples generated from the prior distributions π_0 and π_1, respectively, then

$$\frac{n_0^{-1} \sum_{j=1}^{n_0} f(y|\theta_{0,j})}{n_1^{-1} \sum_{j=1}^{n_1} f(y|\theta_{1,j})} \tag{14.1.5}$$

is a strongly consistent estimator of $B_{01}(y)$.

In most cases, sampling from the prior distribution corresponding to either hypothesis is straightforward and fast. Therefore, the above estimator is extremely easy to derive as a brute-force evaluation of the Bayes factor. However, if any of the posterior distributions is quite different from the corresponding prior distribution—

and it should be for vague priors—the Monte Carlo evaluation of the corresponding evidence is highly inefficient since the sample will be overwhelmingly producing negligible values of $f(y|\theta_{i,j})$. In addition, if $f^2(y|\theta)$ is not integrable against π_0 or π_1, the resulting estimation has an infinite variance. Since importance sampling usually requires an equivalent computation effort, with a potentially highly efficiency reward, crude Monte Carlo approaches of this type are usually disregarded.

Figure 14.1 and Table 14.1 summarize the results based on 100 replications of Monte Carlo approximations of $B_{01}(y)$, using equation (14.1.5) with $n_0 = n_1 = 20,000$ simulations. As predicted, the variability of the estimator is very high, when compared with the other estimates studied in this survey. (Obviously, the method is asymptotically unbiased and, the functions being square integrable in (14.1.4), with a finite variance. A massive simulation effort would obviously lead to a precise estimate of the Bayes factor.)

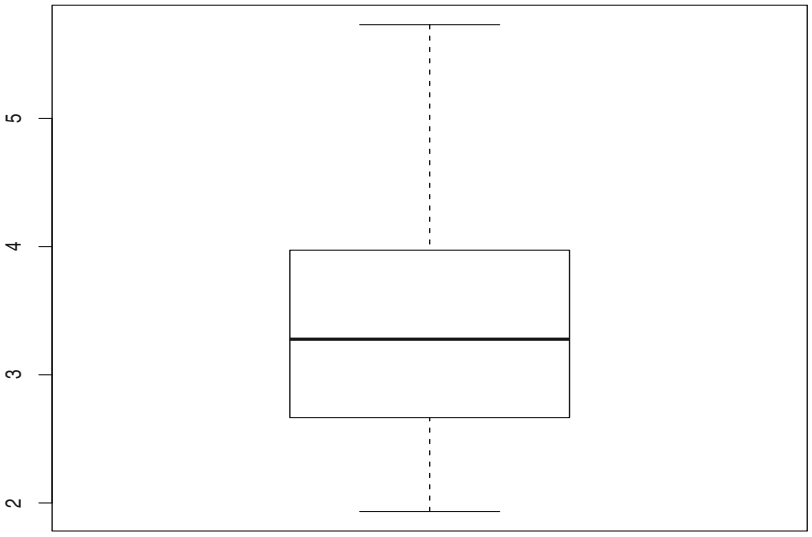


FIGURE 14.1. Pima Indian dataset: boxplot of 100 Monte Carlo estimates of $B_{01}(y)$ based on simulations from the prior distributions, for 2×10^4 simulations.

14.1.3 Usual Importance Sampling Approximations

Defining two importance distributions with densities ϖ_0 and ϖ_1, with the same supports as π_0 and π_1, respectively, we have:

$$B_{01}(y) = \mathbb{E}_{\varpi_0}\left[f(y|\theta)\pi_0(\theta)/\varpi_0(\theta)\right] \Big/ \mathbb{E}_{\varpi_1}\left[f(y|\theta)\pi_1(\theta)/\varpi_1(\theta)\right].$$

Therefore, given two independent samples generated from distributions ϖ_0 and ϖ_1, $\theta_{0,1},\ldots,\theta_{0,n_0}$ and $\theta_{1,1},\ldots,\theta_{1,n_1}$, respectively, the corresponding importance sampling estimate of $B_{01}(y)$ is

$$\frac{n_0^{-1}\sum_{j=1}^{n_0} f_0(x|\theta_{0,j})\pi_0(\theta_{0,j})/\varpi_0(\theta_{0,j})}{n_1^{-1}\sum_{j=1}^{n_1} f_1(x|\theta_{1,j}\pi_1(\theta_{1,j})/\varpi_1(\theta_{1,j})}. \tag{14.1.6}$$

Compared with the standard Monte Carlo approximation above, this approach offers the advantage of opening the choice of the representation (14.1.6) in that it is possible to pick importance distributions ϖ_0 and ϖ_1 that lead to a significant reduction in the variance of the importance sampling estimate. This implies choosing importance functions that provide as good as possible approximations to the corresponding posterior distributions. Maximum likelihood asymptotic distributions or kernel approximations based on a sample generated from the posterior are natural candidates in this setting, even though the approximation grows harder as the dimension increases.

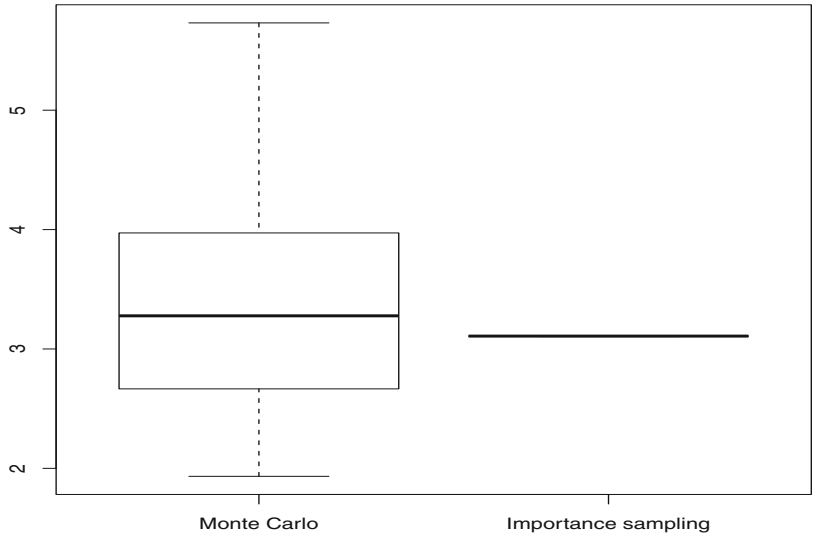


FIGURE 14.2. Pima Indian dataset: boxplots of 100 Monte Carlo and importance sampling estimates of $B_{01}(y)$, based on simulations from the prior distributions, for 2×10^4 simulations.

For the Pima Indian benchmark, we propose for instance to use as importance distributions, Gaussian distributions with means equal to the maximum likelihood (ML) estimates and covariance matrices equal to the estimated covariance matrices

of the ML estimates, both of which are provided by the R glm() function. While, in general, those Gaussian distributions provide crude approximations to the posterior distributions, the specific case of the probit model will show this is an exceptionally good approximation to the posterior, since this leads to the best solution among all those compared here. The results, obtained over 100 replications of the methodology with $n_0 = n_1 = 20{,}000$ are summarized in Figure 14.2 and Table 14.1. They are clearly excellent, while requiring the same computing time as the original simulation from the prior.

14.1.4 Bridge Sampling Methodology

The original version of the bridge sampling approximation to the Bayes factor (Gelman and Meng, 1998; Chen, Shao, and Ibrahim, 2000) relies on the assumption that the parameters of both models under comparison belong to the same space: $\Theta_0 = \Theta_1$. In that case, for likelihood functions f_0 and f_1 under respectively models $\mathfrak{M}_0$ and $\mathfrak{M}_1$, the bridge representation of the Bayes factor is

$$B_{01}(y) = \frac{\int_{\Theta_0} f_0(y|\theta)\pi_0(\theta)\mathrm{d}\theta}{\int_{\Theta_1} f_1(y|\theta)\pi_1(\theta)\mathrm{d}\theta} = \mathbb{E}_{\pi_1}\left[\frac{f_0(y|\theta)\pi_0(\theta)}{f_1(y|\theta)\pi_1(\theta)}\middle| y\right]. \tag{14.1.7}$$

Given a sample from the posterior distribution of θ under model $\mathfrak{M}_1$, $\theta_{1,1},\ldots,\theta_{1,N} \sim \pi_1(\theta|y)$, a first bridge sampling approximation to $B_{01}(y)$ is

$$N^{-1}\sum_{j=1}^{N}\frac{f_0(y|\theta_{1,j})\pi_0(\theta_{1,j})}{f_1(y|\theta_{1,j})\pi_1(\theta_{1,j})}.$$

From a practical perspective, for the above bridge sampling approximation to be of any use, the constraint on the common parameter space for both models goes further in that, not only must both models have the same complexity, but they must also be parameterized on a common ground, i.e. in terms of some specific moments of the sampling model, so that parameters under both models have a common meaning. Otherwise, the resulting bridge sampling estimator will have very poor convergence properties, possibly with infinite variance.

Equation (14.1.7) is nothing but a very special case of the general representation (Torrie and Valleau, 1977).

$$B_{01}(y) = \mathbb{E}_{\varphi}\left[f_0(y|\theta)\pi_0(\theta)/\varphi(\theta)\right] \Big/ \mathbb{E}_{\varphi}\left[f_1(y|\theta)\pi_1(\theta)/\varphi(\theta)\right],$$

which holds for any density φ with a sufficiently large support and which only requires a single sample $\theta_1,\ldots,\theta_N$ generated from φ to produce an importance sampling estimate of the ratio of the marginal likelihoods. Apart from using the *same* importance function φ for both integrals, this method is therefore a special case of importance sampling.

Another extension of this bridge sampling approach is based on the general representation

$$B_{01}(y) = \int f_0(y|\theta)\pi_0(\theta)\alpha(\theta)\pi_1(\theta|y)\mathrm{d}\theta \Big/ \int f_1(y|\theta)\pi_1(\theta)\alpha(\theta)\pi_0(\theta|y)\mathrm{d}\theta$$
$$\approx \frac{{n_1}^{-1}\sum_{j=1}^{n_1} f_0(y|\theta_{1,j})\pi_0(\theta_{1,j})\alpha(\theta_{1,j})}{{n_0}^{-1}\sum_{j=1}^{n_0} f_1(y|\theta_{0,j})\pi_1(\theta_{0,j})\alpha(\theta_{0,j})},$$

where $\theta_{0,1},\ldots,\theta_{0,n_0}$ and $\theta_{1,1},\ldots,\theta_{1,n_1}$ are two independent samples coming from the posterior distributions $\pi_0(\theta|y)$ and $\pi_1(\theta|y)$, respectively. That applies for any positive function α as long as the upper integral exists. Some choices of α lead to very poor performances of the method in connection with the harmonic mean approach (see Section 14.1.5), but there exists a quasi-optimal solution, as provided by Gelman and Meng (1998):

$$\alpha^\star(y) \propto \frac{1}{n_0\pi_0(\theta|y) + n_1\pi_1(\theta|y)}.$$

This optimum cannot be used *per se*, since it requires the normalizing constants of both $\pi_0(\theta|y)$ and $\pi_1(\theta|y)$. As suggested by Gelman and Meng (1998), an approximate version uses iterative versions of $\alpha^\star$, based on iterated approximations to the Bayes factor. Note that this solution recycles simulations from both posteriors, which is quite appropriate since one model is selected via the Bayes factor, instead of using an importance weighted sample common to both approximations. We will see below an alternative representation of the bridge factor that bypasses this difficulty (if difficulty there is!).

Those derivations are, however, restricted to the case where both models have the same complexity and thus they do not apply to embedded models, when $\Theta_0 \subset \Theta_1$ in such a way that $\theta_1 = (\theta, \psi)$, i.e., when the submodel corresponds to a specific value ψ_0 of ψ: $f_0(y|\theta) = f(y|\theta, \psi_0)$.

The extension of the most advanced bridge sampling strategies to such cases requires the introduction of a *pseudo-posterior density,* $\omega(\psi|\theta, y)$, on the parameter that does not appear in the embedded model, in order to reconstitute the equivalence between both parameter spaces. Indeed, if we augment $\pi_0(\theta|y)$ with $\omega(\psi|\theta, y)$, we obtain a joint distribution with density $\pi_0(\theta|y) \times \omega(\psi|\theta, y)$ on Θ_1. The Bayes factor can then be expressed as

$$B_{01}(y) = \frac{\int_{\Theta_1} f(y|\theta,\psi_0)\pi_0(\theta)\alpha(\theta,\psi)\pi_1(\theta,\psi|y)\mathrm{d}\theta\,\omega(\psi|\theta,y)\,\mathrm{d}\psi}{\int_{\Theta_1} f(y|\theta,\psi)\pi_1(\theta,\psi)\alpha(\theta,\psi)\pi_0(\theta|y) \times \omega(\psi|\theta,y)\mathrm{d}\theta\,\mathrm{d}\psi}, \qquad (14.1.8)$$

for all functions $\alpha(\theta, \psi)$, because it is clearly independent from the choice of both $\alpha(\theta, \psi)$ and $\omega(\psi|\theta, y)$. Obviously, the performances of the approximation

$$\frac{(n_1)^{-1}\sum_{j=1}^{n_1} f(y|\theta_{1,j},\psi_0)\pi_0(\theta_{1,j})\omega(\psi_{1,j}|\theta_{1,j},y)\alpha(\theta_{1,j},\psi_{1,j})}{(n_0)^{-1}\sum_{j=1}^{n_0} f(y|\theta_{0,j},\psi_{0,j})\pi_1(\theta_{0,j},\psi_{0,j})\alpha(\theta_{0,j},\psi_{0,j})},$$

where $(\theta_{0,1}, \psi_{0,1}), \ldots, (\theta_{0,n_0}, \psi_{0,n_0})$ and $(\theta_{1,1}, \psi_{1,1}), \ldots, (\theta_{1,n_1}, \psi_{1,n_1})$ are two independent samples generated from distributions $\pi_0(\theta|y) \times \omega(\psi|\theta, y)$ and $\pi_1(\theta, \psi|y)$, respectively, do depend on this completion by the pseudo-posterior as well as on the function $\alpha(\theta, \psi)$. Chen, Shao, and Ibrahim (2000) establish that the asymptotically optimal choice for $\omega(\psi|\theta, y)$ is the obvious one, namely

$$\omega(\psi|\theta, y) = \pi_1(\psi|\theta, y),$$

which most often is unavailable in closed form (especially when considering that the normalizing constant of $\omega(\psi|\theta, y)$ is required in (14.1.8)). However, in latent variable models, approximations of the conditional posteriors often are available, as detailed in Section 14.1.6.

While this extension of the basic bridge sampling approximation is paramount for handling embedded models, its implementation suffers from the dependence on this pseudo-posterior. In addition, this technical device brings the extended bridge methodology close to the cross-model alternatives of Carlin and Chib (1995) and Green (1995), in that both those approaches rely on completing distributions, either locally (Green 1995) or globally (Carlin and Chib, 1995), to link both models under comparison in a bijective relation. The density $\omega(\psi|\theta_0, y)$ is then a pseudo-posterior distribution in Chib and Carlin's (1995) sense, and it can be used as Green's (1995) proposal in the reversible jump MCMC step to move (or not) from model $\mathfrak{M}_0$ to model $\mathfrak{M}_1$. While using cross-model solutions to compare only two models does seem superfluous, given that the randomness in picking the model at each step of the simulation is not as useful as in the setting of comparing a large number or an infinity of models, the average acceptance probability for moving from model $\mathfrak{M}_0$ to model $\mathfrak{M}_1$ is related to the Bayes factor since

$$\mathbb{E}_{\pi_0 \times \omega} \left[\frac{f(y|\theta, \psi)\pi_1(\theta, \psi)}{f(y|\theta, \psi_0)\pi_0(\theta)\omega(\psi|\theta, y)} \right] = B_{01}(y)$$

even though the average

$$\mathbb{E}_{\pi_0 \times \omega} \left[\min \left\{ 1, \frac{f(y|\theta, \psi)\pi_1(\theta, \psi)}{f(y|\theta, \psi_0)\pi_0(\theta)\omega(\psi|\theta, y)} \right\} \right]$$

does not provide a closed form solution.

For the Pima Indian benchmark, we use as pseudo-posterior density $\omega(\theta_3|\theta_1, \theta_2, y)$, the conditional Gaussian density deduced from the asymptotic Gaussian distribution on $(\theta_1, \theta_2, \theta_3)$ already used in the importance sampling solution, with mean equal to the ML estimate of $(\theta_1, \theta_2, \theta_3)$ and with covariance matrix equal to the estimated covariance matrix of the ML estimate. The quasi-optimal solution $\alpha^\star$ in the bridge sampling estimate is replaced with the inverse of an average between the asymptotic Gaussian distribution in model $\mathfrak{M}_1$ and the product of the asymptotic Gaussian distribution in model $\mathfrak{M}_0$ times the above $\omega(\theta_3|\theta_1, \theta_2, y)$. This obviously is a suboptimal choice, but it offers the advantage of providing a non-iterative solution. The results, obtained over 100 replications of the methodology

with $n_0 = n_1 = 20,000$ are summarized in Figure 14.3 and Table 14.1. The left-hand graph shows that this choice of bridge sampling estimator produces a solution whose variation is quite close to the (excellent) importance sampling solution, a considerable improvement upon the initial Monte Carlo estimator. However, the right-hand-side graph shows that the importance sampling solution remains far superior, especially when accounting for the computing time. (In this example, running 20,000 iterations of the Gibbs sampler for the models with both two and three variables takes approximately 32 seconds.)

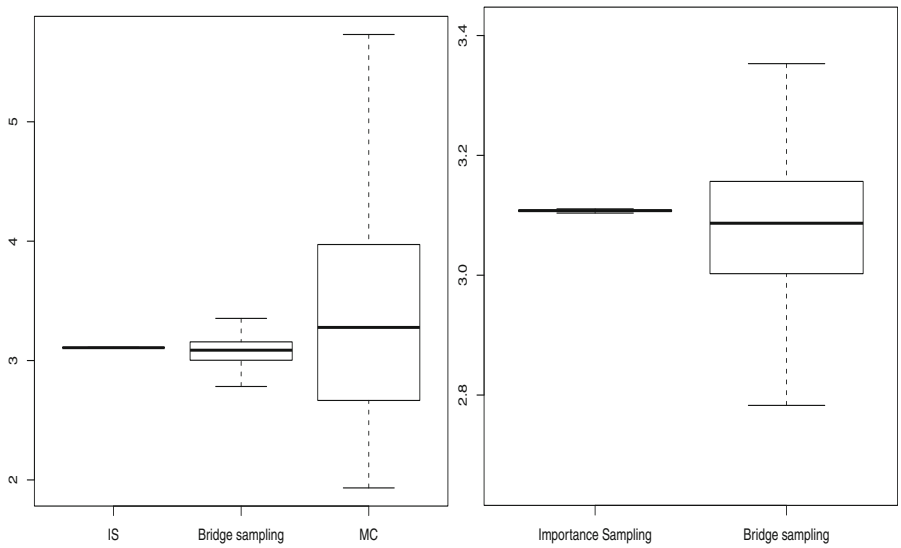


FIGURE 14.3. Pima Indian dataset: *(left)* boxplots of 100 importance sampling, bridge sampling and Monte Carlo estimates of $B_{01}(y)$, based on simulations from the prior distributions, for 2×10^4 simulations; *(right)* same comparison for the importance sampling versus bridge sampling estimates only.

14.1.5 Harmonic Mean Approximations

While using the generic harmonic mean approximation to the marginal likelihood is often fraught with danger (Neal, 1994), the representation (Gelfand and Dey, 1994) $(k = 0, 1)$

$$\mathbb{E}_{\pi_k}\left[\frac{\varphi_k(\theta)}{\pi_k(\theta)f_k(y|\theta)}\middle| y\right] = \int \frac{\varphi_k(\theta)}{\pi_k(\theta)f_k(y|\theta)}\frac{\pi_k(\theta)f_k(y|\theta)}{m_k(y)}\,\mathrm{d}\theta = \frac{1}{m_k(y)} \qquad (14.1.9)$$

holds, no matter what the density $\varphi_k(\theta)$ is—provided $\varphi_k(\theta) = 0$ when $\pi_k(\theta) f_k(y|\theta) = 0$—. This representation is remarkable in that it allows for a direct processing of Monte Carlo or MCMC output from the posterior distribution $\pi_k(\theta|y)$. As with importance sampling approximations, the variability of the corresponding estimator of $B_{01}(y)$ will be small if the distributions $\varphi_k(\theta)$ $(k = 0, 1)$ are close to the corresponding posterior distributions. However, as opposed to usual importance sampling constraints, the density $\varphi_k(\theta)$ must have lighter—rather than fatter—tails than $\pi_k(\theta) f_k(y|\theta)$ for the approximation of the marginal $m_k(y)$

$$\left\{ \frac{1}{N} \sum_{j=1}^{N} \frac{\varphi_k(\theta_{k,j})}{\pi_k(\theta_{k,j}) f_k(y|\theta_{k,j})} \right\}^{-1}$$

to enjoy finite variance. For instance, using $\varphi_k(\theta) = \pi_k(\theta)$ as in the original harmonic mean approximation (Newton and Raftery, 1994) will most usually result in an infinite variance estimator, as discussed by Neal (1994). On the opposite, using φ_k's with constrained supports derived from a Monte Carlo sample, like the convex hull of the simulations corresponding to the 10% or to the 25% HPD regions—that again is easily derived from the simulations—is both completely appropriate and implementable (Robert and Wraith, 2009).

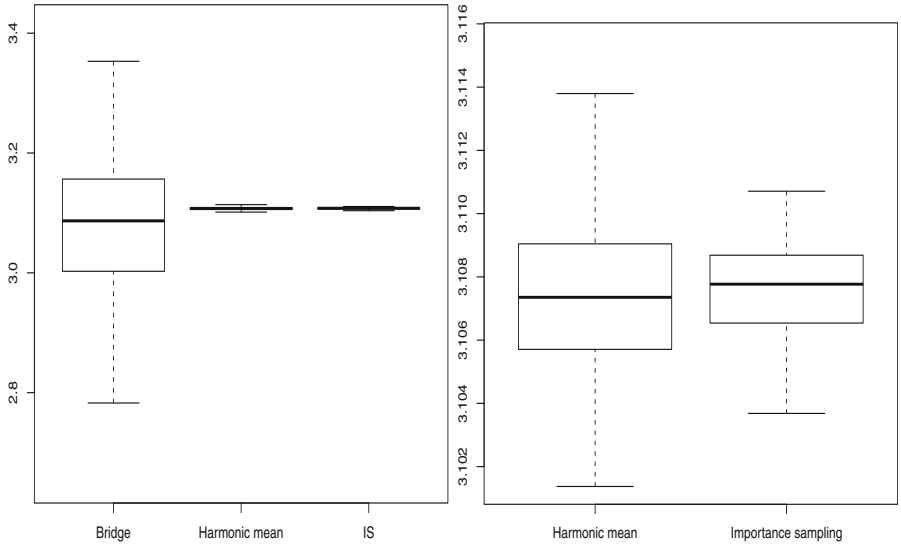


FIGURE 14.4. Pima Indian dataset: *(left)* boxplots of 100 bridge sampling, harmonic mean and importance sampling estimates of $B_{01}(y)$, based on simulations from the prior distributions, for 2×10^4 simulations; *(right)* same comparison for the harmonic mean versus importance sampling estimates only.

However, for the Pima Indian benchmark, we propose to use instead as our distributions $\varphi_k(\theta)$ the very same distributions as those used in the above importance sampling approximations, that is, Gaussian distributions with means equal to the ML estimates and covariance matrices equal to the estimated covariance matrices of the ML estimates. The results, obtained over 100 replications of the methodology with $N = 20,000$ simulations for each approximation of $m_k(y)$ $(k = 0, 1)$ are summarized in Figure 14.4 and Table 14.1. They show a very clear proximity between both importance solutions in this special case and a corresponding domination of the bridge sampling estimator, even though the importance sampling estimate is much faster to compute. This remark must be toned down by considering that the computing time due to the Gibbs sampler should not necessarily be taken into account into the comparison, since samples are generated under both models.

14.1.6 Exploiting Functional Equalities

Chib's (1995) method for approximating a marginal (likelihood) is a direct application of Bayes' theorem: given $y \sim f_k(y|\theta)$ and $\theta \sim \pi_k(\theta)$, we have that

$$m_k = \frac{f_k(y|\theta)\,\pi_k(\theta)}{\pi_k(\theta|y)},$$

for all θ's (since both the lhs and the rhs of this equation are constant in θ). Therefore, if an arbitrary value of θ, say θ_k^*, is selected and if a good approximation to $\pi_k(\theta|y)$ can be constructed, denoted $\hat{\pi}(\theta|y)$, Chib's (1995) approximation to the evidence is

$$m_k = \frac{f_k(y|\theta_k^*)\,\pi_k(\theta_k^*)}{\hat{\pi}_k(\theta_k^*|y)}. \tag{14.1.10}$$

In a general setting, $\hat{\pi}(\theta|y)$ may be the Gaussian approximation based on the MLE, already used in the importance sampling, bridge sampling and harmonic mean solutions, but this is unlikely to be accurate in a general framework. A second solution is to use a nonparametric approximation based on a preliminary MCMC sample, even though the accuracy may also suffer in large dimensions. In the special setting of latent variables models (like mixtures of distributions but also like probit models), Chib's (1995) approximation is particularly attractive as there exists a natural approximation to $\pi_k(\theta|y)$, based on the Rao–Blackwell (Gelfand and Smith, 1990) estimate

$$\hat{\pi}_k(\theta_k^*|y) = \frac{1}{T}\sum_{t=1}^{T} \pi_k(\theta_k^*|y, z_k^{(t)}),$$

where the $z_k^{(t)}$'s are the latent variables simulated by the MCMC sampler. The estimate $\hat{\pi}_k(\theta_k^*|y)$ is a parametric unbiased approximation of $\pi_k(\theta_k^*|y)$ that converges with rate $O(\sqrt{T})$. This Rao–Blackwell approximation obviously requires the full

conditional density $\pi_k(\theta_k^*|y,z)$ to be available in closed form (constant included) but, as already explained, this is the case for the probit model.

Figure 14.5 and Table 14.1 summarize the results obtained for 100 replications of Chib's approximations of $B_{01}(y)$ with $T = 20,000$ simulations for each approximation of $m_k(y)$ $(k = 0, 1)$. While Chib's method is usually very reliable and dominates importance sampling, the incredibly good approximation provided by the asymptotic Gaussian distribution implies that, in this particular case, Chib's method is dominated by both the importance sampling and the harmonic mean estimates.

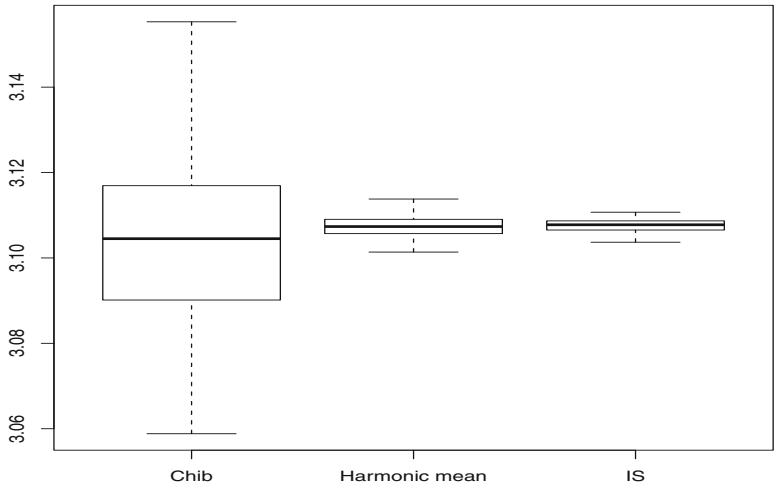


FIGURE 14.5. Pima Indian dataset: boxplots of 100 Chib's, harmonic mean and importance estimates of $B_{01}(y)$, based on simulations from the prior distributions, for 2×10^4 simulations.

TABLE 14.1. Pima Indian dataset: Performances of the various approximation methods used in this survey.

	Monte Carlo	Importance sampling	Bridge sampling	Harmonic mean	Chib's approximation
Median	3.277	3.108	3.087	3.107	3.104
Standard deviation	0.7987	0.0017	0.1357	0.0025	0.0195
Duration in seconds	7	7	71	70	64

14.1.7 Conclusion

In this short evaluation of the most common estimations to the Bayes factor, we have found that a particular importance sampling and its symmetric harmonic mean counterpart are both very efficient in the case of the probit model. The bridge sampling estimate is much less efficient in this example, due to the approximation error resulting from the pseudo-posterior. In most settings, the bridge sampling is actually doing better than the equivalent importance sampler Robert and Wraith (2009), while Chib's method is much more generic than the four alternatives. The recommendation resulting from the short experiment above is therefore to look for handy approximations to the posterior distribution, whenever available, but to fall back on Chib's method as a backup solution providing a reference or better.

Acknowledgments: J.-M. Marin and C.P. Robert are supported by the 2009–2012 grant ANR-09-BLAN-0218 "Big'MC".

14.2 Bayesian Computation and the Linear Model

Matthew J. Heaton and James G. Scott

The linear model is a venerable topic, and one that may even seem passé in light of the past decade's revolution in applied Bayesian nonparametric modeling. Yet despite its apparent simplicity, the linear model remains as important as ever to the practice of modern Bayesian statistics, for at least three reasons.

First, many data sets are simply too high-dimensional to be modeled using the slickest, newest methods. Computers run out of memory; Markov chains fail to converge; priors become prohibitively difficult to elicit or choose in a default way. Already this is a problem with data that arises in genetics and SNP association studies. Yet these data sets are small compared to those concerning Internet traffic that, for example, Google or Microsoft encounter every day. When a linear model is all that can be fit, it should be fit using the best available statistical and computational tools.

Second, many practitioners will fit a linear model to their data as a first pass, and will never make, or never be able to publish, a second pass. Indeed, many decisions in public health and policy are made using the results of a linear regression, with choices of great consequence coming down to the question of whether a particular term is "significant." Echoing the above: when a linear model is all that *will* be fit, it should be fit using the best available statistical and computational tools.

Finally, some nonparametric, nonlinear models can be recast as parametric, linear ones. For example, many kernel-regression problems correspond to expanding a function as a linear combination of basis elements given by the orthonormal eigenfunctions of an integral operator. Similarly, methods based on wavelets, splines, Fourier polynomials, and many other "dictionaries" of basis functions can be treated as little more than linear regression, and yet are capable of fitting highly nonlinear

functions. A hypothetical Bayesian who knew only how to fit "$Y = X\beta$ + error" could still handle a vast array of problems, simply by being clever about the choice of X.

Complicating matters is the fact that Bayesian linear modeling can generate many potential summaries for a high-dimensional data set, and that each summary corresponds, in some sense, to a different inferential goal. These summaries can include posterior means or medians of regression coefficients, variable-inclusion probabilities, and the posterior probabilities of models themselves (where a model is a specific combination of coefficients being identically zero). Section 14.2.3, for example, considers a data set where ozone-concentration levels around Los Angeles are regressed upon 65 possible atmospheric predictors. One could ask at least three different, scientifically relevant, questions concerning this data:

1. **Which subset of atmospheric variables best accounts for observed variation in past ozone levels?** This question can, in principle, be answered by finding the model with the highest posterior probability, given the data and prior assumptions.
2. **Which subset of atmospheric variables should be used to predict future ozone levels?** It is known that model-averaged predictions are generally best, but this is unsatisfactory if one must choose a single model to use for prediction. In orthogonal and nested-model settings, the best model to use for prediction is the median probability model (Barbieri and Berger, 2004). But in general, it is unknown whether there exists a single best model to use for prediction.
3. **What numbers should be used to yield the best estimate of the marginal effect of each variable on ozone levels?** Here, as above, the model-averaged estimates of the coefficients are generally best.

As this list suggests, different methodological approaches may work better for different questions. This paper seeks to reach a complementary understanding of different computational strategies. Indeed, we find that no single computational strategy works best for all of these problems—a fact that is both interesting and surprising, given that all strategies are, fundamentally, trying to reconstruct the same joint distribution over data, models, and parameters. In light of this fact, it is important to understand each strategy's strengths and weaknesses.

Our approach differs from existing review papers on Bayesian linear models in two main ways:

1. We focus less on well-established material regarding traditional MCMC, and more on recent innovations involving stochastic search and adaptive MCMC.
2. We provide a computational and methodological overview of "pure shrinkage" solutions, which have been the subject of a recent surge in research activity. An example of a pure-shrinkage solution is to place exchangeable double-exponential priors on the regression coefficients, a tactic which often goes by the name of "the Bayesian LASSO" (Park and Casella, 2008).

Additionally, we also review some recent developments about shrinkage and variable selection that are not explicitly computational in nature. We include these re-

sults in an attempt to give a current picture of the "state of the art" for Bayesian linear modeling.

14.2.1 Bayesian Linear Models

14.2.1.1 Notation

Given a vector Y of n responses and an $n \times p$ design matrix X, suppose we wish to select a subset of k predictors, zeroing out the remaining $p-k$ coefficients. This yields a sparse linear model of the form

$$Y_i = \alpha + X_{ij_1}\beta_{j_1} + \ldots + X_{ij_k}\beta_{j_k} + \varepsilon_i\,, \tag{14.2.1}$$

for some $\{j_1,\ldots,j_k\} \subset \{1,\ldots,p\}$, where $\varepsilon_i \overset{iid}{\sim} N(0,\phi^{-1})$.

We follow the convention of treating the intercept α differently, since all models will include this term. Let H_0 denote the null model with only an intercept, and let H_F denote the full model with all covariates included. The full model thus has parameter vector $\theta' = (\alpha,\beta')$, $\beta' = (\beta_1,\ldots,\beta_p)'$.

Each model H_γ is indexed by a binary vector γ of length p indicating a set of $k_\gamma \leq p$ nonzero regression coefficients β_γ:

$$\gamma_i = \begin{cases} 0 \text{ if } \beta_i = 0 \\ 1 \text{ if } \beta_i \neq 0\,. \end{cases}$$

In Bayesian model selection, γ itself is a random variable that takes values in the discrete space $\{0,1\}^p$, which has 2^p members. Inference relies upon the prior probability of each model, $p(H_\gamma)$, along with the marginal likelihood of the data under each model:

$$f(\mathbf{Y} \mid H_\gamma) = \int f(\mathbf{Y} \mid \theta_\gamma,\phi)\pi(\theta_\gamma,\phi)d\theta_\gamma d\phi, \tag{14.2.2}$$

where $\pi(\theta_\gamma,\phi)$ is the prior for model-specific parameters. These together define, up to a constant, the posterior probability of a model:

$$p(H_\gamma \mid \mathbf{Y}) \propto p(H_\gamma) f(\mathbf{Y} \mid H_\gamma). \tag{14.2.3}$$

Let $\mathbf{X}_\gamma$ denote the columns of the full design matrix $\mathbf{X}$ given by the nonzero elements of γ, and let $\mathbf{X}^*_\gamma$ denote the concatenation $(\mathbf{1}\ \mathbf{X}_\gamma)$, where $\mathbf{1}$ is a column of ones corresponding to the intercept α. For simplicity, assume that all covariates have been centered so that $\mathbf{1}$ and $\mathbf{X}_\gamma$ are orthogonal. Also assume that the common choice $\pi(\alpha) = 1$ is made for the parameter α in each model (see Berger, Pericchi, and Varshavsky (1998) for a justification of this choice of prior).

Often all models will have small posterior probability, in which case more useful summaries of the posterior distribution are quantities such as the posterior inclusion probabilities of the individual variables:

$$w_i = \Pr(\gamma_i \neq 0 \mid \mathbf{Y}) = \sum_{\gamma} 1_{\gamma_i=1} \cdot p(H_\gamma \mid \mathbf{Y}). \tag{14.2.4}$$

These quantities also define the median-probability model, which is the model that includes those covariates having posterior inclusion probability of at least 1/2 (Barbieri and Berger, 2004).

14.2.1.2 Choosing Priors for Variable Selection

An extensive body of literature confronts the difficulties of Bayesian model choice in the face of weak prior information. These difficulties arise due to the obvious dependence of the marginal likelihoods in (14.2.2) upon the choice of priors for model-specific parameters. In general one cannot use improper priors on these parameters, since this leaves the resulting Bayes factors defined only up to an arbitrary multiplicative constant.

One class of methods for dealing with this issue involves training a noninformative prior using some function of the data, and then using the remaining data to compute Bayes factors under the induced family of prior distributions. This class includes the fractional Bayes factors of O'Hagan (1995) and the intrinsic Bayes factors of Berger and Pericchi (1996a). An extensive discussion can be found in Berger and Pericchi (2001). Another promising recent method due to Ray et al. (2007) known as PBIC offers a default specification in terms of the principal components of the observed information matrix, and seems to offer an interesting alternative to the well known Bayesian information criterion (Schwarz, 1978), which can also be used to compute a set of pseudo-marginal likelihoods.

Other authors have sidestepped this problem by defining default proper priors that are appropriate for model selection and that explicitly aim to minimize the effect of the prior. One such example is the g-prior and its robust variants, where

$$(\beta_\gamma \mid g, \phi) \sim N\left\{0, \frac{g}{\phi}(X_\gamma^t X_\gamma)^{-1}\right\},$$

and where g is either chosen outright, given a prior, or estimated by marginal maximum likelihood.

The existence of simple expressions for marginal likelihoods has made the use of g-priors very popular. They can also be defended on foundational Bayesian grounds, since they automatically adjust the predictive distribution of a model to account for observed co-linearities in the variables (precisely the kind of behavior one would expect from a carefully done subjective elicitation).

Additionally, some recent authors have overcome one of the major problems of g-priors—namely, a type of unsettling behavior known as the "information paradox,"

a phenomenon that was noted by Jeffreys (1961) in the context of testing normal means. It turns out that robustifying the *g*-prior by giving it heavier-than-normal tails seems to solve this problem. Moreover, it does so in a way that does not make marginal likelihoods all that much more difficult to compute. Key references here are Zellner and Siow (1980), Zellner (1986), George and Foster (2000), and Liang et al. (2008). Another overview of *g*-type priors can be found in the appendix of Scott and Berger (2008).

What of the prior probabilities for models themselves? One might reasonably consider a set of subjective prior model probabilities in smaller problems. But the complexity of such an elicitation means that default methods must be developed as a practical matter for high-dimensional problems, or when the appearance of objectivity is important. In such cases, there seems to be wide agreement surrounding the use of so-called "variable selection priors," where the the p-dimensional vector γ is assumed to arise as a sequence of exchangeable Bernoulli trials with common success probability w.

In such cases, we find it natural to think of specifying prior model probabilities as an opportunity to apportion mass across model space in a way that solves the implicit problem of multiple hypothesis testing posed by variable selection. The key intuition when using these priors is to let the data estimate w. This yields an automatic penalty for multiple testing, in that the introduction of spurious covariates will cause the posterior mass of w to concentrate near 0, making it harder for all variables to overcome the increasingly strong prior belief in their irrelevance (Scott and Berger, 2006). George and Foster (2000) propose estimating w by empirical-Bayes methods, but they note that this can prove computationally overwhelming in large problems with nonorthogonal design. Cui and George (2008) consider the fully Bayesian specification, whereby w is marginalized away before computing the posterior probability of a model. Finally, Scott and Berger (2008) offer some theoretical and numerical comparisons of the empirical-Bayes and fully Bayes approaches. They show that the fully Bayes solution offers an automatic improvement over empirical Bayes, in that it can avoid a particular form of degeneracy that arises when the empirical-Bayes solution collapses to the boundary of the parameter space.

14.2.2 Algorithms for Variable Selection and Shrinkage

14.2.2.1 Traditional MCMC

Computational algorithms for variable selection took flight beginning with the seminal work of George and McCulloch (1993) and followed by, among others, Geweke (1996), Clyde, Desimone, and Parmigiani (1996), and George and McCulloch (1997). These algorithms construct a Markov chain to simulate a sequence $\gamma^{(1)}, \gamma^{(2)}, \ldots, \gamma^{(T)}$ such that

$$\gamma^{(t)} \xrightarrow{\mathcal{D}} p(H_\gamma \mid \mathbf{Y})$$

as $t \to \infty$. The majority of these algorithms assume conjugate prior distributions to implement a Gibbs sampler (Gelfand and Smith, 1990) over the model space, since these allow marginal likelihoods to be computed in closed form. Several algorithms, however, allow non-conjugate priors to be used by employing Metropolis proposals (see, e.g., Madigan and York, 1995). In all cases, the inclusion probabilities (14.2.4) are estimated using the simulated γ sequence, with,

$$\hat{w}_i = \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}_{\gamma_i^{(t)}=1}. \tag{14.2.5}$$

An eloquent overview of these methods is given in Clyde and George (2004). Due to their intuitive construction and ease of implementation, MCMC techniques for variable selection surged in popularity during the 1990s and at the turn of the century.

In recent years, however, variable-selection techniques based upon traditional Markov-chain algorithms have come under scrutiny for a few reasons. First, for large p, the posterior distribution $p(H_\gamma \mid \mathbf{Y})$ is highly multimodal, and there are no trustworthy diagnostics that can effectively recognize a lack of Markov-chain convergence in such complex situations. The usual "rules of thumb" for MCMC, as we shall see on the examples later in the paper, can lead one badly astray when assessing convergence.

Second, many years of testing and implementation of such algorithms have shown that for a finite (and computationally practical) run time T, the chain often completely misses large modes in the model space. This potentially renders (14.2.5) a poor estimator of w_i; see Scott and Carvalho (2008) for particularly stark example of this kind, along with the examples in subsequent sections. For many researchers that have studied this issue, it is very difficult to understand how a procedure can correctly estimate marginal distributions when it misses such large, obvious modes of the joint distribution. Even now, despite years of research on computational approaches for variable selection, it is simply not known whether the estimated inclusion probabilities that arise from MCMC on large problems are even approximately correct.

Third, it is very unlikely that the Markov chain will visit any model frequently enough to allow model probabilities to be estimated by frequency of occurrence in the Monte Carlo sample. In fact, in large problems, it will almost always be the case that all models (even the best one) will have posterior probabilities significantly smaller than $1/T$, which is the smallest nonzero model probability that can arise from an MCMC of length T.

14.2.2.2 Stochastic Search Algorithms for Variable Selection

For these and other reasons, some researchers have become skeptical of "vanilla" MCMC, and the popularity of these techniques as an active research area has dwindled. These older techniques have, however, paved the way for the emergence of

newer stochastic-search (SS) algorithms, which focus on rapidly discovering models with high posterior probability. These algorithms use the information from previously visited models, such as estimated inclusion probabilities, to guide the search over the model space.

MCMC can, of course, be viewed as a form of stochastic search. But the SS algorithms discussed here pay little, if any, attention to the goal of converging to the posterior distribution $p(H_\gamma \mid \mathbf{Y})$. Rather, SS algorithms focus on finding the models with the highest posterior probability. Their output is simply a list of models visited, together with a score for each one—typically an un-normalized posterior probability. There is no sense which the estimated inclusion probabilities "converge" to the true ones, unless all of the models are eventually enumerated.

A simple SS algorithm proposed by Berger and Molina (2005), for example, uses online estimates of posterior model and inclusion probabilities to orient the search. Let $p^{(t)}(H_\gamma \mid \mathbf{Y})$ and $w_i^{(t)}$ be the estimates of $p(H_\gamma \mid \mathbf{Y})$ and w_i at the t^{th} iteration of the SS algorithm, respectively. At iteration t, the algorithm proceeds by:

1. Resampling one of the $t-1$ previously sampled models in proportion to their estimated probabilities $p^{(t)}(H_\gamma \mid \mathbf{Y})$, and setting this to be the current model.
2. Flipping a coin to decide whether to add or delete a variable to the current model.
3. Adding [removing] variable i with probability $(w_i^{(t)} + \varepsilon)/(1 - w_i^{(t)} + \varepsilon)$ [or $(1 - w_i^{(t)} + \varepsilon)/(w_i^{(t)} + \varepsilon)$ in the case of a deletion], where $\varepsilon > 0$ is small and bounds $w_i^{(t)}$ away from 0 or 1.

Berger and Molina (2005) suggest updating $p^{(t)}(H_\gamma \mid \mathbf{Y})$ via a path-based estimate of the Bayes factors between models. Certainly the algorithm cannot explore all 2^p models, but the hope is that a majority of visited models will have high posterior probability.

Many stochastic-search algorithms have this general flavor. The key ingredient to visiting good models seems to be to use the inclusion probabilities to guide the search—an approach that also works in far more general classes of models and features. For example, Scott and Carvalho (2008) propose a SS algorithm called FINCS (feature-inclusion stochastic search). This algorithm, which builds upon the insight of Berger and Molina (2005) regarding the importance of the inclusion probabilities, interweaves local moves (adding or deleting a variable from $\gamma^{(t)}$), resampling moves (selection from among one of $\gamma^{(1)}, \ldots, \gamma^{(t-1)}$), and global moves that attempt to avoid getting stuck in local modes in model space. Their application of the algorithm is to Gaussian graphical models, but the approach is in principle quite straightforward to use in linear models, as well.

A SS algorithm of a slightly different nature was proposed by Hans, Dobra, and West (2007), and is known as shotgun stochastic search (SSS). Hans, Dobra, and West (2007) consider constructing a neighborhood of models around $\gamma^{(t)}$ denoted by $\partial\gamma^{(t)} = \{\gamma_+^{(t)}, \gamma_0^{(t)}, \gamma_-^{(t)}\}$, where $\gamma_+^{(t)}, \gamma_0^{(t)}, \gamma_-^{(t)}$ is the set of all models which add, replace, or remove one element from $\gamma^{(t)}$, respectively. Each model in $\partial\gamma^{(t)}$ is given a "score" (e.g., AIC, BIC, or a posterior probability), and a set $\mathscr{S}^{(t)}$ is adapted to contain the B highest scoring models of $\{\partial\gamma^{(t)}, \mathscr{S}^{(t-1)}\}$ such that $\mathscr{S}^{(T)}$ contains

the B best models after T iterations. To iterate the algorithm, $\gamma^{(t+1)}$ is sampled from $\partial\gamma^{(t)}$ proportional to the assigned scores. (Obviously, the SSS algorithm is computationally demanding and works best in a parallel computing environment, which it was designed to exploit.)

Clyde, Ghosh, and Littman, M. (2009), astutely observe that for models with tractable marginal likelihoods, resampling a model provides no additional information for estimating posterior model probabilities. They go on to develop a Bayesian adaptive sampling (BAS) algorithm which samples without replacement from the 2^p models. This too is accomplished by sampling models one variable at a time in a manner that is guided by the estimated inclusion probabilities. If $\mathscr{S}^{(t)}$ is the set of sampled models, then the estimated inclusion probability for the i^{th} variable is given by

$$\hat{w}_i^{(t)} = \frac{\sum_{\gamma\in\mathscr{S}^{(t)}} p(\mathbf{Y} \mid H\gamma)\gamma_i}{\sum_{\gamma\in\mathscr{S}^{(t)}} p(\mathbf{Y} \mid H\gamma)}, \tag{14.2.6}$$

and $\hat{w}_i^{(t)} \to w_i$ as $t \to 2^p$ because $\mathscr{S}^{(t)}$ becomes the set of all 2^p models. Sampling without replacement is ensured by subtracting the mass of model $\gamma^{(t)}$ from the total mass of $\pi(\gamma \mid \mathbf{Y})$.

14.2.2.3 Adaptive MCMC Algorithms for Variable Selection

Similar to SS algorithms, the key idea of adaptive MCMC (AMCMC) is to inform and adapt the proposal distribution of a Metropolis-Hastings algorithm using past draws. Specifically, if $X^{(1:t)} = \{X^{(i)} : i = 1\ldots,t\}$ is the set of realizations of the Markov chain $X^{(t)}$ up to time up time t, then AMCMC would adapt the proposal distribution $q(X^{(t)},\cdot;\psi^{(t)})$ iteratively by adapting the parameter vector $\psi^{(t)} = f(X^{(1:t)})$ for some function f.

As a simple example, suppose that the proposal density is given by $q(X^{(t)},\cdot;\psi^{(t)})$ $= N(\cdot;X^{(t)},\sigma^{(t)})$. Then an AMCMC algorithm could adapt $\psi^{(t)} = \sigma^{(t)}$ iteratively via the update equation $\psi^{(t)} = \sqrt{Var(X^{(1:t)})}$. While this is a simple example, it illustrates the appeal of AMCMC in that the tuning of the proposal distribution is done automatically.

Because AMCMC algorithms use all past states $X^{(1:t)}$ to construct the proposal distribution (i.e., estimate $\psi^{(t)}$), the resulting algorithms no longer satisfy the Markov property: the past and future are no longer conditionally independent, given the present. Nevertheless, due to the recent theoretical work of, for example, Haario, Saksman, and Tamminen (2001), Atchadé and Rosenthal (2005), Andrieu and Moulines (2006), Andrieu and Atchadé (2007), Roberts and Rosenthal (2007), and Atchadé et al. (2009), simple and intuitive conditions have been established which, if met, guarantee that an AMCMC algorithm will converge to the desired posterior distribution. Using these conditions, practically useful algorithms have emerged for a variety of models and situations. These include Haario, Saksman, and Tamminen (2001, 2005), Roberts and Rosenthal (2009), Pasarica and Gelman (2010), and Craiu, Rosenthal, and Yang (2009).

Recently, some AMCMC methods for variable selection have begun to emerge. One of the first was proposed by Nott and Kohn (2005), who made clever use of the fact that $Pr(\gamma_i = 1 \mid \gamma_{iC}, \mathbf{Y}) = \mathbb{E}(\gamma_i \mid \gamma_{iC}, \mathbf{Y})$, where $\gamma_{iC} = \{\gamma_j \in \gamma : j \neq i\}$. Specifically, Nott and Kohn (2005) adaptively estimate $\bar{\gamma}^{(t)} = t^{-1} \sum_i \gamma^{(i)}$ and $\Gamma^{(t)} = Cov(\gamma \mid \mathbf{Y})$ at each step of the AMCMC algorithm. They do so using the best linear unbiased estimator of $\mathbb{E}(\gamma_i \mid \gamma_{iC}, \mathbf{Y})$,

$$\hat{\mathbb{E}}(\gamma_i \mid \gamma_{iC}, \mathbf{Y}) = \bar{\gamma}_i + \Gamma^{(t)}_{i,iC} \left[\Gamma^{(t)}_{iC,iC} \right]^{-1} (\gamma_{iC} - \bar{\gamma}_{iC}), \tag{14.2.7}$$

as a proposal distribution in the MCMC algorithm, where $\Gamma^{(t)}_{i,iC}$ is the i^{th} row of $\Gamma^{(t)}$ with the i^{th} column removed, and $\Gamma^{(t)}_{iC,iC}$ is $\Gamma^{(t)}$ with the i^{th} row and column removed. Using (14.2.7) as a proposal distribution, variables with a high *conditional* inclusion probabilities are frequently added to the model.

The algorithm of Nott and Kohn (2005) uses conjugate prior distributions such that the coefficients and error precision can be integrated out, allowing a closed-form expression for $f(\mathbf{Y} \mid H_\gamma)$. An alternative AMCMC algorithm proposed by Ji and Schmidler (2009) use a point-mass mixture prior for the coefficients, e.g.

$$p(\beta_i) = (1 - w)\delta_0(\beta_i) + wN(\beta_i; 0, s_i^2), \tag{14.2.8}$$

where $\delta_0(\cdot)$ is the Dirac measure at 0 and s_i^2 is a known prior variance. Ji and Schmidler (2009) then consider adapting proposal distributions of the form,

$$q(\beta_i^{(t)}, \cdot; \psi^{(t)}) = \lambda q_0(\cdot; \tilde{\psi}) + (1 - \lambda) \left[(1 - \omega^{(t)})\delta_0(\cdot) + \omega^{(t)} N(\cdot; \hat{\beta}_i^{(t)}, \hat{\Sigma}_i^{(t)}) \right], \tag{14.2.9}$$

where $0 < \lambda < 1$ is fixed and known, and $q_0(\cdot; \tilde{\psi})$ is fixed (non-adaptive) to ensure that the bounded convergence condition in Theorem 13 of Roberts and Rosenthal (2007) is satisfied. Ji and Schmidler (2009) then use the stochastic approximation algorithm of Robbins and Monro (1951) to develop an adaptive scheme for $\omega^{(t)}$, $\hat{\beta}_j^{(t)}$, and $\widehat{\Sigma}^{(t)}$ which minimizes the Kullback-Leibler (KL)-divergence between the target distribution $\pi(\beta_j \mid \mathbf{Y})$ and the proposal distribution (14.2.9).

One potential limitation of this algorithm is that it depends upon being able to write an exchangeable joint prior for the regression coefficients in a given model, where $\beta_i \sim w \cdot p(\beta_i) + (1 - w) \cdot \delta_0$ as in (14.2.8). This restriction excludes the possibility of using g-like priors, since these cannot be expressed using an exchangeable model for each coefficient. Since there are strong (non-computational) reasons to prefer g-like priors for variable selection in situations with non-orthogonal designs, this limitation may be a significant one.

14.2.2.4 Shrinkage-based Alternatives

All of the models discussed so far place nonzero probability mass upon the hypothesis that each coefficient β_i is zero. As we have seen, this results in a combinatorial explosion in the number of discrete models that must be considered, leading to a very difficult problem in stochastic computation.

Recently, many researchers have become interested in an alternative approach based upon "pure shrinkage" priors, which do not place positive probability at zero. There is, of course, an established tradition of evaluating such priors on foundational Bayesian grounds (see, e.g., Pericchi and Smith, 1992). Yet much of the more recent activity has arisen as a Bayesian rejoinder to the neo-classical literature on penalized least squares, which offers a very different perspective on variable selection. Indeed, many well-studied priors are enjoying newfound prosperity in their second careers as "penalty functions," which yield solutions that can be interpreted as posterior modes.

These priors are often in the family of multivariate scale mixtures of normals, which is very general and has many nice analytical properties:

$$\begin{aligned}(Y \mid \beta, \sigma^2) &\sim N(X\beta, \sigma^2 I) \\ (\beta_j \mid \lambda_j, \tau, \sigma^2) &\sim N(0, \lambda_j^2 \tau^2 \sigma^2) \\ \lambda_j &\sim g(\lambda_j) \\ \tau &\sim h(\tau)\,.\end{aligned}$$

The λ_j's are known as the "local" shrinkage parameters, while τ is known as the "global" shrinkage parameter.

The following list is by no means comprehensive, but gives a sense of the strong level of activity in this area:

1. The horseshoe prior of Carvalho, Polson, and Scott (2008) assumes a half-Cauchy prior on the local scales, $\lambda_j \sim \mathrm{C}^+(0,1)$, which is equivalent to an $\mathrm{F}(1,1)$ prior on the local variances λ_j^2. Polson and Scott (2009) generalize this prior to a wider class of hypergeometric–beta mixtures, while Scott (2009) proposes two methods for fitting models in this family: one based on importance sampling, and an alternative MCMC algorithm that involves a slice-sampling step for the local shrinkage parameters.
2. The Student-t prior is defined by an inverse-gamma mixing density, $\lambda_j^2 \sim \mathrm{IG}(\xi/2, \xi\tau^2/2)$. Tipping (2001) uses this model for sparsity by finding posterior modes under the assumption that $\xi \to 0$.
3. The double-exponential prior uses an exponential mixing density: $p(\lambda_j^2 \mid \tau^2) \propto \exp\{\lambda_j^2/2\tau^2\}$. The standard Markov-chain Monte Carlo algorithm for working with this model is from Carlin and Polson (1991b), and uses the fact that the local variance parameters are conditionally inverse-Gaussian, given the data and other parameters. More recently, Park and Casella (2008) and Hans (2008) have revitalized interest in this prior as a Bayesian alternative to the LASSO (Tibshirani, 1996).

4. The normal–Jeffreys prior has been studied by Figueiredo (2003) and Bae and Mallick (2004). This improper prior is induced by placing Jeffreys' prior upon each variance term, $p(\lambda_j^2) \propto 1/\lambda_j^2$, leading to $p(\beta_j) \propto |\beta_j|^{-1}$ independently.
5. The normal–exponential–gamma family of priors proposed by Griffin and Brown (2005) is also based upon the exponential mixing density, but uses a $\text{Ga}(c, d^2)$ density rather than an inverse-gamma for the global scale term τ. The two hyperparameters allow control over tail weight (c) and scale (d). This leads to

$$p(\lambda_j^2) = \frac{c}{d^2}\left(1 + \frac{\lambda_j^2}{d^2}\right)^{-(c-1)}.$$

Clearly, options abound. A discussion of some general principles to help guide this choice can be found in Carvalho, Polson, and Scott (2008), who compare many of the above possibilities at great length. Their conclusion is that, in order to be appropriate for sparse problems, the prior for λ_j should have positive density at zero, and should decay no faster than λ_j^{-2}. (These same guidelines also apply to the prior on τ.)

To be sure, pure-shrinkage solutions can never provide a truly sparse solution, in the sense that they will never allocate positive posterior probability at zero. Nonetheless, there is a growing body of empirical evidence to suggest that it *is* possible to use pure-shrinkage priors to get estimates and predictions very close to those that arise under Bayesian model averaging. This is an active and fast-moving area of research, and the exciting possibility of "BMA mimicry" using shrinkage priors is just one of many open problems here.

14.2.3 Examples

In this section, the approaches described above are evaluated on three examples: a very simple, simulated orthogonal problem; a data set on the long-term economic growth rates of 88 countries; and a data set of daily maximum ozone measurements near Los Angeles. We address four questions that are relevant to comparing MCMC and stochastic search/adaptive sampling, the two general classes of variable-selection algorithms that have been considered here:

1. Does either class of methods systematically find better models?
2. Do the classes systematically differ in their estimates of inclusion probabilities?
3. Does either class yield better out-of-sample performance?
4. How do pure-shrinkage solutions compare to full-blown model averaging?

14.2.3.1 Orthogonal Simulation Study

Our first experiment is designed to test the algorithms in a situation where the model space is too large to enumerate, but where everything else remains as simple as possible. Hence we construct an orthogonal problem with no unknown hyperparameters, where all true inclusion probabilities are known exactly, and where the identity of the top model is known.

Specifically, we let

$$Y_i \overset{iid}{\sim} N(\mu_i \gamma_i, \sigma^2) \tag{14.2.10}$$

for $i = 1, \ldots, n$. Here γ_i is either 1 or 0, designating signal or noise. We chose $n = 50$ and $\sigma^2 = 1$, and we assume that the nonzero means follow $\mu_i \sim N(0,1)$, and that $Pr(\gamma_i = 1) = 0.5$ independently for all i. Even though the full model space has 2^{50} members and is too big to enumerate, the structure of the problem allows inclusion probabilities to be computed exactly. The marginal likelihood of the data under a model configuration γ is

$$f(\mathbf{Y} \mid H\gamma) = \prod_i N(Y_i \mid 0, 1+\gamma_i),$$

where $N(x \mid m, v)$ is the normal p.d.f. with mean m and variance v evaluated at x. Meanwhile, under this simple design, the true inclusion probabilities are

$$w_i = \frac{N(Y_i \mid 0,2)}{N(Y_i \mid 0,1) + N(Y_i \mid 0,2)}, \tag{14.2.11}$$

and the highest posterior probability model is the median probability model.

We actually simulated three data sets under (14.2.10) with low, medium, and high signal-to-noise (STN) ratios. The low-STN data set takes $\mu_i = i$ for $i = 1, \ldots, 5$; the medium STN data set takes $\mu_i = i/2$ for $i = 1, \ldots, 10$; and the high STN data set takes $\mu_i = i/5$ for $i = 1, \ldots, 25$. (All other means are set to zero.)

TABLE 14.2. Sum of absolute error in inclusion probabilities for orthogonal simulation study.

Data	SSVS	FINCS	AMCMC
Low Density	0.284	12.213	0.384
Medium Density	0.221	10.135	0.394
High Density	0.208	8.391	0.372

For each simulated data set, we attempted to reconstruct the posterior distribution for γ using the stochastic-search algorithm of Berger and Molina (2005) (FINCS), the AMCMC algorithm of Nott and Kohn (2005), and the SSVS algorithm of George and McCulloch (1993). Each MCMC was run for $T = 5000$ iterations after discarding an initial 500 iterations for burn-in, while FINCS was run for 250,000 iterations. (These numbers mean that each algorithm evaluated the same number of marginal likelihoods, making the comparison a fair one.) Table 14.2 displays the sum of ab-

solute errors for inclusion probabilities $SAE_w = \sum_i |w_i - \hat{w}_i|$ for the three data sets, and Figure 14.6 displays corresponding boxplots of $\log f(\mathbf{Y} \mid \gamma)$ of the top visited models.

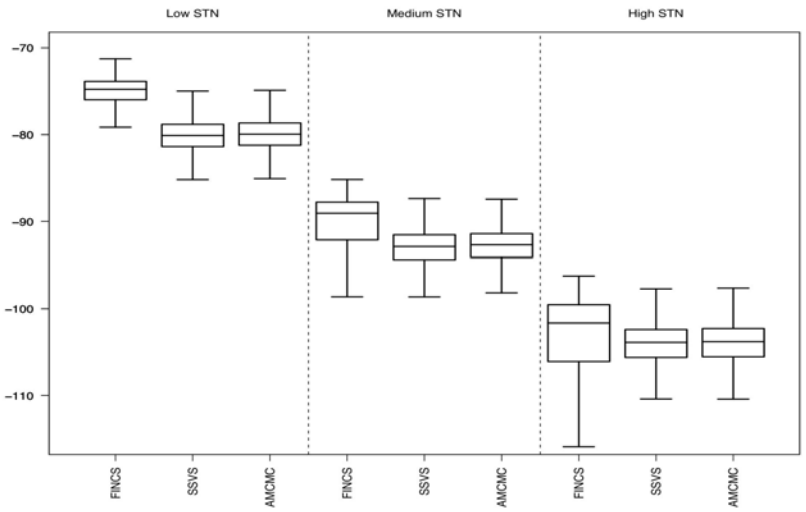


FIGURE 14.6. $\log f(\mathbf{Y} \mid H\gamma)$ of the explored models using FINCS, SSVS, and AMCMC for the low, medium, and high STN data sets.

Additionally, the full set of inclusion probabilities for the medium STN experiment can be found in Table 14.3. The table is a bit dense but repays close inspection, since together with Figure 14.6 it tells a very interesting story. On the one hand, the FINCS algorithm is quite poor at estimating inclusion probabilities compared to AMCMC or SSVS, even on this simple orthogonal problem. In particular, it seems to overestimate w_i for "good" variables, and to underestimate w_i for "bad" variables. (This was also true for the low- and high-STN data sets, though these tables are omitted.) This systematic bias is interesting but perhaps not too surprising: FINCS is not concerned with exploring all models, nor with re-constructing any marginal distributions.

Meanwhile, both SSVS and AMCMC get the inclusion probabilities essentially correct. Yet paradoxically, the explored models under FINCS have a higher marginal likelihood than those found under either AMCMC or FINCS. Indeed, FINCS finds dozens of models that are better than the single best one discovered by either SSVS or AMCMC. This fact is much harder to understand: how is it that, at least in this case, both MCMC methods are able to reconstruct the correct marginal distributions while missing large pockets of probability in the joint distribution from which all these marginals are derived?

TABLE 14.3. True inclusion probabilities for the 50 simulated coefficients in the medium signal-to-noise-ratio configuration, along with estimates arising from three algorithms: stochastic search using inclusion probabilities (FINCS), Gibbs sampling over models (SSVS), and adaptive Markov-chain Monte Carlo (AMCMC). The results are rounded to two decimal places and are ordered by the absolute value of the observation Y_i.

Rank	Y	True w_i	FINCS	SSVS	AMCMC
1	5.98	1.00	1.00	1.00	1.00
2	4.94	1.00	1.00	1.00	1.00
3	4.30	1.00	1.00	1.00	1.00
4	3.80	0.99	1.00	0.99	0.99
5	3.38	0.97	1.00	0.97	0.97
6	3.10	0.95	1.00	0.94	0.95
7	2.75	0.90	0.99	0.90	0.90
8	-2.71	0.89	0.99	0.89	0.89
9	-2.14	0.74	0.95	0.75	0.74
10	-1.87	0.65	0.92	0.65	0.66
11	1.67	0.59	0.86	0.59	0.59
12	-1.63	0.58	0.75	0.57	0.58
13	-1.60	0.57	0.74	0.57	0.58
14	1.57	0.56	0.75	0.55	0.55
15	-1.55	0.55	0.69	0.55	0.55
16	1.44	0.52	0.55	0.52	0.53
17	1.42	0.52	0.69	0.51	0.52
18	-1.29	0.48	0.36	0.48	0.49
19	1.28	0.48	0.70	0.48	0.49
20	1.26	0.48	0.49	0.48	0.48
21	1.23	0.47	0.31	0.46	0.48
22	-1.23	0.47	0.43	0.46	0.48
23	1.13	0.45	0.22	0.44	0.44
24	1.03	0.43	0.14	0.42	0.42
25	-0.85	0.40	0.18	0.40	0.39
26	0.74	0.38	0.09	0.39	0.36
27	0.71	0.38	0.08	0.38	0.36
28	-0.69	0.37	0.08	0.38	0.37
29	-0.66	0.37	0.08	0.37	0.37
30	0.65	0.37	0.07	0.36	0.36
31	0.61	0.36	0.09	0.36	0.35
32	0.59	0.36	0.08	0.37	0.35
33	0.58	0.36	0.07	0.35	0.34
34	0.57	0.36	0.08	0.36	0.35
35	-0.55	0.36	0.16	0.35	0.35
36	0.50	0.36	0.06	0.36	0.35
37	-0.50	0.36	0.08	0.36	0.37
38	-0.46	0.35	0.07	0.34	0.35
39	0.42	0.35	0.10	0.36	0.34
40	-0.36	0.34	0.07	0.35	0.33
41	-0.33	0.34	0.06	0.34	0.34
42	0.26	0.34	0.06	0.33	0.32
43	0.22	0.34	0.07	0.34	0.33
44	-0.21	0.34	0.06	0.33	0.32
45	0.19	0.34	0.06	0.34	0.32
46	-0.18	0.34	0.07	0.33	0.32
47	-0.14	0.34	0.06	0.34	0.34
48	-0.07	0.33	0.06	0.34	0.31
49	0.02	0.33	0.05	0.33	0.33
50	-0.01	0.33	0.06	0.33	0.32

14.2.3.2 GDP Growth Data

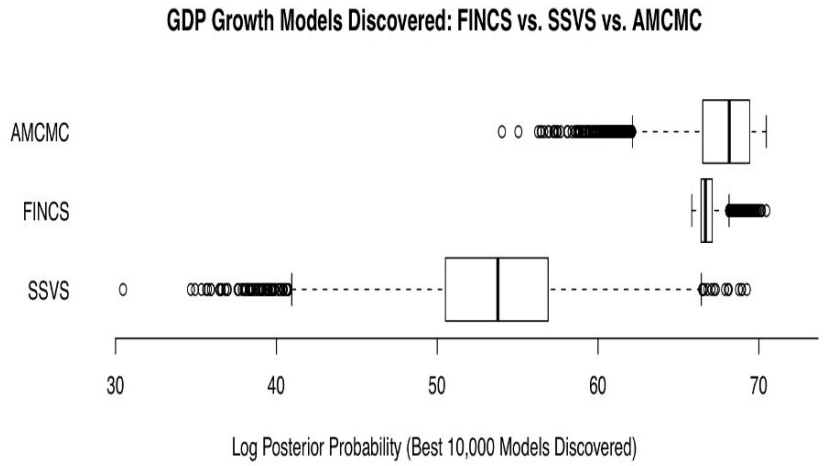


FIGURE 14.7. Log marginal likelihoods of models discovered by the three algorithms on the GDP growth example.

We next ran a similar experiment on a real data set that was collected in an attempt to understand the determinants of long-term economic growth. Here Y is annualized GDP growth since 1960 for 88 countries, and X represents a battery of 67 possible socio-economic, political, and geographical predictors of growth. This data set has been previously analyzed by Fernandez, Ley, and Steel (2001), Sala-i Martin, Doppelhofer, and Miller (2004), and Ley and Steel (2007). We assume g-priors for the coefficients; unlike in the orthogonal problem, the true inclusion probabilities are unknown.

Surprisingly, a very different pattern emerged. Before, SSVS and AMCMC agreed (both with each other and with the truth), while FINCS disagreed despite visiting better models. On this problem, however, FINCS and AMCMC tend to agree with each other—though not perfectly—while SSVS disagrees with both of them. As Table 14.4 shows, this disagreement can be stark. For example, SSVS estimates the inclusion probability of the East Asian dummy variable to be 50%, while neither of the other methods estimate this probability to be larger than 4%.

Given a sufficient burn-in period, both SSVS and AMCMC are fairly stable from run to run. (The burn period tends to be quite long for AMCMC, but not untenably so.) This creates the illusion that each has independently converged to the posterior distribution. Yet at least one of them certainly has not, and it is impossible to know which one it is using existing tools.

TABLE 14.4. Estimated inclusion probabilities for the top 50 (out of 67) variables in the GDP growth data set. Results are given for SSVS, FINCS, and AMCMC. AMCMC was replicated three times to ensure stability, with results displayed for all three runs.

Variable	SSVS	FINCS	AMCMC 1	AMCMC 2	AMCMC 3
Investment Price	0.98	1.00	1.00	1.00	1.00
GDP in 1960 (log)	0.97	1.00	1.00	1.00	1.00
Primary Schooling in 1960	0.94	0.99	1.00	1.00	1.00
Fraction Confucian	0.73	1.00	1.00	1.00	0.99
Fraction GDP in Mining	0.72	1.00	0.99	0.99	0.99
Public Investment Share	0.64	0.99	0.98	0.98	0.98
African Dummy	0.55	1.00	0.96	0.97	0.98
Fraction Buddhist	0.52	0.93	0.94	0.96	0.95
East Asian Dummy	0.50	0.02	0.04	0.03	0.03
Fraction Speaking Foreign Language	0.48	0.85	0.83	0.81	0.84
Life Expectancy in 1960	0.47	0.51	0.59	0.56	0.56
Fraction Muslim	0.43	0.12	0.16	0.14	0.13
Fraction of Tropical Area	0.41	0.13	0.13	0.16	0.12
Latin American Dummy	0.41	1.00	0.96	0.96	0.98
Population Density Coastal in 1960s	0.39	0.09	0.11	0.12	0.10
Population Density 1960	0.37	0.05	0.03	0.03	0.03
Real Exchange Rate Distortions	0.34	0.07	0.07	0.06	0.05
Nominal Gov. GDP Share 1960s	0.33	0.10	0.09	0.09	0.07
Gov. Consumption Share 1960s	0.31	0.30	0.24	0.35	0.29
Real Gov. GDP Share in 1960s	0.30	0.54	0.54	0.42	0.52
Revolutions and Coups	0.28	0.40	0.26	0.29	0.31
Fraction Catholic	0.27	0.07	0.04	0.05	0.04
Openess measure 1965-74	0.27	0.10	0.07	0.08	0.06
Fertility in 1960s	0.25	0.14	0.09	0.07	0.09
Hydrocarbon Deposits in 1993	0.24	0.04	0.02	0.03	0.02
Fraction Hindus	0.24	0.04	0.05	0.04	0.04
European Dummy	0.22	0.03	0.04	0.03	0.02
Ethnolinguistic Fractionalization	0.21	0.02	0.01	0.01	0.02
Outward Orientation	0.20	0.17	0.08	0.07	0.09
Fraction Protestants	0.20	0.02	0.01	0.01	0.01
Spanish Colony	0.02	0.03	0.03	0.03	0.02
Fraction Population In Tropics	0.20	0.06	0.05	0.04	0.04
Political Rights	0.20	0.02	0.01	0.01	0.01
Civil Liberties	0.20	0.04	0.02	0.02	0.02
Years Open 1950-94	0.20	0.02	0.01	0.01	0.01
Primary Exports 1970	0.20	0.05	0.03	0.03	0.03
Fraction Population Over 65	0.19	0.03	0.03	0.03	0.02
Colony Dummy	0.17	0.02	0.01	0.01	0.01
Air Distance to Big Cities	0.17	0.02	0.01	0.01	0.01
Higher Education 1960	0.17	0.02	0.01	0.01	0.01
Education Spending Share, 1960s	0.17	0.05	0.02	0.02	0.02
Socialist Dummy	0.17	0.06	0.02	0.03	0.03
Malaria Prevalence in 1960s	0.17	0.05	0.03	0.03	0.03
Capitalism	0.16	0.04	0.02	0.02	0.02
Population in 1960	0.16	0.03	0.01	0.01	0.01
Absolute Latitude	0.16	0.02	0.01	0.01	0.01
Fraction Land Near Navigable Water	0.16	0.03	0.01	0.02	0.01
Fraction Population Less than 15	0.16	0.03	0.01	0.01	0.01
British Colony Dummy	0.16	0.03	0.01	0.01	0.01
Landlocked Country Dummy	0.14	0.02	0.01	0.01	0.01

A final fact worth noting is that, as before, SSVS fails to visit many high-probability models (Figure 14.7). Indeed, the cumulative posterior probability of all models discovered by SSVS is only 0.6% that of the top 10,000 models visited by FINCS.

14.2.3.3 Ozone Data and Out-of-Sample Performance

The ozone data set consists of $n = 178$ daily measurements of the maximum ozone concentration near Los Angeles. This data set has become a standard benchmark in the regression literature, and has been recently analyzed by, among others, Casella and Moreno (2005), Berger and Molina (2005), and Liang et al. (2008). For this study, 10 atmospheric predictor variables are considered (see Casella and Moreno (2005) for a description), along with all squared terms and all 45 second-order interactions. This yields $p = 65$ potential variables that could be included in the model. Enumerating all 2^{65} models is impossible—to store all of the binary vectors on a computer would require 300 million terabytes of memory.

For this study, we performed 100 different "train–test" splits of the data set: a random sample of 134 data points were used the fit the model, with the remaining 44 used to compare out of sample predictive performance. The test subjects were: the AMCMC algorithm of Nott and Kohn (2005), the SSVS algorithm of George and McCulloch (1993), the BAS algorithm of Clyde, Ghosh, and Littman, M. (2009), the horseshoe (HS) method of Carvalho, Polson, and Scott (2008), the Bayesian lasso, and the classical lasso. Zellner's g-prior was used with $g = n$ for the AMCMC, BAS, and SSVS algorithms.

Figure 14.8 displays box plots of the sum of predictive squared errors for the 100 repetitions,

$$SPSE = \sum_{i \in \mathscr{V}} (Y_i - \hat{Y}_i)^2,$$

where $\mathscr{V}$ is set of indices of the 44 data points in the test data set, and $\hat{Y}_i$ is one of either the model-averaged estimate of Y_i when using AMCMC, BAS, and SSVS; the posterior predictive mean when using the Bayesian lasso and the horseshoe; or the posterior predictive mode when using the lasso.

Each of the methods performed very similarly in terms of prediction, with median SPSE's of 833.99 for AMCMC, 794.50 for SSVS, 847.27 for BAS, 837.51 for the Bayesian lasso, 821.56 for the HS, and 898.58 for the classical lasso. Clearly the between-sample variation is much larger than the between-method variation.

While the methods were similar at predicting, however, they differed greatly in average model size. Adaptive MCMC, with an average model size of 6.46, found the most parsimonious models; SSVS, with an average model size of 16.18, found the most complex models. Bayesian adaptive sampling and the classical lasso were intermediate, yielding average model sizes of 11.56 and 11.84, respectively. Thus, while each method is performing similarly in terms of SPSE, the models being chosen are quite different. It is particularly difficult to explain why this is the case for AMCMC and SSVS, since both algorithms are theoretically converging to the same

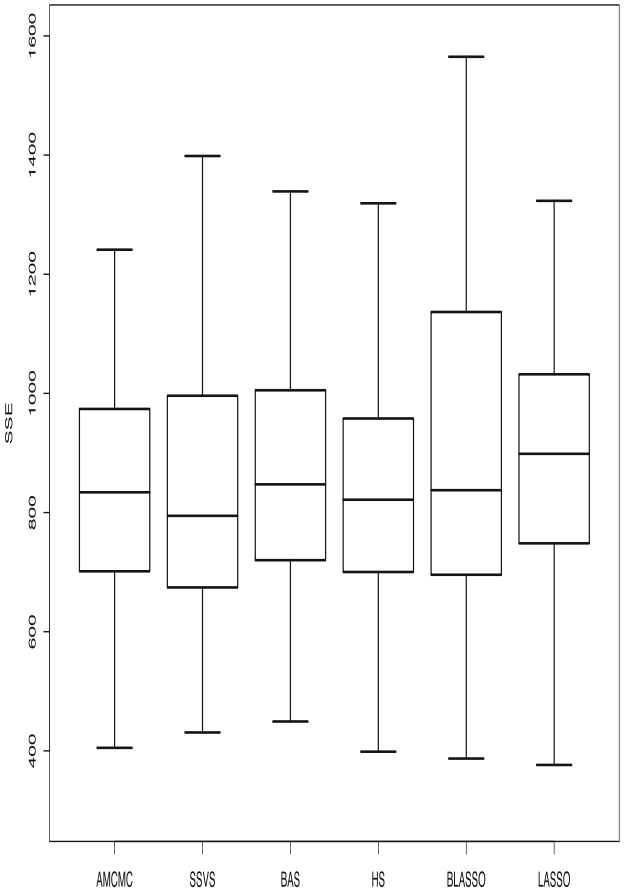


FIGURE 14.8. Box plots of sum of predictive squared errors for ozone cross-validation study.

posterior distribution. Also noteworthy is that the pure-shrinkage solutions are competitive with Bayesian model-averaging in terms of out-of-sample predictions. It is not at all clear, however, how one would select a model or construct a measure of variable importance under pure-shrinkage priors. Sparse solutions based on the posterior mode are clearly dubious from a Bayesian point of view; except for the very rare case of a true "0–1" loss function, there is no deeper justification for choosing *any* point in $\mathscr{R}^p$ with zero posterior probability, beyond a simple desire to induce sparsity.

14.2.4 Final Remarks

Where do these results leave those interested in fitting a Bayesian linear model? We cannot, unfortunately, give an unqualified recommendation for any algorithm. We can only point to specific areas in which they fail.

First, it is clear that if one's goal is to find models with high posterior probability, then stochastic search is preferred to either AMCMC or SSVS. This message emerges again and again from our simulation studies, and from those of other authors: MCMC in γ space simply misses too many good models.

Second, as a general matter, it remains unclear how one should compute posterior inclusion probabilities. It is true that Gibbs and AMCMC are able to reconstruct these probabilities in orthogonal settings. But no one knows whether this fidelity of reconstruction holds in high-dimensional, nonorthogonal problems. The experiment involving GDP-growth data suggests that it may not. Perhaps the best we can hope for at present is to estimate a set of *conditional* inclusion probabilities—conditioning, of course, on the set of models actually visited, and working hard to ensure that this set is a good one.

14.3 MCMC for Constrained Parameter and Sample Spaces

Merrill W. Liechty, John C. Liechty, and Peter Müller

Due to computational concerns, Bayesian methods are often used with models consisting of sampling distributions, priors, and hyperpriors of well-known distributions with known normalizing constants. In these cases the constants are trivial to find by simply recognizing the kernel of the density function when the distributions are conjugate.

The typical Bayesian analysis can be describe as above. However, even in these cases, sometimes the distributions of interest are truncated over a region that depends on (hyper-) parameters in the next level of the model. The normalization constant then becomes a function of these hyperparameters, and evaluation of the posterior distribution requires the evaluation of this normalization constant. This can severely impact the efficiency of Markov chain Monte Carlo methods (MCMC). In some instances, the dimension over which the normalizing constant is defined may be small enough that the researcher can use numerical integration at each step in the MCMC to evaluate the required normalization constant. These numerical methods can be quite computationally expensive, and although they may be feasible for small dimensions, as the dimension increases, they quickly become prohibitively expensive.

Consider as a classic example the positive definiteness constraint for a random covariance matrix θ, with a hyperprior, $p(\theta|\eta) \propto g(\theta|\eta) I\{\theta \in \mathscr{A}\}$, where $\mathscr{A}$ denotes the set of positive definite matrices and $g(\cdot)$ denotes an (unconstrained) prob-

ability distribution for θ. The constraint introduces a new normalization constant, $1/g(\mathscr{A}|\eta)$. The complete conditional posterior distribution for η includes the normalization constant $1/g(\mathscr{A}|\eta)$ that might require an analytically intractable integral. The same problem arises when there is a constraint on the sampling model $p(y|\theta)$.

There are several approaches that have been proposed to address this problem through specific applications. Yet few general methods are available. These include Chen and Schmeiser (1993) who propose the hit-and-run sampler as a generic posterior simulation method that is particularly suitable for constrained parameter spaces. See Chen, Shao, and Ibrahim (2000) for a review. Chib and Greenberg (1998) and McCulloch, Polson, and Rossi (2000) use a random walk Metropolis sampler to accommodate a constrained parameter space.

An alternative approach is the use of auxiliary variables, proposed by Møller et al. (2006), who introduce an auxiliary variable that has a well chosen distribution. When sampled jointly with the constrained parameter of interest, exact inference is achieved. Another method is approximate Bayesian computation (ABC), proposed in Marjoram et al. (2003) and Reeves and Pettitt (2005). This approach draws from the prior and posterior, respectively, and use these as a basis for generation of "data" from the model which leads to the acceptance/rejection of the proposed parameters. Gourieroux, Monfort, and Renault (1993) propose similar techniques for econometrics applications. Several exact and approximate approaches to drawing MCMC samples from distributions with intractable normalizing constants-referred to as "doubly intractable" are explored by Murray (2007). These approaches extend the work of Marjoram et al. (2003) and Møller et al. (2006).

In Liechty, Liechty, and Müller (2009), we propose the shadow prior construction as a practical approach to implement posterior inference for problems with constrained parameter or sample spaces. The shadow prior is used to simplify computation by strategically inserting an additional level in the hierarchical structure of the model. The purpose of this is to move the constraints on a particular parameter to a level of the hierarchy where it can be dealt with more easily. Introducing the shadow prior into a model also introduces additional complexity, as well as a slight departure from the original model. However, in our experience these limitations are more than compensated by the great savings both in computation time and in the implementation effort. We have found no substantive difference in the reported inference due to the approximation that is introduced under the shadow prior. Thus, the shadow prior can be a quick and easy fix to a difficult problem.

This section briefly reviews the shadow prior construction and its merits and limitations. For a more in depth development, with more examples, and further guidance on making implementation worthwhile see Liechty, Liechty, and Müller (2009). In Section 14.3.1 we introduce the proposed mechanism. In Section 14.3.2 we show how the shadow prior is used in the example of the correlation matrix. Section 14.3.3 describes the results of a simulation study where we investigate the nature of the implied approximation. In Section 14.3.4 we give specific guidelines for implementation of the shadow prior in general. Section 14.3.5 concludes with a final discussion.

14.3.1 The Shadow Prior

We illustrate the usefulness of the shadow prior by considering a generic Bayesian model with truncation in the sampling distribution and/or the prior.

$$p(y|\theta) \propto g(y|\theta) I\{y \in \mathscr{A}\} = \frac{g(y|\theta)}{g(\mathscr{A}|\theta)} I\{y \in \mathscr{A}\}, \tag{14.3.1}$$

$$p(\theta|\eta) \propto g(\theta|\eta) I\{\theta \in \mathscr{B}\} = \frac{g(\theta|\eta)}{g(\mathscr{B}|\eta)} I\{\theta \in \mathscr{B}\}, \tag{14.3.2}$$

$$p(\eta).$$

Here we generically use $g(\cdot)$ for an unconstrained probability model. the truncation in (14.3.1) could be, for example, a monotonicity constraint for a time series $y = (y_1, \ldots, y_T)$. A truncation in (14.3.2) could be, for example, a positive definiteness constraint for a covariance matrix. The constraints in the sampling model and/or prior cause computational complications when posterior inference requires the evaluation of the normalizing constants $g(\mathscr{A}|\theta)$ and $g(\mathscr{B}|\eta)$. These probabilities can be analytically intractable, depending on the form of the restrictions $\mathscr{A}$ and $\mathscr{B}$.

The computational problems arising from the truncation can be side stepped by introducing shadow priors. Consider first the case of a constraint in the sampling model. The idea of the shadow prior construction is to insert an additional layer $p(\delta|\theta)$ in the hierarchical model, between levels (14.3.1) and (14.3.2), and to change $p(y|\theta)$ to $p(y|\delta)$, so that y is dependent on a new intermediate parameter δ and y and θ are conditionally independent given δ. Let $\theta = (\theta_1, \ldots, \theta_d)$ denote a $d-$dimensional parameter vector. In many cases a multivariate normal shadow prior $\delta_i \sim N(\theta_i, v^2)$, $i = 1, \ldots, d$, independently, can be used, replacing (14.3.1) by

$$p(y|\delta) = \frac{g(y|\delta)}{g(\mathscr{A}|\delta)} I\{y \in \mathscr{A}\},$$
$$p(\delta|\theta) = \prod N(\delta_i;\, \theta_i, v^2),$$

and unchanged prior $p(\theta|\eta)$ and hyperprior $p(\eta)$. Here $N(x;\, m, s^2)$ denotes a normally distributed random variable x with mean m and variance s^2. This leaves the conditional posterior for θ free of the analytically intractable normalization constant. Of course the shadow prior does not entirely remove the effects of the truncation; now the truncation shows up in updating δ. In many significant problems the shadow prior can be set up in a manner such that the truncation minimally affects the conditional posterior of δ, as will be seen in the examples that follow. This is of particular interest when $p(\theta)$ has a complicated structure or θ is high dimensional. The idea behind the shadow prior is for δ to 'shadow' y. The user defined variance v^2 of the shadow prior acts as a tuning parameter. In particular, the shadow prior is useful when there are highly efficient algorithms for posterior inference, if the constraint were removed. Examples are non-parametric regression with wavelet pri-

ors (see, for example, Vidakovic, 1998) or normal dynamic linear models (West and Harrison, 1997), with constraints on the sample space. (See Liechty, Liechty, and Müller (2009) for details on implementation of the shadow prior in normal dynamic linear models.)

The case of truncation in the prior, i.e., a constraint to $\mathscr{B}$ in (14.3.2), is handled similarly, replacing (14.3.2) by

$$p(\theta|\delta) = \frac{q(\theta|\delta)}{q(\mathscr{B}|\delta)} I\{\theta \in \mathscr{B}\} \text{ and } p(\delta|\eta)$$

with the shadow prior $q(\theta|\delta)$. For example, for a d-dimensional parameter vector the shadow prior might be $q(\theta|\delta) = \prod N(\theta_i;\ \delta_i, \nu^2)$. The constraint moves from the prior indexed by η to the new shadow prior, facilitating posterior updating of η.

Whether the constraint is on the sampling model or the prior, using the shadow prior does not remove the computational challenge of the constraint, it only moves it to the conditional posterior distribution $p(\delta|\dots)$, and the problem now arises in updating δ. See Liechty, Liechty, and Müller (2009) for discussion of some significant problems where this tradeoff is worthwhile. These include problems with complicated prior structures for θ, for example, if $\theta = (\theta_1, \dots, \theta_n)$ and the prior on θ is a mixture model and η are the parameters of the mixture. Another scenario where shadow priors simplify posterior simulation are problems with a high dimensional parameter θ. Assume $\theta = (\theta_1, \dots, \theta_d)$ is a d-dimensional parameter vector with prior $p(\theta|\eta) \propto g(\theta|\eta) I\{\theta \in \mathscr{B}\}$, subject to a constraint $\theta \in \mathscr{B}$, resulting in a normalization constant $g(\mathscr{B}|\eta) = \int_{\mathscr{B}} g(\theta|\eta) d\theta$. Adding a shadow prior we can assume independence, $q(\theta|\delta) = \prod q(\theta_i|\delta_i)$, and $\delta \sim g(\delta|\eta)$. Posterior inference in the augmented model requires d univariate integrations, one for each δ_i, instead of the d-dimensional integration to evaluate $g(\mathscr{B}|\eta)$.

In summary, the shadow prior mechanism replaces a constraint in the sampling model or prior by an additional level in the hierarchical model. The constraint is moved from the sampling model or prior to this additional level. Inference under the augmented model approximates inference under the original model. With a well chosen shadow prior the extent of the approximation can be kept minimal and the augmented model can be set up to greatly simplify posterior inference. Using the shadow prior involves several implementation decisions. Foremost is how closely the shadow follows the truncated component of the hierarchical model. The "tightness" of the shadow prior parameter is modulated by specifying the variance of the distribution of the shadow prior. We have only so far implemented the shadow prior in the form of the normal distribution, with the tightness formalized by the variance parameter ν^2. Other distributions can also be used. Larger values of ν^2 reduce the posterior correlation of δ and θ and lead to faster-mixing Markov chains, the trade off is a greater departure from the original model. Smaller values of ν^2 improve the approximation to the original model at the cost of slower-mixing chains. Choice of tightness is problem specific. In the next section and in Section 14.3.3 we investigate implications of different values for ν^2. We find that if ν^2 is set small enough, in

practice the constraint on δ can be ignored, essentially removing the computational difficulty of the constraint.

14.3.2 Example: Modeling Correlation Matrices

A very good example of the usefulness of the shadow prior are priors for correlation matrices. For example, in Liechty, Liechty, and Müller (2004) we propose, among other models, a model for a correlation matrix $R = [r_{ij}]$ assuming independent normal priors for r_{ij}, $i < j$, subject to R being positive definite. Let $\mathscr{B}$ denote the set of positive definite correlation matrices. We assume

$$p(R|\mu,\sigma^2) = \frac{\prod_{i<j}\exp\left\{-1/(2\sigma^2)(r_{ij}-\mu)^2\right\}}{\int_{R\in\mathscr{B}}\prod_{i<j}\exp\left\{-1/(2\sigma^2)(r_{ij}-\mu)^2\right\}dr_{ij}} I\{R\in\mathscr{B}\}. \quad (14.3.3)$$

Hyperpriors $\mu \sim N(0,\tau^2)$, and $\sigma^2 \sim IG(\alpha = shape, \beta = scale)$ with known τ^2, α, and β complete the model. The indicator function in (14.3.3) guarantees positive definiteness of the correlation matrix and induces dependence of the r_{ij}'s. The truncation region $\mathscr{B}$ is a very convoluted subspace of the d-dimensional hypercube $[-1,1]^d$ (see Molenberghs and Rousseeuw (1994) for more details of the shape of this region). The normalization constant of interest depends directly on μ and σ^2. To find this constant analytically is not feasible. This causes a serious problem because it is needed to sample from the full conditional distributions for the r_{ij}'s, μ, and σ^2.

Model (14.3.3) is exactly the setup of (14.3.2) with $\theta = R$ and $\eta = (\mu,\sigma^2)$. Using the shadow prior here works very nicely by placing δ_{ij}'s between the r_{ij}'s and μ, replacing (14.3.3) by

$$p(R|\delta) \propto \prod N(r_{ij};\ \delta_{ij},v^2)I(R\in\mathscr{B}) \text{ and } p(\delta|\mu,\sigma^2) = \prod N(\delta_{ij};\ \mu,\sigma^2) \text{ for } i<j. \quad (14.3.4)$$

In the resulting model δ_{ij} does not inherit the positive definiteness constraint from the correlation coefficients r_{ij}, greatly simplifying the complete conditional posterior distribution of (μ,σ^2).

To further illustrate the issue at hand, consider again the full conditional density for μ in (14.3.3). The hyperparameter μ is hopelessly entangled in the normalizing constant of the prior:

$$C^{-1}(\mu,\sigma^2) = \int_{R\in\mathscr{B}}\prod_{i<j}\exp\left\{-1/(2\sigma^2)(r_{ij}-\mu)^2\right\}dr_{ij}. \quad (14.3.5)$$

Without the shadow prior we must use a Metropolis-Hastings step to update μ. As a proposal density we could use, for example, the normal distribution that results if $I\{R\in\mathscr{B}\}$ is removed from (14.3.3). The acceptance probability can be written, using μ^* to denote the proposal, as

$$\alpha = \min\left\{1, C(\mu^*, \sigma^2)/C(\mu, \sigma^2)\right\}. \tag{14.3.6}$$

We avoid evaluation of the normalizing constant C by using (14.3.4) to insert shadow parameters δ_{ij}'s into the model hierarchy between the r_{ij}'s and the prior moments (μ, σ^2). The full conditional densities for μ and σ^2 become

$$p(\mu|\ldots) \propto \prod_{i<j} \exp\left\{-1/(2\sigma^2)(\delta_{ij}-\mu)^2\right\} \exp\left\{-1/(2\tau^2)(\mu)^2\right\}$$

and

$$p(\sigma|\ldots) \propto \prod_{i<j} \exp\left\{-1/(2\sigma^2)(\delta_{ij}-\mu)^2\right\} \left(1/\sigma^2\right)^{\alpha-1} \exp\left(-\beta/\sigma^2\right),$$

and are now conjugate normal and gamma distributions, respectively. The full conditional density for δ_{ij} is similar to the full conditional for μ in the original model (14.3.3), requiring a Metropolis-Hastings step. But there are two important differences. First, the value of ν^2 can be set to any value deemed suitable. As ν^2 approaches zero, the ratio in (14.3.6), now with $C(\delta, \nu^2)$ replacing $C(\mu, \sigma^2)$, approaches one. So it is reasonable to set ν^2 to a small number and assume $C(\delta^*, \nu^2)/C(\delta, \nu^2) = 1$. The intuition behind doing this is that the full conditional density of R essentially lies inside the constrained space $\mathscr{B}$, which allows the unconstrained normalizing constant to be a reasonably good approximation of the constrained normalizing constant. Second, the dependence of the normalizing constant on μ is dispersed between the many δ_{ij}'s, making the impact of the normalizing constant less than it would be without the shadow prior.

There are a few drawbacks to using the shadow prior. For instance, the choice of ν^2 can strongly influence the mixing properties of the Markov chain. In this particular example, the chain may mix very slowly for very small ν^2 since the r_{ij}'s and δ_{ij}'s are so closely related. Also, inference under the augmented model with the shadow prior is an approximation of inference under the original model.

14.3.3 Simulation Study

To further investigate the impact of the model augmentation with the shadow prior we performed a simulation study. Using a common correlation matrix, we generated 500 samples from an eight dimensional multivariate normal distribution. We then implemented posterior inference using the shadow prior version of (14.3.3) using different values of ν^2.

The first question is about the impact of choosing very small ν^2. We expect that the mixing properties of the MCMC algorithm would suffer. This however was not the case for this example. Examination of autocorrelation and the marginal posterior density of the parameters shows us that there are no mixing issues (see Liechty, Liechty, and Müller (2009) for relevant figures and further discussion).

The appropriate choice of v^2 is case dependent. A researcher should carry out some preliminary investigation to see how the ratio of normalizing constants changes with each value of v^2. In this example, for each value of v^2 we calculated the mean of the estimated ratio of normalizing constants $C(\delta^*, v^2)/C(\delta, v^2)$. From Figure 14.9, we see that as v^2 becomes small, the average ratio approaches 1.0 and the variance of this ratio approaches zero. This suggests that for small values of v^2 we can assume $C(\delta^*, v^2)/C(\delta, v^2) = 1$. This greatly simplifies posterior inference. For $v^2 < 0.01$ we can entirely avoid evaluation of the normalization constant $C(\cdot)$.

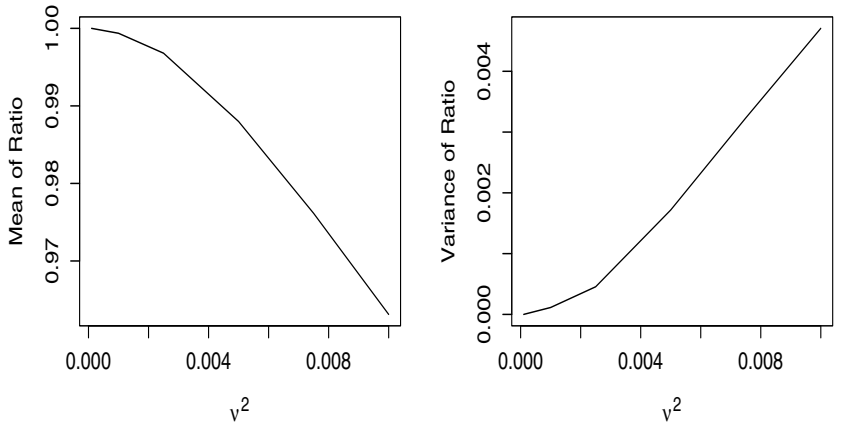


FIGURE 14.9. Correlation matrix model (14.3.3). The figure plots the ratio of normalizing constants $C(\delta^*, v^2)/C(\delta, v^2)$ for $d = 8$ dimensional covariance matrices and varying values for v^2. The left panel shows the average of computed normalizing constants. The right panels plots the empirical variance of the same ratios. As v^2 becomes small the average ratio becomes one and the variance of this ratio approaches zero. For $v^2 < 0.001$ we can assume $C(\delta^*, v^2)/C(\delta, v^2) = 1$.

14.3.4 Classes of Models Suitable for Shadow Prior Augmentations

There are several classes of problems where shadow priors could greatly simplify posterior inference. Here we give guidance on how to implement the shadow prior in these classes. The models include constraints on y or θ as indicated.

Constraint on y and high-dimensional parameter θ. The shadow prior mechanism is attractive for problems with constraints in the sampling model when the prior would be conjugate in the corresponding unconstrained model. The use of shadow priors can become particularly advantageous when the unconstrained model would

allow highly efficient posterior inference for a high dimensional parameter vector θ. This case arises, for example, in wavelet regression. The constraint could be, for example, a monotonicity constraint on y. The shadow prior replaces the model with $p(y|\delta)I\{y \in \mathscr{A}\}$, $p(\delta|\theta) = \prod_{i=1}^{d} N(\delta_i; \theta_i, v)$ and prior $p(\theta)$ (possibly still indexed by hyperparameters η). In the augmented model the conditional posterior $p(\theta|\delta)$ allows us to exploit all the advantages of posterior inference in the unconstrained model. Updating the δ_j parameters involves the evaluation of a normalization constant (in the sampling model). But now the multivariate problem is reduced to many univariate problems.

Constraints on y induce posterior dependence. A special case of the previously described class arises when a constraint in the sampling model induces posterior dependence in a model that would otherwise imply *a posteriori* independent parameters.

Constraints on θ hinder efficient posterior simulation. Consider for example a normal dynamic linear model (West and Harrison, 1997) with a constraint on the state vector θ. The constraint prevents the use of the highly efficient forward filtering and backward smoothing (FFBS) algorithm for posterior inference. Replacing the original prior by an augmented model with a shadow prior mitigates the problem. Using sufficiently small v^2 we can ignore the additional normalization constant introduced by the constraint and proceed with the FFBS algorithm as in the unconstrained model.

Other examples in this class are wavelet regression models with constraints on the mean function and in general conjugate models with an additional constraint on a multivariate parameter θ.

Constraints on θ hinder efficient posterior simulation for η. This was the case of the example in Section 14.3.2. The constraint to positive definite matrices $\mathscr{B}$ greatly complicated the complete conditional posterior for $\eta = (\mu, \sigma)$. The use of an additional shadow prior in the hierarchical model separates the hyperprior from the constraint on R and allowed efficient posterior simulation. Even more simplification is achieved when v^2 is chosen sufficiently small to ignore the normalization constants when updating δ_{ij}.

Similar simplifications are possible for mixture models $p(\theta|\eta)$ with constraints on θ. The use of a shadow prior removes the constraint that hinders the use of efficient posterior simulation schemes based on latent indicators to replace the mixture.

14.3.5 Conclusion

The proposed shadow prior mechanism addresses the computational challenge that arises from the evaluation of normalization constants in constrained data and/or parameter problems. The augmentation of the model with the proposed shadow prior trades off computational difficulties against a minor approximation and possibly slower mixing of the resulting MCMC simulation.

The main advantages of the shadow prior mechanism is that it empowers investigators to freely use constraints that are suggested by the scientific problem of interest. Many applications allow the specification of meaningful constraints that could greatly improve the relevance of statistical inference, but are usually avoided for the sake of computational simplicity.

Some limitations remain. The construction of shadow prior augmentations is not automated. It typically requires some preliminary investigation to assure an acceptable level of approximation. Also posterior simulation has to be closely watched to identify possible convergence issues due to slowly mixing Markov chains.

In summary, the shadow prior mechanism offers a potentially powerful new tool to researchers, but requires some expert judgment for proper implementation and interpretation.

References

Aarts, E. and Korst, J. (1989). *Simulated Annealing and Boltzmann Machines*. New York: Wiley.

Abraham, B. and Box, G.E.P. (1979). Bayesian analysis of some outlier problems in time series. Biometrika **66**, 229–236.

Adcock, C.J. (1997). Sample size determination: A review. *The Statistician: Journal of the Institute of Statisticians* **46**, 261–283.

Adler, R.J. (1981). *The Geometry of Random Fields*. Chichester, UK: Wiley.

Agresti, A. (1993). Distribution-free fitting of logit models with random effects for repeated categorical responses. *Statistics in Medicine* **12**, 1969–1987.

Agresti, A. and Coull, B.A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician* **52**, 119-126.

Ahn, J., Mukherjee, B., Banerjee, M., and Cooney, K.A. (2009). Bayesian inference for the stereotype regression model: Application to a case-control study of prostate cancer. *Statistics in Medicine* **28**, 3139–3157.

Airoldi, E.M. (2007). Getting started in probabilistic graphical models. *PLoS Computational Biology* **3**, e252.

Airoldi, E.M., Blei, D.M., Fienberg, S.E., and Xing, E.P. (2008). Mixed-membership stochastic blockmodels. *Journal of Machine Learning Research* **9**, 1981–2014.

Airoldi, E.M., Fienberg, S.E., Joutard, C.J., and Love, T.M. (2006). Discovery of latent patterns with hierarchical Bayesian mixed-membership models and the issue of model choice. *Technical Report*, School of Computer Science, Carnegie Mellon University.

Airoldi, E.M., Fienberg, S.E., Joutard, C.J., Love, T.M., and Shringapure, S. (2009). Re-conceptualizing the classification of PNAS articles. *Technical Report*, School of Computer Science, Carnegie Mellon University.

Aitchison, J. (1975). Goodness of prediction fit. *Biometrika* **62**, 547–554.

Aitchison, J. (1990). On coherence in parametric density estimation. *Biometrika* **77**, 905–908.

Aitkin, M. (1991). Posterior Bayes factors (with Discussion). *Journal of the Royal Statistical Society Series B.* **53** 111–142.

M.-H. Chen et al. (eds.), *Frontiers of Statistical Decision Making and Bayesian Analysis*, DOI 10.1007/978-1-4419-6944-6,

Aitkin, M. and Rubin, D.B. (1985). Estimation and hypothesis testing in finite mixture models. *Journal of the Royal Statistical Society, Series B* **47**, 67–75.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *The Second International Symposium on Information Theory* (Eds. B.N. Petrov and F. Csaki). Budapest: Akademia Kiado, pp. 267–281.

Akaike, H. (1983). Information measures and model selection. In *Proceedings of the International Statistical Institute 44th Session*. The Hague, Netherlands: International Statistical Institute, pp. 277–291.

Albert, J.H. (1992). Bayesian estimation of normal ogive item response curve using Gibbs sampling. *Journal of Educational Statistics* **17**, 251–269.

Albert, J.H. and Bennett, J. (2003). *Curve Ball*. New York: Springer.

Albert, J.H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88**, 669–679.

Aldous, D.J. (1985). Exchangeability and related topics. In *Ecole d'Eté de Probabilités de Saint-Flour XIII–1983* (Ed. P.L. Hennequin). Berlin: Springer, pp. 1–198.

Aldous, D.J. (1987). On the Markov chain simulation method for uniform combinatorial distributions and simulated annealing. *Probability in Engineering and Information Systems* **1**, 33–46.

Allen, D.M. (1971). The prediction sum of squares as a criterion for selecting prediction variables. *Technical Report 23*, Department of Statistics, University of Kentucky.

Altham, P.M.E. (1971). The analysis of matched proportions. *Biometrika* **58**, 561–576.

Amari, S. (1990). *Diffrential-Geometrical Methods in Statistics*. Second Edition. Lecture Notes in Statistics **28**. Berlin: Springer.

Andersen, E.B. (1970). Asymptotic properties of conditional maximum likelihood estimators. *Journal of the Royal Statistical Society, Series B* **32**, 283–301.

Anderson, T.W. (2003). *An Introduction to Multivariate Statistical Analysis*. New York: Wiley.

Andrews, D.F. and Herzberg, A.M. (1985). *Data*. New York: Springer.

Andrieu, C. and Atchadé, Y.F. (2007). On the efficiency of adaptive MCMC algorithms. *Electronic Communications in Probability* **12**, 336–349.

Andrieu, C. and Moulines, E. (2006). On the ergodicity properties of some adaptive MCMC algorithms. *Annals of Applied Probability 16*, 1462–1505.

Ané, C., Larget, B., Baum, D.A., Smith, S.D., and Rokas, A. (2007). Bayesian estimation of concordance among gene trees. *Molecular Biology and Evolution* **24**, 412–426.

Antoniak, C. (1974). Mixtures of Dirichlet processes with application to Bayesian non-parametric problems. *Annals of Statistics* **2**, 1152–1174.

Aragones, E. and Palfrey, T. (2002). Mixed equilibrium in Downsian model with a favored candidate. *Journal of Economic Theory* **103**, 131–142.

Aranda-Ordaz, F.J. (1981). On two families of transformations to additivity for binary response data. *Biometrika* **68**, 357–364.

Arellano-Valle, R.B. and Bolfarine, H. (1995). On some characterizations of the t-distribution. *Statistics and Probability Letters* **25**, 79–85.

Armstrong, B.G. (2003). Fixed factors that modify the effects of time-varying factors: Applying the case-only approach. *Epidemiology* **14**, 467–472.

Ashby, D., Hutton, J.L., and McGee, M.A. (1993). Simple Bayesian analyses for case-controlled studies in cancer epidemiology. *Statistician* **42**, 385–389.

Atchadé, Y.F., Fort, G., Moulines, E., and Priouret, P. (2009). Adaptive Markov chain Monte Carlo: Theory and methods. *Technical Report*, Department of Statistics, University of Michigan.

Atchadé, Y.F. and Rosenthal, J.S. (2005). On adaptive Markov chain Monte Carlo algorithms. *Bernoulli* **11**, 815–828.

Aslan, M. (2006). Asymptotically minimax Bayes predictive densities. *Annals of Statistics* **34**, 2921–2938.

Aurenhammer, F. (1991). Voronoi diagrams a study of a fundamental geometric data structure. *ACM Computing Surveys* **23**, 345–405.

Avramov, D. (2002). Stock return predictability and model uncertainty. *Journal of Financial Economics* **64**, 423–458.

Azzalini, A. and Capitanio, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t distribution. *Journal of the Royal Statistical Society, Series B* **65**, 367–389.

Badache, A. and Gonçalves, A. (2006). The ErbB2 signaling network as a target for breast cancer therapy. *Journal of Mammary Gland Biology and Neoplasia* **11**, 13–25.

Bae, K. and Mallick, B.K. (2004). Gene selection using a two-level hierarchical Bayesian model. *Bioinformatics* **20**, 3423–30.

Baker, G.A. and Crosbie, P.J. (1993). Measuring food safety preferences: Identifying consumer segments. *Research in Agriculture & Applied Economics* **18**, 277–287.

Banerjee, S., Carlin, B.P., and Gelfand, A.E. (2004). *Hierarchical Models and Analysis for Spatial Data*. Boca Raton, FL: Chapman & Hall / CRC Press.

Banerjee, S. and Gelfand, A.E. (2003). On smoothness properties of spatial processes. *Journal of Multivariate Analysis* **84**, 85–100.

Banerjee, S. and Gelfand, A.E. (2006). Bayesian wombling: Curvilinear gradient assessment under spatial process models. *Journal of the American Statistical Association*, **101**, 1487–1501.

Banerjee, S., Gelfand, A.E., and Sirmans, C.F. (2003). Directional rates of change under spatial process models. *Journal of the American Statistical Association* **98**, 946–954.

Banfield, J.D. and Raftery, A.E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49**, 803–821.

Banks, H.T., Karr, A.F., Nguyen, H.K., and Samuels Jr., J.R. (2008). Sensitivity to noise variance in a social network dynamics model. *Quarterly of Applied Mathematics* **66**, 233–247.

Baranchik, A.J. (1964). Multiple regression and estimation of the mean of a multivariate normal distribution. *Technical Report No. 51*, Department of Statistics, Stanford University.

Barberis, N. (2000). Investing for the long run when returns are predictable. *Journal of Finance* **55**, 225–264.

Barbieri, M.M. and Berger, J.O. (2004). Optimal predictive model selection. *Annals of Statistics* **32**, 870–897.

Barnard, K., Duygulu, P., de Freitas, N., Forsyth, D., Blei, D.M., and Jordan, M.I. (2003). Matching words and pictures. *Journal of Machine Learning Research* **3**, 1107–1135.

Barron, A.R., Rissanen, J., and Yu, B. (1998). The minimum description length principle in coding and modelling. *IEEE Transaction on Information Theory* **44**, 2743–2760.

Barron, A.R., Schervish, M.J., and Wasserman, L. (1999). The consistency of posterior distributions in nonparametric problems. *Annals of Statistics* **27**, 536–561.

Basu, S. (1995). Ranges of posterior probability over a distribution band. *Journal of Statistical Planning and Inference* **44** 149–166.

Basu, S. (1996). Local sensitivity, functional derivatives and nonlinear posterior quantities. Statist. *Statistics & Decisions* **14**, 405–418.

Basu, S. (1999). Posterior sensitivity to the sampling distribution and the prior: More than one observation. *Annals of the Institute of Statistical Mathematics* **51**, 499–513.

Bates, D. (2005). Fitting linear models in R using the lme4 package. *R News* **5**, 27–30.

Battese, G.E., Harter, R.M., and Fuller, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association* **83**, 28–36.

Baudry, J.P., Raftery, A.E., Celeux, G., Lo, K., and Gottardo, R. (2008). Combining mixture components for clustering. *Technical Report 540*, Department of Statistics, University of Washington.

Bayarri, M.J. and Berger, J.O. (2000). P values for composite null models. *Journal of the American Statistical Association* **95** 1127–1142.

Bayarri, M.J., Berger, J.O., Molina, G., Rouphail, N.M., and Sacks, J. (2004). Assessing uncertainties in traffic simulation: A key component in model calibration and validation. *Transportation Research Record* **1876**, 32–40.

Bayarri, M.J., Berger, J.O., Paulo, R., Sacks, J., Cafeo, J.A., Cavendish, J., Lin, C.H., and Tu, J. (2005). A framework for validation of computer models. *Technical Report 162*, National Institute of Statistical Sciences.

Bayarri, M.J., Berger, J.O., Cafeo, J.A., Garcia-Donato, G., Liu, F., Palomo, J., Parthasarathy, R.J., Paulo, R., Sacks, J., and Walsh, D. (2007a). Computer model validation with functional output. *Annals of Statistics* **35**, 1874–1906.

Bayarri, M.J., Berger, J.O., Paulo, R., Sacks, J., Cafeo, J.A., Cavendish, J., Lin, C.H., and Tu, J. (2007b). A framework for validation of computer models. *Technometrics* **49**, 138–154.

Bayarri, M.J., Berger, J.O., Calder, E., Dalbey, K., Lunagomez, S., Patra, A.K., Pitman, E.B., Spiller, E.T., and Wolpert, R.L. (2009a). Using Statistical and Computer Models to Quantify Volcanic Hazards. *Technometrics, 51*, 402–413.

Bayarri, M.J., Berger, J.O., Kennedy, M.C., Kottas, A., Paulo, R., Sacks, J., Cafeo, J.A., Lin, C.H., and Tu, J. (2009b). Predicting vehicle crashworthiness: Validation of computer models for functional and hierarchical data. *Journal of the American Statistical Association* **104**, 929–943.

Bayarri, M.J. and Castellanos, M. (2007). Bayesian checking of the second levels of hierar- chical models. *Statistical Science* **22**, 322–342.

Bayarri, M.J. and Morales, J. (2003). Bayesian measures of surprise for outlier detection. *Journal of Statistical Planning and Inference* **111**, 3–22.

Bayes, C.L. and Branco, M.D. (2007). Bayesian inference for the skewness parameter of the scalar skew-normal distribution. *Brazilian Journal of Probability and Statistics* **21**, 141–163.

Bellinger, F.P., He, Q., Bellinger, M., Lin, Y., Raman, A.V., White, L.R., and Berry, M. (2008). Association of selenoprotein P with Alzheimers pathology in human cortex. *Journal of Alzheimers Disease* **15**, 465–472.

Bellman, R. and Kalaba, R. (1958). On communication processes involving learning and random duration. *IRE National Convention Record* **4**, 16–20.

Ben-Dor, A. and Yakhini, Z. (1999). Clustering gene expression patterns. In *Proceedings of the Third Annual International Conference on Computational Molecular Biology*. Lyon, France: ACM Press, PP. 33–42.

Benzecri, J.P. (1992). *Correspondence Analysis Handbook*. New York: Marcel Dekker.

Berg, J. and Lassig, M. (2004). Local graph alignment and motif search in biological networks. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 14689–14694.

Berg, J. and Lassig, M. (2006). Cross-species analysis of biological networks by Bayesian alignment. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 10967–10972.

Berger, J.O. (1975). Minimax estimation of location vectors for a wide class of densities. *Annals of Statistics* **3**, 1318–1328.

Berger, J.O. (1976). Admissible minimax estimation of multivariate normal mean with arbitrary quadratic loss. *Annals of Statistics* **4**, 223–226.

Berger, J.O. (1980). A robust generalized Bayes estimator and confidence region for a multivariate normal mean. *Annals of Statistics* **8**, 716–761.

Berger, J.O. (1984). The robust Bayesian viewpoint (with Discussion). In *Robustness of Bayesian Analysis* (Ed. J.B. Kadane). Amsterdam: North-Holland, pp. 63–144.

Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Second Edition. New York: Springer.

Berger, J.O. (1990). Robust bayesian analysis: Sensitivity to the prior. *Journal of Statistical Planning and Inference* **25**, 303–328.

Berger, J.O. (1994). An overview of robust bayesian analysis. *Test* **3**, 5–58.

Berger, J.O. (2006). The case for objective Bayesian analysis (with Discussion). *Bayesian Analysis* **1**, 385–402.

Berger, J.O. and Bernardo, J.M. (1989). Estimating a product of means: Bayesian analysis with reference priors. *Journal of the American Statistical Association* **84**, 200–207.

Berger, J.O. and Bernardo, J.M. (1992a). On the development of reference priors (with Discussion). In *Bayesian Statistics* **4** (Eds. J.M. Bernado, J.O. Berger, A.P. Dawid, and A.F.M. Smith). Oxford: Oxford University Press, pp. 35–60.

Berger, J.O. and Bernardo, J.M. (1992b). Ordered group reference priors with application to the multinomial. *Biometrika* **25**, 25–37.

Berger, J.O. and Bernardo, J.M. (1992c). Reference priors in a variance components problem. In *Bayesian Inference in Statistics and Econometrics* (Eds. P.K. Goel and N.S. Iyengar). New York: Springer, pp. 177–194.

Berger, J.O., Bernardo, J.M., and Mendoza, M. (1989). On priors that maximize expected information. In *Recent Developments in Statistics and Their Applications* (J. Klein and J.C. Lee). Seoul: Freedom Academy Publishing, pp. 1–20.

Berger, J.O., Bernardo, J.M., and Sun, D. (2009). The formal definition of reference priors. *Annals of Statistics* **37**, 905–938.

Berger, J.O., Bock, M.E., Brown, L., Casella , G., and Gleser, L. (1977). Minimax estimation of a normal mean vector for arbitrary quadratic loss and unknown covariance matrix. *Annals of Statistics* **5**, 763–771.

Berger, J.O. and Delampady, M. (1987). Testing precise hypotheses (with Discussion). *Statistical Science* **2**, 317–352.

Berger, J.O., De Oliveira, V., and Sansó, B. (2001). Objective Bayesian analysis of spatially correlated data. *Journal of the American Statistical Association* **96**, 1361–1374.

Berger, J.O. and Guglielmi, A. (2001). Bayesian and conditional frequentist testing of a parametric model against nonparametric alternatives. *Journal of the American Statistical Association* **96**, 174–184.

Berger, J.O., Liseo, B., and Wolpert, R.L. (1999). Integrating likelihood methods for eliminating nuisance parameters. *Statistical Science* **14**, 1–28.

Berger, J.O. and Molina, G. (2005). Posterior model probabilities via path-based pairwise priors. *Statistica Neerlandica* **59**, 3–15.

Berger, J.O. and Moreno, E. (1994). Bayesian robustness in bidimensional models: prior independence. *Journal of Statistical Planning and Inference* **20**, 1697–1710.

Berger, J.O. and Pericchi, L. (1996a). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association* **91**, 109–122.

Berger J.O. and Pericchi L. (1996b). The intrinsic Bayes factor for linear models. In *Bayesian Statistics* **5** (Eds. J.M. Bernado, J.O. Berger, A.P. Dawid, and A.F.M. Smith). Oxford: Oxford University Press, pp. 23–42.

Berger J.O. and Pericchi L. (1996c). On the Justification of Default and Intrinsic Bayes Factors. In *Modelling and Prediction* (Eds. J.C. Lee, W. Johnson, and A. Zellner). New York: Springer, pp. 276–293.

Berger, J.O. and Pericchi, L. (2001). Objective Bayesian methods for model selection: Introduction and comparison (with Discussion). In *Model Selection* (Ed. P. Lahiri). Institute of Mathematical Statistics Lecture Notes - Monograph Series, **38**, Beachwood, Ohio: Institute of Mathematical Statistics, pp. 135–207.

Berger, J.O., Pericchi, L., and Varshavsky, J. (1998). Bayes factors and marginal distributions in invariant situations. *Sankhyā, Series A* **60**, 307–321.

Berger, J.O., Ríos Insua, D., and Ruggeri, F. (2000). Bayesian robustness. In *Robust Bayesian Analysis*. *Lecture Notes in Statistics* **152** (Eds. D. Ríos Insua and F. Ruggeri). New York: Springer, pp. 1-31.

Berger, J.O. and Sellke, T. (1987). Testing of a point null hypothesis: The irreconcilability of signifiance levels and evidence (with dDscussion). *Journal of the American Statistical Association* **82**, 112–139.

Berger, J.O. and Strawderman, W.E. (1996). Choice of hierarchical priors: Admissibility in estimation of normal means. *Annals of Statistics* **24**, 931–951.

Berger, J.O., Strawderman, W.E., and Tang, D. (2005). Posterior propriety and admissibility of hyperpriors in normal hierarchical models. *Annals of Statistics* **33**, 606–646.

Berger, J.O. and Sun, D. (2008). Objective priors for the bivariate normal model. *Annals of Statistics* **36**, 963–982.

Bernardo, J.M. (1979). Reference posterior distributions for Bayesian inference (with Discussion). *Journal of the Royal Statistical Society, Series B* **41**, 113–147.

Bernardo, J.M. (2005). Reference analysis. *Bayesian Thinking: Modeling and Computation*. *Handbook of Statistics* **25** (Eds. D.K. Dey and C.R. Rao). Amsterdam: Elsevier, pp. 17–90.

Bernardo, J.M. and Smith, A.F.M. (1994). *Bayesian Theory*. New York: Wiley.

Berry, D.A. and Ho, C.-H. (1988). One-sided sequential stopping boundaries for clinical trials: A decision-theoretic approach. *Biometrics* **44**, 219–227.

Bertucci, F., Borie, N., Ginestier, C., Groulet, A., Charafe-Jauffret, E., Adélaïde, J., Geneix, J., Bachelart, L., Finetti, P., Koki, A., Hermitte, F., Hassoun, J., Debono, S., Viens, P., Fert, V., Jacquemier, J., and Birnbaum, D. (2004). Identification and validation of an ERBB2 gene expression signature in breast cancers. *Oncogene* **23**, 2564–2575.

Betro, B., Meczarski, M., and Ruggeri, F. (1994). Robust bayesian analysis under generalized moments conditions. *Journal of Statistical Planning and Inference* **41**, 257–266.

Bhat, K., Haran, M., Tonkonojenkov, R., and Keller, K. (2009). Inferring likelihoods and climate system characteristics from climate models and multiple tracers. *Technical Report*, Department of Statistics, Pennsylvania State University.

Bhattacharya, S., Ghosh, J.K., and Samanta, T. (2009). On Bayesian "central clustering": Application to landscape classification of Western Ghats. *Technical Report*, Department of Statistics, Purdue University.

Biernacki, C., Celeux, G., and Govaert, G. (1998). Assessing a mixture model for clustering with the integrated classification likelihood. *Technical Report No. 3521*, Institut National de Recherche en informatique et en automatique.

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated complete likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**, 719–725.

Bild, A.H., Yao, G., Chang, J.T., Wang, Q., Potti, A., Chasse, D., Joshi, M., Harpole, D., Lancaster, J.M., Berchuck, A., Olson Jr. , J.A., Marks, J.R., Dressman, H.K., West, M., and Nevins, J.R. (2006). Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* **439**, 353–357.

Bingham, C. (1974). An antipodally symmetric distribution on the sphere. *Annals of Statistics* **2**, 1201–1225.

Bininda-Emonds, O.R.P. (2005). Supertree construction in the genomic age. *Methods in Enzymology* **395**, 747–757.

Birmiwal, L. and Dey, D.K. (1993). Measuring local inflence of posterior features under contaminated classes of priors. *Statistics & Decisions* **11**, 377–390.

Birnbaum, A. (1969). Statistical theory for logistic mental test models with a prior distribution of ability. *Journal of Mathematical Psychology* **6**, 258–276.

Bishop, C.M. (1998). Bayesian PCA. *Neural Information Processing Systems* **11**, 382–388.

Bishop, Y.M., Fienberg, S.E., and Holland, P.W. (2007). *Discrete Multivariate Analysis, Theory and Practice*. New York: Springer.

Black, F. and Litterman, R. (1991). Asset allocation: Combining investor views with market equilibrium. *Journal of Fixed Income* **1**, 7–18.

Blackwell, D. and MacQueen, J.B. (1973). Ferguson distributions via Pólya urn schemes. *Annals of Statistics* **1**, 353–355.

Blanquart, S. and Lartillot, N. (2008). A site- and time-heterogeneous model of amino acid replacement. *Molecular Biology and Evolution* **25**, 842–858.

Blei, D.M., Griffiths, T.L., and Jordan, M.I. (2010). The nested Chinese restaurant process and Bayesian inference of topic hierarchies. *Journal of the ACM* **57**. In press.

Blei, D.M., Ng, A.Y., and Jordan, M.I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research* **3**, 993–1022.

Bloomquist, E.W., Dorman, K.S., and Suchard M.A. (2009). Stepbrothers: Inferring partially shared ancestries among recombinant viral sequences. *Biostatistics* **10**, 106–120.

Bloomquist, E.W. and Suchard, M.A. (2009). Unifying vertical and non-vertical evolution. *Systematic Biology*. In press.

Blyth, S. (1994). Local divergence and association (Corr: 95V82 p667). *Biometrika* **81**, 579–584.

Bock, M.E. (1988). Shrinkage estimators: Pseudo Bayes estimators for normal mean vectors. In *Statistical Decision Theory IV* (Eds. S.S. Gupta and J.O. Berger). New York: Springer, pp 281-298.

Bock, R.D. and Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika* **35**, 179–197.

Bockenholt, U., Barlas, S., and der Heijden, P.V. (2009). Do randomized response designs eliminate response biases? An empirical study of noncompliance behavior. *Journal of Applied Econometrics* **24**, 377–392.

Boratinska, A. (1996). On bayesian robustness with the ε-contamination class of priors. *Statistics and Probability Letters* **26**, 323–328.

Borg, I. and Groenen, P.J.F. (2005). *Modern Multidimensional Scaling: Theory and Applications*. Second Edition. New York: Springer.

Bose, S. (1990). Bayesian robustness with shape-constrained priors and mixtures of priors. *Unpublished Ph.D. Dissertation*, Department of Statistics, Purdu University.

Bose, S. (1994). Bayesian robustness with mixture classes of priors. *Annals of Statistics* **22**, 652–667.

Boudoukh, J., Michaely, R., Matthew Richardson, M., and Michael Roberts, M. (2007). On the importance of measuring payout yield: Implications for empirical asset pricing. *Journal of Finance* **62**, 877–916.

Box, G.E.P. (1980). Sampling and Bayes inference in scientifi modelling and robustness. *Journal of the Royal Statistical Soceity, Series A* **143**, 383–430.

Box, G.E.P. and Tiao, G.C. (1968). A Bayesian approach to some outlier problems. *Biometrika* **55**, 119–129.

Box, G.E.P. and Tiao, G.C. (1973). *Bayesian Inference in Statistical Analysis*. New York: Wiley.

Bradlow, E. and Zaslavsky, A.M. (1997). Case inflence analysis in Bayesian inference. *Journal of Computational and Graphical Statistics* **6**, 314–331.

Brandt, M.W., Goyal, A., Santa-Clara, P., and Stroud, J.R. (2005). A simulation approach to dynamic portfolio choice with an application to learning about predictability. *Review of Financial Studies* **18**, 831–873.

Brandt, M.W. (2009). Portfolio choice Problems. In *Handbook of Financial Econometrics* (Eds. Y. Ait-Sahalia and L.P. Hansen). Amsterdam: Elsevier. In press.

Brandwein A.C. (1979). Minimax estimation of the mean of spherically symmetric distributions under quadratic loss. *Journal of Multivariate Analysis* **9**, 579-588.

Brandwein, A.C. and Strawderman, W.E. (1978). Minimax estimation of location parameters for spherically symmetric unimodal distributions. *Annals of Statistics* **6**, 377–416.

Brandwein, A.C. and Strawderman, W.E. (1990). Stein estimation: The spherically symmetric case. *Statistical Science* **5**, 356–369.

Brandwein, A.C. and Strawderman, W.E. (1991). Generalizations of James-Stein estimators under spherical symmetry. *Annals of Statistics* **19**, 1639-1650.

Brandwein, A.C. and Strawderman, W.E. (2005). Bayesian estimation of multivariate location parameters. In *Bayesian Thinking: Modelling and Computation*. *Handbook of Statistics* **25** (Eds. C.R. Rao and D.K. Day). Amsterdam: Elsevier, pp. 221–244.

Breiman, L. (1996). Stacked regression. *Machine Learning* **24**, 49–64.

Breslow, N.E. (1996). Statistics in epidemiology: The case-control study. *Journal of the American Statistical Association* **91**, 14–28.

Breslow, N.E. and Cain, K.C. (1988). Logistic regression for two-stage case-control data. *Biometrika* **75**, 11-20.

Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9–25.

Breslow, N.E. and Day, N.E. (1980). *Statistical Methods in Cancer Research*, Volume 1. Lyon, International Agency for Research on Cancer.

Breslow, N.E., Day, N.E., Halvorsen, K.T., Prentice, R.L., and Sabai, C. (1978). Estimation of multiple relative risk functions in matched case-control studies. *American Journal of Epidemiology* **108** , 299–307

Broad Institute (2007). Molecular signatures data base. http://www.broad.mit.edu/gsea/msigdb/.

Broadie, M., Chernov, M., and Johannes, M. (2007). Model specification and risk premia: Evidence from SP500 futures options. *Journal of Finance* **62**, 1453–1490.

Broadie, M., Chernov, M., and Johannes, M. (2009). Understanding index option returns. *Review of Financial Studies*. In press.

Brown, L.D. (1966). On the dmissibility of invariant estimators of one or more location parameters. *Annals of Mathematical Statistics* **37**, 1087–1136.

Brown, L.D. (1971). Admissible estimators, recurrent diffusions, and insoluble boundary value problems. *Annals of Mathematical Statistics* **42**, 855-903.

Brown, L.D., George, E.I., and Xu, X. (2008). Admissible predictive density estimation. *Annals of Statistics* **36**, 1156–1170.

Brown, L.D. and Hwang, J.T. (1982) A unified admissibility proof. In *Statistical Decision Theory and Related Topics III* **1** (Eds. S.S. Gupta and J.O. Berger). New York: Academic Press, pp. 205–230.

Brown, P.J. (2001). The generalized inverted Wishart distribution. In *Encylopedia of Environmetrics* (Eds. A.H. El-Shaarawi and W.W. Piegorsch). New York: Wiley, pp. 1079–1083.

Brown, P.J., Fearn, T., and Vannucci, M. (1999). The choice of variables in multivariate regression: A non-conjugate Bayesian decision theory approach. *Biometrika* **86**, 635–648.

Brown, P.J., Vannucci, M., and Fearn, T. (1998). Multivariate Bayesian Variable Section and Prediction. *Journal of the Royal Statistical Society, Series B* **60**, 627–641.

Browne, S. and Whitt, W. (1996). Portfolio choice and the Bayesian kelly criterion. *Advances in Applied Probability* **28**, 1145–1176.

Bruss, T. and Ferguson, T.S. (2002). High risk and competitive investment models. *Annals of Applied Probability* **12**, 1202–1226.

Campbell, J.Y. and Thompson, S.B. (2008). Predicting excess stock returns out of sample: Can anything beat the historical average. *Review of Financial Studies* **21**, 1509–1531.

Campbell, J.Y. and Viceira, L. (1999). Consumption and portfolio decisions when expected returns are time-varying. *Quarterly Journal of Economics* **114**, 433–495.

Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics* **35**, 2313–2351.

Cappé, O., Godsill, S., and Moulines, E. (2007). An overview of existing methods and recent advances in sequential Monte Carlo. *Proceedings of the IEEE* **95**, 899–924.

Carlin, B.P. and Chib, S. (1995). Bayesian model choice through Markov chain Monte Carlo. *Journal of the Royal Statistical Society, Series B* **57**, 473–484.

Carlin, B.P., Kadane, J.B., and Gelfand, A.E. (1998). Approaches for optimal sequential decision analysis in clinical trials. *Biometrics* **54**, 964–975.

Carlin, B.P. and Louis, T.A. (2005). *Bayes and Empirical Bayes Methods for Data Analysis*. Boca Raton, FL: Chapman & Hall / CRC Press.

Carlin, B.P. and Polson, N.G. (1991a). An expected utility approach to inflence diagnostics. *Journal of the American Statistical Association* **86**, 1013–1021.

Carlin, B.P. and Polson, N.G. (1991b). Inference for nonconjugate Bayesian models using the gibbs sampler. *Canadian Journal of Statistics* **19**, 399–405.

Carlin, B.P., Polson, N.G., and Stoffer, D. (1992). A Monte Carlo approach to nonnormal and nonlinear state-space modeling. *Journal of the American Statistical Association*, **87**, 493–500.

Carota, C. and Parmigiani, G. (1996). On Bayes factors for nonparametric alternatives. In *Bayesian Statistics* **5** (Eds. J.M. Bernado, J.O. Berger, A.P. Dawid, and A.F.M. Smith). Oxford: Oxford University Press, pp. 507–511.

Carroll, J.D. (1980). Models and methods for multidimensional analysis of preferential choice (or other dominance) data. In *Similarity and Choice* (Eds., E.D. Lantermann and H. Feger). Vienna: Hans Huber Publishers, pp. 234–289.

Carroll, J.D. and Green, P.E. (1997). Psychometric methods in marketing research: Part II, multidimensional scaling. *Journal of Marketing Research* **34**, 193–204.

Carroll, R.J., Wang, S., and Wang, C.Y. (1995). Prospective analysis of logistic case-control studies. *Journal of the American Statistical Association* **90**, 157–169.

Carter, C. and Kohn, R. (1994). On Gibbs sampling for state space models. *Biometrika* **81**, 541–553.

Carvalho, C.M., Lucas, J.E., Wang, Q., Chang, J., Nevins, J.R., and West, M. (2008). High-dimensional sparse factor modelling: Applications in gene expression genomics. *Journal of the American Statistical Association* **103**, 1438–1456.

Carvalho, C.M., Johannes, M., Lopes, H., and Polson, N.G. (2009). Particle learning and smoothing. *Discussion Paper*, Booth School of Business, University of Chicago.

Carvalho, C.M., Polson, N.G., and Scott, J.G. (2008). The horseshoe estimator for sparse signals. *Technical Report*, Department of Statistical Science, Duke University.

Casella, G. and Hwang, J.T. (1983). Empirical Bayes confidence sets for the mean of a multivariate normal distribution. *Journal of the American Statistical Association* **78**, 688–697.

Casella, G. and Moreno, E. (2005). Objective Bayesian variable selection. *Technical Report*, Department of Statistics, Univeristy of Florida.

Casella, G. and Moreno, E. (2009). Assessing robustness of intrinsic test of independence in two-way contingency Tables. *Technical Report*, Department of Statistics, University of Florida.

Celeux, G. and Diebolt, J. (1992). A stochastic approximation type EM algorithm for the mixture problem. *Stochastics and Stochastics Reports* **41**, 127–146.

Celeux, G., Forbes, F., Robert, C.P., and Titterington, D.M. (2006). Deviance information criteria for missing data models (with Discussion). *Bayesian Analysis* **1**, 651–706.

Celeux, G. and Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification* **13**, 195–212.

Cellier, D. and Fourdrinier, D. (1995). Shrinkage estimators under spherical symmetry for the general linear model. *Journal of Multivariate Analysis* **52**, 338–351.

Chakrabarti, A. and Ghosh, J.K. (2007). Some aspects of Bayesian model selection for prediction. In *Bayesian Statistics* **8** (Eds. J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith, and M. West). Oxford: Oxford University Press, pp. 51–90.

Chaloner, K. (1991). Bayesian residual analysis in the presence of censoring. *Biometrika* **78**, 637–644.

Chaloner, K. and Brant, R. (1988). A Bayesian approach to outlier detection and residual analysis. *Biometrika* **75**, 651–659.

Chamberlain, G. (2000). Econometric applications of maxmin expected utility. *Journal of Applied Econometrics* **15**, 625–644.

Chan, K.S. and Ledolter, J. (1995). Monte Carlo EM estimation for time series models involving observations. *Journal of the American Statistical Association* **90**, 242–252.

Chang, I.H., Kim, B.H., and Mukerjee, R. (2003). Probability matching priors for predicting unobservable random effects with application to ANOVA models. *Statistics and Probability Letters* **62**, 223–228.

Chang, J., Carvalho, C.M., Mori, S., Bild, A., Wang, Q., West, M., and Nevins, J.R. (2009). Decomposing cellular signaling pathways into functional units: A genomic strategy. *Molecular Cell* **34**, 104–114.

Chang, K. and Ghosh, J. (1998). Principal curve classifier - A nonlinear approach to pattern classification. *Proceeding of IEEE International Joint Conference on Neural Networks*, 695–700.

Chen, C.F. (1985). On asymptotic normality of limiting density functions with Bayesian implications. *Journal of he Royal Statistical Society, Series B* **97**, 540–546.

Chen, M.-H., Dey, D.K., and Ibrahim, J.G. (2004). Bayesian criterion based model assessment for categorical data. *Biometrika* **91**, 45–63.

Chen, M.-H., Dey, D.K., and Shao, Q.-M. (1999). A new skewed link model for dichotomous quantal response data. *Journal of the American Statistical Association* **94**, 1172–1186.

Chen, M.-H. and Shao, Q.-M. (1997). On Monte Carlo methods for estimating ratios of normalizing constants. *Annals of Statistics* **25**, 1563–1594.

Chen, M.-H., Shao, Q.-M., and Ibrahim, J.G. (2000). *Monte Carlo Methods in Bayesian Computation*. New York: Springer.

Chen, M.-H. and Schmeiser, B.W. (1993). Performance of the Gibbs, Hit-and-Run, and Metropolis samplers, *Journal of Computational and Graphical Statistics* **2**, 251–272.

Chen, J.L., Lucas, J.E., Schroeder, T., Mori, S., Nevins, J., Dewhirst, M., West, M., and Chi, J.T. (2008). The genomic analysis of lactic acidosis and acidosis response in human cancers. *PLoS Genetics* **4**, e1000293.

Chen, L.S. and Buja, A. (2009). Local multidimensional scaling for nonlinear dimension reduction, graph drawing, and proximity analysis. *Journal of the American Statistical Association* **104**, 209–219.

Chen, R. and Liu, J.S. (2000). Mixture Kalman filters. *Journal of Royal Statistical Society, Series B* **62**, 493–508.

Cheng, K.F. and Chen, J.H. (2005). Bayesian models for population based case-control studies when the population is in Hardy-Weinberg equilibrium. *Genetic Epidemiology* **28**, 183–192.

Chernov, M. and Ghysels, E. (2000). A study towards a unified approach to the joint estimation of objective and risk neutral measures for the purpose of options valuation. *Journal of Financial Economics* **56**, 407–458.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association* **90**, 1313–1321.

Chib, S. and Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika* **85**, 347–361.

Chib, S. and Jeliazkov, I. (2001). Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association* **96**, 270–281.

Chickering, D.W and Heckerman, D. (2000). A comparison of scientific and engineering criteria for Bayesian model selection. *Statistics and Computing* **10**, 55–62.

Chilés, J.P. and Delfiner, P. (1999). *Geostatistics: Modeling Spatial Uncertainty*. New York: Wiley.

Chipman, H., George, E.I., and McCulloch, R.E. (2001). The practical implementation of Bayesian model selection. In *Model Selection*, IMS Lecture Notes - Monograph Series Volume **38** (Ed. P. Lahiri). Beachwood, OH: Institute of Mathematical Statistics, pp. 67–134.

Cho, H. (2009). Bayesian influence diagnostic methods for parametric regression models. *Unpublished Ph.D. Dissertation*, Department of Biostatistics, University of North Carolina at Chapel Hill.

Cho, H., Ibrahim, J.G., Sinha, D., and Zhu, H. (2009). Bayesian case inflence diagnostics for survival models. *Biometrics* **65**, 116–124.

Cho, R.J., Campbell, M.J., Winzeler, E.A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T.G., Gabrielian, A.E., Landsman, D., Lockhart, D.J., and Davis, R.W. (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell* **2**, 65–73.

Choi, T. and Ramamoorthi, R.V. (2008). Remarks on consistency of posterior distributions. In *IMS Collections 3. Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh* (Eds. S. Ghosal and B. Clarke). Beachwood, OH: Institute of Mathematical Statistics, pp. 170–186.

Chopin, N. and Robert, C.P. (2007). Contemplating evidence: Properties, extensions of, and alternatives to nested sampling. *Technical Report 2007-46*, CEREMADE, Université Paris Dauphine. ArXiv:0801.3887.

Chopra, V.K. and Ziemba, W.T. (1993). The effect of errors in means, variances, and covariances on optimal portfolio choice. *Journal of Portfolio Management* **19**, 6–11.

Choudhuri, N., Ghosal, S., and Roy, A. (2005). Bayesian models for function estimation. *Bayesian Thinking: Modeling and Computation*. *Handbook of Statistics* **25** (Eds. D.K. Dey and C.R. Rao). Amsterdam: Elsevier, pp. 373–414.

Christen, J.A., Müller, P., Wathen, J.K., and Wolf, J. (2004). A bayesian randomized clinical trial: A decision theoretic sequential design. *Canadian Journal of Statistics* **32**, 1–16.

Christensen, O.F. and Ribeiro Jr., P.J. (2002). **geoRglm:** A package for generalised linear spatial models. *R-NEWS* **2**, 26–28.

Christensen, R. (1997). *Log-linear Models and Logistic Regression*. New York: Springer.

Ciuleanu, T.E., Brodowicz, T., Belani, C.P., Kim, J., Krzakowski, M., Laack, E., Wu, Y., Peterson, P., Adachi, S., and Zielinski, C.C. (2008). Maintenance pemetrexed plus best supportive care (bsc) versus placebo plus bsc: A phase III study. *Journal of Clinical Oncology (Meeting Abstracts)* **26**, 8011.

Clarke, B. and Gustafson, P. (1998). On the overall sensitivity of the posterior distributions to its inputs. *Journal of Statistical Planning and Inference* **71**, 137–150.

Clarke, B. and Sun, D. (1997). Reference priors under the chi-square distance. *Sankhyā* **59**, 215–231.

Clarke, B. and Yuan, A. (1999). An informative criterion for likelihood selections. *IEEE Transaction on Information Theory* **45**, 562–571.

Clarke, B. and Yuan, A. (2004). Partial information reference priors: Derivation and interpretations. *Journal of Statistical Planning and Inference* **123**, 313–345.

Clevensen, M.L. and Zidek, J. (1975). Simultaneous estimation of the mean of independent Poisson laws. *Journal of the American Statistical Association* **70**, 698–705.

Clinton, J., Jackman, S., and Rivers, D. (2004). The statistical analysis of roll call data. *American Political Science Review* **98**, 355–370.

Clyde, M., Desimone, H., and Parmigiani, G. (1996). Prediction via orthogonalized model mixing. *Journal of the American Statistical Association* **91**, 1197–208.

Clyde, M. and George, E.I. (2004). Model uncertainty. *Statistical Science 19*, 81–94.

Clyde, M., Ghosh, J., and Littman, M. (2009). Bayesian adaptive sampling for variable selection. *Technical Report*, Department of Statistical Science, Duke University.

Coles, S.G. (2001). *An Introduction to Statistical Modeling of Extreme Values*. New York: Springer.

Cook, D.I., Gebski, V.J., and Keech, A.C. (2004). Subgroup analysis in clinical trials. *The Medical Journal of Australia* **180**, 289–291.

Cook, R. (1986). Assessment of local inflence (with Discussion). *Journal of the Royal Statistical Society, Series B* **48**, 133–169.

Cook, R. and Weisberg, S. (1982). *Residuals and Inflence in Regression*. London: Chapman and Hall.

Copas, J. and Eguchi, S. (2005). Local model uncertainty and incomplete data bias (with Discussion). *Journal of the Royal Statistical Society Series B.* **67**, 459–512.

Corduneanu, A. and Bishop, C.M. (2001). Variational Bayesian model selection for mixture distributions. In *Artificial Intelligence and Statistics* (Eds. T. Jaakkola and T. Richardson). San Fransisco, CA: Morgan Kaufmann, pp. 27–34.

Cowans, P.J. (2004). Information retrieval using hierarchical Dirichlet processes. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York: Association for Computing Machinery, pp. 564–565.

Cox, D.R. (1975). A note on partially Bayes inference and the linear model. *Biometrika* **92**, 399-418.

Cox, T.F. and Cox, M.A.A. (2001). *Multidimensional Scaling*. Second Edition. Boca Raton, FL: Chapman & Hall / CRC Press.

Craiu, R.V., Rosenthal, J.S., and Yang, C. (2009). Learn from thy neighbor: Parallel-chain and regional adaptive MCMC. *Journal of the American Statistical Association* **104**, 1454–1466.

Cremers, M. (2002). Stock return predictability: A Bayesian model selection procedure. *Review of Financial Studies* **15**, 1223–1249.

Cressie, N.A. (1993). *Statistics for Spatial Data*. Second Edition. New York: Wiley.

Csiszár, I. (1967). Information-type measures of diffrence of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica* **2**, 299–318.

Cui, W. and George, E.I. (2008). Empirical Bayes vs. fully Bayes variable selection. *Journal of Statistical Planning and Inference* **138**, 888–900.

Czado, C. and Santner, T.J. (1992). The effect of link misspecification on binary regression inference. *Journal of Statistical Planning and Inference* **33**, 213–231.

D'Amico, A.V., Whittington, R., Malkowicz, S.B., Cote, K., Loffredo, M., Schultz, D., Chen, M.-H., Tomaszewski, J.E., Renshaw, A.A., Wein, A., and Richie, J. P. (2002), Biochemical outcome following radical prostatectomy or external beam radiation therapy for clinically localized prostate cancer in the PSA era. *Cancer* **95**, 281–286.

Damien, P., Wakefield, J.C., and Walker, S.G. (1999). Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. *Journal of the Royal Statistical Society, Series B* **61**, 331–344.

Dasgupta, A. and Raftery, A.E. (1998). Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association* **93**, 294–302.

Dass, S.C. and Lee, J. (2004). A note on the consistency of Bayes factors for testing point null versus non-parametric alternatives. *Journal of Statistical Planning Inference* **119**, 143–152.

Datta, G.S. (2009). Model-based approach to small area estimation. In *Sample Surveys: Inference and Analysis*. *Handbook of Statistics* **29B** (Eds. D. Pfeffermann and C.R. Rao). Amsterdam: North-Holland, pp. 251–288.

Datta, G.S., Fay, R.E., and Ghosh, M. (1991). Hierarchical and empirical multivariate Bayes analysis in small area estimation. In *Proceedings of the Seventh*

Annual Research Conference of the Bureau of the Census. Arlington, VA, pp. 63–79.

Datta, G.S. and Ghosh, M. (1991). Bayesian prediction in linear models: Applications to small area estimation. *Annals of Statistics* **19**, 1748–1770.

Datta, G.S., Ghosh, M., and Mukerjee, R. (2000). Some new results on probability matching priors. *Calcutta Statistical Association Bulletin* **50**, 179–192.

Datta, G.S. and Lahiri, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica* **10**, 613–627.

Datta, G.S. and Mukerjee, R. (2003). Probability matching priors for predicting a dependent variable with application to regression models. *Annals of the Institute of Statistical Mathematics* **55**, 1–6.

Datta, G.S. and Mukerjee, R. (2004). *Probability Matching Priors: Higher Order Asymptotics*. New York: Springer.

Datta, G.S., Mukerjee, R., Ghosh, M., and Sweeting, T.J. (2000). Bayesian prediction with approximate frequentist validity. *Annals of Statistics* **28**, 1414–1426.

Datta, G.S., Rao, J.N.K., and Smith, D.D. (2005). On measuring the variability of small area estimators under a basic area level model. *Biometrika* **92**, 183–196.

Datta, G.S. and Sweeting, T.J. (2005). Probability matching priors. In *Bayesian Thinking: Modeling and Computation*. *Handbook of Statistics* **25** (Eds. D.K. Dey and C.R. Rao). Amsterdam: North-Holland, pp. 91–114.

Dawid, A.P. and Stone, M. (1982). The functional-model basis of fiducial inference. *Annals of Statistics* **10**, 1054–1074.

de Finetti, B. (1940). The problem of full-risk insurances. Reprinted in: *Journal of Investment Management, 2006* **4**, 19–43.

DeGroot, M.H. (1970). *Optimal Statistical Decisions*. New York: McGraw-Hill.

de la Horra, J. and Rodriguez-Bernal, M.T. (2001). Posterior predicitive p-values: What they are and what they are not. *Test* **10**, 75–86.

Delyon, B., Lavielle, M., and Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *Annals of Statistics* **27**, 94–128.

DeMiguel, V., Garlappi, L., and Uppal, R. (2009). Optimal versus naive diversification: How inefficient is the $1/N$ portfolio strategy. *Review of Financial Studies* **22**, 1915–1953.

De Oliveira, V. (2007). Objective Bayesian analysis of spatial data with measurement error. *The Canadian Journal of Statistics* **35**, 283–301.

De Oliveira, V. (2010). Bayesian analysis of conditional autoregressive models. *Annals of the Institute of Statistical Mathematics*. In press.

De Oliveira, V., Kedem, B. and Short, D.A. (1997). Bayesian prediction of transformed Gaussian random fields. *Journal of the American Statistical Association* **92**, 1422–1433.

De Oliveira, V. and Song, J.J. (2008). Bayesian analysis of simultaneous autoregressive models. *Sankhyā, Series B* **70**, 323–350.

DeRobertis, L. and Hartigan, J.A. (1981). Bayesian inference using intervals of measure. *Annals of Statistics* **1**, 235–244.

Deroo, B.J. and Korach, K.S. (2006). Estrogen receptors and human disease. *The Journal of Clinical Investigation* **116**, 561–570.

De Santis, F. (2007). Using historical data for Bayesian sample size determination. *Journal of the Royal Statistical Society, Series A* **170**, 95–113.

De Santis, F., Perone Pacifico, M., and Sambcini, V. (2001). Optimal predictive sample size for case-control studies. *Tecnical Report, #17*, Dipartmento di Statistica, Probabilita e Statistiche Applicate, Universita di Roma.

DeSarbo, W.S., Fong, D.K.H., Liechty, J.C., and Saxton, K. (2004). A hierarchical Bayesian procedure for two-mode cluster analysis. *Psychometrika* **69**, 547–572.

DeSarbo, W.S., Kim, Y., Wedel, M., and Fong, D.K.H. (1998). A Bayesian approach to the spatial representation of market structure from consumer choice data. *European Journal of Operational Research* **111**, 285–305.

DeSarbo, W.S., Kim, Y., and Fong, D.K.H. (1999). A Bayesian multidimensional scaling procedure for the spatial analysis of Revealed Choice Data. *Journal of Econometrics* **89**, 79–108.

DeSarbo, W.S. and Rao, V.R. (1986). A constrained unfolding methodology for product positioning. *Marketing Science* **5**, 1–19.

DeSarbo, W.S. and Wu, J. (2001). The joint spatial representation of multiple variable batteries collected in marketing research. *Journal of Marketing Research* **38**, 244–253.

Dey, D.K., Ghosh, S.K., and Lou, K.R. (1996). On local sensitivity measures in Bayesian (with Discussion). In *Bayesian Robustness*, *IMS Lecture Notes-Monograph Series* **29** (Eds. J.O. Berger, B. Betró, E. Moreno, L. Pericchi, F. Ruggeri, G. Salinetti, and L. Wasserman). Beachwood, OH: Institute of Mathematical Statistics, pp. 21–39.

D'haeseleer, P., Liang, S., and Somogyi, R. (2000). Genetic network inference: From co-expression clustering to reverse engineering. *Bioinformatics* **16**, 707–726.

Diaconis, P. and Freedman, D. (1986). On the consistency of Bayes estimates. *Annals of Statistics* **14**, 1–26.

Diaconis, P., Goel, S., and Holmes, S. (2008). Horseshoes in multidimensional scaling and local kernel methods. *Annals of Applied Statistics* **2**, 777–807.

Diebolt, J. and Robert, C.P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society, Series B* **56**, 363–375.

Diggle, P.J., Morris, S.E. and Wakefield, J.C. (2000). Point-source modeling using matched case-control data. *Biostatistics* **1**, 89–105.

Diggle, P.J. and Ribeiro Jr., P.J. (2007). *Model-Based Geostatistics*. New York: Springer.

Diggle, P.J., Tawn, J.A., and Moyeed, R.A. (1998). Model-based geostatistics (with Discussion). *Applied Statistics* **47**, 299–350.

Dixon, W. and Massey, F. (1983). *Introduction to Statistical Analysis*. New York: McGraw-Hill.

Drummond, A.J., Ho, S.Y.W., Phillips, M.J., and Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. *PLoS Biology* **4**, 699–710, 2006.

Duda, R. and Hart, P., and Stork, D. (2000). *Pattern Classification*. New York: Wiley.

Duffie, D. (1996). State-space models of the term structure of interest rates. *Stochastic Analysis and Related Topics V* (Eds. H. Körezlioglu, B. Øksendal, and A. Üstünel). The Silivri Workshop, Boston: Birkhauser.

East V5.0 (2007). Cytel statistical software. Boston, MA.

Eaton, M.L. (1989). *Group Invariance Applications in Statistics*. Beachwood, Ohio: Institute of Mathematical Statistics.

Ecker, M.D. and Gelfand, A.E. (1997). Bayesian variogram modeling for an isotropic spatial process. *Journal of Agricultural, Biological and Environmental Statistics* **2**, 347–369.

Edgar, R.C. and Batzoglou, S. (2006). Multiple sequence alignment. *Current Opinion in Structural Biology* **16**, 368–373.

Edwards, S.V., Liu, L., and Pearl, D.K. (2007). High-resolution species trees without concatenation. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 5936–5941.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R.J. (2004). Least angle regression. *Annals of Statistics* **32**, 407–499.

Efron, B. and Morris, C. (1973). Stein's estimation rule and its competitors - an empirical Bayes approach. *Journal of the American Statistical Association* **68**, 117–130.

Efron, B. and Morris, C. (1977). Stein's paradox in statistics. *Scientific American* **236**, 119–127.

Efroymson, M.A. (1960). Multiple regression analysis. In *Mathematical Methods for Digital Computers*(Eds. A. Ralston and H.S. Wilf). New York: Wiley, pp. 191–203.

Eisen, M.B., Spellman, P.T., Brown, P.O., and Bolstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 14863–14868.

Ellis, B. and Wong, W.H. (2008). Learning causal Bayesian network structures from experimental data. *Journal of the American Statistical Association* **103**, 778–789.

Elmore, J.R., Heymann, H., Johnson, J., and Hewett, J.E. (1999). Preference mapping: Relating acceptance of "Creaminess" to a descriptive sensory map of semi-solid. *Food Quality and Preference* **10**, 465–475.

Ennis, S., Jomary, C., Mullins, R., Cree, A., Chen, X., MacLeod, A., Jones, S., Collins, A., Stone, E., and Lotery, A. (2008). Association between the SERPING1 gene and age-related macular degeneration: A two-stage case-control study. *The Lancet* **372**, 1828–1834.

Erosheva, E.A. (2002a). Grade of membership and latent structure models with application to disability survey data. *Unpublished Ph.D. Dissertation*, Department of Statistics, Carnegie Mellon University.

Erosheva, E.A. (2002b). Partial membership models with application to disability survey data. In *Proceedings of Conference on the New Frontiers of Statistical Data Mining* (Ed. H. Bozdogan). Boca Raton, FL: CRC Press, pp. 117–134.

Erosheva. E.A. (2003). Bayesian estimation of the grade of membership model. In *Bayesian Statistics* **7** (Eds. J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith, and M. West). Oxford: Oxford University Press, pp. 501–510.

Erosheva, E.A. and Fienberg, S.E. (2005). Bayesian mixed membership models for soft clustering and classification. In *Classification–the Ubiquitous Challenge* (Eds. C. Weihs and W. Gaul). New York: Springer, pp. 11–26.

Erosheva, E.A., Fienberg, S.E., and Joutard, C.J. (2007). Describing disability through individual-level mixture models for multivariate binary data. *Annals of Applied Statistics* **1**, 502–537.

Erosheva, E.A., Fienberg, S.E., and Lafferty, J. (2004). Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 11885–11892.

Escobar, M.D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90**, 577–588.

Everitt, B.S. and Hand, D.J. (1981). *Finite Mixture Distributions*. London: Chapman and Hall.

Everson, R. and Roberts, S. (2000). Inferring the eigenvalues of covariance matrices from limited, noisy data. *IEEE Transactions on Signal Processing* **48**, 2083–2091.

Fagan, W.F., Fortin, M.J., and Soykan, C. (2003). Integrating edge detection and dynamic modeling in quantitative analyses of ecological boundaries. *BioScience* **53**, 730–738.

Fahrmeir, L. and Lang, S. (2001). Bayesian inference for generalized additive mixed models based on Markov random field priors. *Journal of the Royal Statistical Society, Series C* **50**, 201–220.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360.

Fang, K.-T. and Wang, Y. (1994). *Number-theoretic Methods in Statistics*. London: Chapman and Hall.

Fanurik, D., Zeltzer, L.K., Roberts, M.C., and Blount, R.L. (1993). The relationship between children's coping styles and psychological interventions for cold pressor pain. *Pain* **53**, 213–222.

Fay, R.E., and Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association* **74**, 269–277.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution* **17**, 368–376.

Felsenstein, J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* **39**, 783–791.

Felsenstein, J. (2003). *Inferring Phylogenies*. Sunderland, MA: Sinauer Associates.

Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics* **1**, 209–230.

Ferguson, T.S. (1974). Prior distributions on spaces of probability measures. *Annals of Statistics* **2**, 615–629.

Ferguson, T.S. (1983). Bayesian density estimation by mixtures of Normal distributions. In *Recent Advances in Statistics* (Eds. M. Rizvi, J. Rustagi, and D. Siegmund). New York: Academic Press, pp. 287–302.

Ferguson, T.S. (1996). *A Course in Large Sample Theory*. London: Chapman and Hall.

Ferguson, T.S. and Gilstein, C.Z. (1985). A general investment model. *Working Paper*, Department of Mathematics, University of California at Los Angeles.

Fernandez, C., Ley, E., and Steel, M. (2001). Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics* **16**, 563–76.

Ferreira, M.A.R. and De Oliveira, V. (2007). Bayesian reference analysis for Gaussian Markov random fields. *Journal of Multivariate Analysis* **98**, 789–812.

Ferreira, M.A.R. and Suchard, M.A. (2008). Bayesian analysis of elapsed times in continuous-time Markov chains. *Canadian Journal of Statistics* **36**, 355–368.

Ferreira da Silva, A.R. (2009). Bayesian mixture models of variable dimension for image segmentation. *Computer Methods and Programs in Biomedicine* **94**, 1–14.

Figueiredo, M. (2003). Adaptive sparseness for supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **25**, 1150–1159.

Finley, A., Banerjee, S., and McRoberts, R. (2008). A bayesian approach to quantifying uncertainty in multi-source forest area estimates. *Environmental and Ecological Statistics* **15**, 241–258.

Flegal, J., Haran, M., and Jones, G.L. (2008). Markov chain Monte Carlo: Can we trust the third significant figure? *Statistical Science* **23**, 250–260.

Florens, J.P. , Richard, J.F., and Rolin, J.M. (1996). Bayesian encompassing specification tests of a parametric model against a nonparametric alternative. *Technical Report 96.08*, Institut de Statistique, Université Catholi que de Louvain.

Fong, D.K.H., DeSarbo, W.S., Park, J., and Scott, C.J. (2010). A Bayesian vector multidimensional scaling for the analysis of ordered preference data. *Journal of the American Statistical Association* **105**. In press.

Fonseca, T., Ferreira, M.A.R., and Migon, H. (2008). Objective Bayesian analysis for the Student-*t* regression model. *Biometrika* **95**, 325–333.

Forest, C., Stone, P., Sokolov, A., Allen, M., and Webster, M. (2002). Quantifying uncertainties in climate system properties with the use of recent climate observations. *Science* **295**, 113–117.

Fortini, S. and Ruggeri, F. (1994a). Concentation functions and Bayesian robustness (with Discussion). *Journal of Statisical Planning and Inference* **40**, 205–220.

Fortini, S. and Ruggeri, F. (1994b). On defiing neighbourhoods of measures through the concentration function. Sankhyā, Series A **56**, 444–457.

Foster, D.P. and George, E.I. (1994). The risk inflation criterion for multiple regression. *Annals of Statistics* **22**, 1947–1975.

Fotheringham, A.S., Brunsdon, C., and Charlton, M. (2006). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. West Sussex, UK: Wiley.

Fourdrinier, D., Kortbi, O., and Strawderman, W.E. (2008). Bayes minimax estimators of the mean of a variance mixture of multivariate normal distributions. *Journal of Multivariate Analysis* **99**, 74–93.

Fourdrinier, D. and Strawderman, W.E. (1996). A paradox concerning shrinkage estimators: Should a known scale parameter be replaced by an estimated value in the shrinkage factor. *Journal of Multivariate Analysis* **59**, 109–140.

Fourdrinier, D. and Strawderman, W.E. (2008). Generalized Bayes Minimax estimators of the location vector for spherically symmetric distributions. *Journal of Multivariate Analysis* **99**, 735–750.

Fourdrinier, D., Strawderman, W.E., and Wells, M.T. (1998). On the construction of Bayes minimax estimators. *Annals of Statistics* **26**, 660–671.

Fox, E.B., Sudderth, E.B., and Willsky, A.S. (2007). Hierarchical Dirichlet processes for tracking maneuvering targets. In *Proceedings of the International Conference on Information Fusion*, Quebec, Canada, July 2007.

Fraley, C. and Raftery, A.E. (1998). How many clusters? Which clustering method? - Answers via model-based cluster analysis. *The Computer Journal* **41**, 578–588.

Fraley, C. and Raftery, A.E. (2002). Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association* **97**, 611–631.

Fraley, C. and Raftery, A.E. (2007). Bayesian regularization for normal mixture estimation and model-based clustering. *Journal of Classification* **24**, 155–181.

Friel, N. and Pettitt, A.N. (2008). Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society, Series B* **70**, 589–607.

Frühwirth-Schnatter, S. (1994). Data augmentation and dynamic linear models. *Journal of Time Series Analysis* **15**, 183–202.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. New York: Springer.

Furnival, G.M. and Wilson, R.W. (1974). Regression by leaps and bounds. *Technometrics* **16**, 499–511.

Gail, M.H., Pee, D., and Carroll, R.J. (2001). Effects of violations of assumptions on likelihood methods for estimating the penetrance of an autosomal dominant mutation from kin-cohort studies. *Journal of Statistical Planning and Inference* **96**, 167–177.

Galtier, N., Gascuel, O., and Jean-Marie, A. (2005). Markov models in molecular evolution. In *Statistical Methods in Molecular Evolution* (Ed. R. Nielsen). New York: Springer, pp. 3–25.

Gamerman, D. and Lopes, H. (2007). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Second Edition. Baton Rouge: Chapman & Hall/CRC.

Ganesh, N. and Lahiri, P. (2008). A new class of average moment matching priors. *Biometrika* **95**, 514–520.

Garktze, E. (2007). The capitalist peace. *American Journal of Political Science* **51**, 166–191.

Gaspari, G. and Cohn, S.E. (1999). Construction of correlation functions in two and three dimensions. *Quarterly Journal of the Royal Metereological Society* **125**, 723–757.

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association* **70**, 320–328.

Geisser, S. (1993). *Predictive Inference: An Introduction*. Boca Raton, FL: CRC Press.

Geisser, S. and Cornfield, J. (1963). Posterior distributions for multivariate normal parameters. *Journal of the Royal Statistical Society, Series B* **25**, 368–376.

Gelfand, A.E. and Dey, D.K. (1991). On Bayesian robustness of contaminated classes of priors. *Statistics & Decisions* **9**, 63–80.

Gelfand, A.E. and Dey, D.K. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society, Series B* **56**, 501–514.

Gelfand, A.E., Dey, D.K., and Chang, H. (1992). Model determinating using predictive distributions with implementation via sampling-based methods (with Discussion). In *Bayesian Statistics* **4** (Eds. J.M. Bernado, J.O. Berger, A.P. Dawid, and A.F.M. Smith). Oxford: Oxford University Press, pp. 147–167.

Gelfand, A.E. and Ghosh, S.K. (1998). Model choice: A minimum posterior predictive loss approach. *Biometrika* **85**, 1–11.

Gelfand, A.E., Kim, H.K., Sirmans, C.F., and Banerjee, S. (2003). Spatial modelling with spatially varying coefficient processes. *Journal of the American Statistical Association* **98**, 387–396.

Gelfand, A.E., Kottas, A., and MacEachern, S.N. (2005). Bayesian nonparametric spatial modelling with Dirichlet process mixing. *Journal of the American Statistical Association* **100**, 1021–1035.

Gelfand, A.E., Schmidt, A.M., Banerjee, S., and Sirmans, C.F. (2004). Nonstationary multivariate process modelling through spatially varying coregionalization (with Discussion). *Test* **13**, 1–50.

Gelfand, A.E. and Smith, A.F.M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398–409.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (2003). *Bayesian Data Analysis*. Second Edition. Boca Raton, FL: Chapman & Hall / CRC Press.

Gelman, A. and King, G. (1990). Estimating the electoral consequences of legislative redistricting. *Journal of the American Statistical Association* **85**, 274–282.

Gelman, A. and King, G. (1994). A unified model for evaluating electoral systems and redistricting plans. *American Journal of Political Science* **38**, 514–554.

Gelman, A., King, G., and Liu, C. (1998). Multiple imputation for multiple surveys. *Journal of the American Statistical Association* **93**, 846–874.

Gelman, A. and Little, T.C. (1997). Poststratification into many categories using hierarchical logistic regression. *Survey Methodology* **23**, 127–135.

Gelman, A. and Meng, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science* **13**, 163–185.

Gelman, A., Meng, X.-L., and Stern, H.S. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with Discussion). *Statistica Sinica* **6**, 733–807.

Gelman, A., Park, D., Shor, B., Bafumi, J., and Cortina, J. (2008a). *Rich State, Poor State, Red State, Blue State: Why Americans Vote the Way They Do*. Princeton, NJ: Princeton University Press.

Gelman, A. and Raghunathan, T.E. (2001). Using conditional distributions for missing-data imputation. *Statistical Science* **15**, 268–269.

Gelman, A., Shor, B., Bafumi, J., and Park, D. (2008b). Rich state, poor state, red state, blue state: What's the matter with Connecticut? *Quarterly Journal of Political Science* **2**, 345–367.

Genkin, A., Lewis, D.D., and Madigan, D. (2007). Large-scale Bayesian logisitic regression for text categorization. *Technometrics* **49**, 291–304.

George, E.I. (1986a). Minimax multiple shrinkage estimators. *Annals of Statistics* **14**, 188-205.

George, E.I. (1986b). Combining minimax shrinkage estimators. *Journal of the American Statistical Association* **81**, 437–445.

George, E.I. (1986c). A formal Bayes multiple shrinkage estimator. *Communications in Statistics: Part A - Theory and Methods (Special Issue "Stein-type Multivariate Estimation")* **15**, 2099–2114.

George, E.I. (1999). Discussion of Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T, Bayesian model averaging: A tutorial. *Statistical Science* **14**, 409–412.

George, E.I. (2000). The variable selection problem. *Journal of the American Statistical Association* **95**, 1304–1308.

George, E.I. and Foster, D.P. (2000). Calibration and empirical Bayes variable selection. *Biometrika* **87**, 731–747.

George, E.I., Liang, F., and Xu, X. (2006). Improved minimax prediction under Kullback-Leibler loss. *Annals of Statistics* **34**, 78–91.

George, E.I. and McCulloch, R.E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* **88**, 881–889.

George, E.I. and McCulloch, R.E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica* **7**, 339–374.

George, E.I. and Xu, X. (2008). Predictive density estimation for multiple regression. *Econometric Theory* **24**, 528–544.

Geweke, J. (1996). Variable selection and model comparison in regression. In *Bayesian Statistics* **5** (Eds. J.M. Bernado, J.O. Berger, A.P. Dawid, and A.F.M. Smith). Oxford: Oxford University Press, pp. 609–620.

Ghosal, S., Ghosh, J.K., and Ramamoorthi, R.V. (1999a). Consistency issues in Bayesian nonparametrics. In *Asymptotic Nonparametrics and Time Series: A Tribute to Madan Lal Puri* (Ed. S. Ghosh). New York: Marcel Dekker, pp. 639–667.

Ghosal, S., Ghosh, J.K., and Ramamoorthi, R.V. (1999b). Posterior consistency of Dirichlet mixtures in density estimation. *Annals of Statistics* **27**, 143–158.

Ghosal, S., Ghosh, J.K., and Samanta, T. (1995). On convergence of posterior distribution. *Annals of Statistics* **23**, 2145–2152.

Ghosal, S., Ghosh, J.K., van der Vaart, A.W. (2000). Convergence rates of posterior distributions. *Annals of Statistics* **28**, 500–531.

Ghosal, S., Lember, J., and van der Vaart, A.W. (2008). Nonparametric Bayesian model selection and averaging. *Electronic Journal of Statistics* **2**, 63–89.

Ghosal, S. and van der Vaart, A.W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Annals of Statistics* **29**, 1233–1263.

Ghosal, S. and van der Vaart, A.W. (2007). Posterior convergence rates of Dirichlet mixtures of normal distributions for smooth densities. *Annals of Statistics* **35**, 192–223.

Ghosh, J.K. (1994). *Higher Order Asymptotics*. NSF-CBMS Regional Conference Series, Vol. 5. Beachwood, OH: Institute of Mathematical Statistics.

Ghosh, J.K. and Mukerjee, R. (1992). Non-informative priors (with discussion). In: *Bayesian Statistics* **4** (Eds.: J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith). Oxford: Oxford University Press, pp. 195–210.

Ghosh, J.K. and Ramamoorthi, R.V. (2003). *Bayesian Nonparametrics*. New York: Springer.

Ghosh, M. and Chen, M.-H. (2002). Bayesian inference for matched case-control studies. *Sankhyā, Series B* **64**, 107–127.

Ghosh, M., Chen, M.-H., Ghosh, A., and Agresti, A. (2000a). Hierarchical Bayesian analysis of binary matched pairs data. *Statistica Sinica* **10**, 647–657.

Ghosh, M., Ghosh, A., Chen, M.-H., and Agresti, A. (2000b). Noninformative priors for one-parameter item response models. *Journal of Statistical Planning and Inference* **88**, 99–115.

Ghosh, M., Mergel, V., and Datta, G.S. (2008). Estimation, prediction and the Stein phenomenon under divergence loss. *Journal of Multivariate Analysis* **99**, 1941–1961.

Ghosh, M., Mergel, V., and Liu, R. (2009). A general divergence criterion for prior selection. *Annals of the Institute of Statistical Mathematics*. In press.

Ghosh, M. and Mukerjee, R. (1998). Recent developments on probability matching priors. In *Applied Statistical Science III* (Eds. S.E. Ahmed, M. Ahsanullah, and B.K. Sinha). New York: Nova Science Publishers, pp. 227–252.

Ghosh, M., Zhang, L., and Mukherjee, B. (2006). Equivalence of posteriors in the Bayesian analysis of the multinomial-Poisson transformation. *Metron - International Journal of Statistics* **LXIV**, 19–28.

Gifi, A. (1990). *Nonlinear Multivariate Analysis*. Chichester, England: Wiley.

Gilks, W.R., Best, N.G., and Tan, K. (1995). Adaptive rejection Metropolis sampling within Gibbs sampling (Corr: 97V46, pp. 541–542 with R.M. Neal). *Applied Statistics* **44**, 455–472.

Gilks, W.R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics* **41**, 337–348.

Gillespie, J.H. (1991). *The Causes of Molecular Evolution*. New York: Oxford University Press.

Giron, F.J., Martinez, M.L., and Moreno, E. (2003). Bayesian analysis of matched pairs. *Journal of Statistical Planning and Inference* **113**, 49–66.

Gleser, L. (1979). Minimax estimation of a normal mean vector when the covariance matrix is unknown. *Annals of Statistics* **7**, 838–846.

Gnedin, A. and Pitman, J. (2006). Exchangeable Gibbs partitions and Stirling triangles. *Journal of Mathematical Sciences*, **138**, 5674–5685.

Gneiting, T. (2002). Nonseparable, stationary covariance functions for space-time data. *Journal of the American Statistical Association*, **97**, 590–601.

Godambe, V.P. (1976). Conditional likelihood and unconditional optimum estimating equations. *Biometrika* **63**, 277–284.

Goes, M., Urban, N., Tonkonojenkov, R., Haran, M., and Keller, K. (2009). The skill of different ocean tracers in reducing uncertainties about projections of the Atlantic Meridional Overturning Circulation. *Technical Report*, Department of Geosciences, Pennsylvania State University.

Godsill, S., Doucet, A., and West, M. (2004). Monte Carlo smoothing for nonlinear time series. *Journal of the American Statistical Association* **99**, 156–168.

Good, I.J. (1950). *Probability and Weighing of Evidence*. London: Charles Griffin.

Good, I.J. (1952). Rational decisions. *Journal of the Royal Statistical Society, Series B*, **14**, 107–114.

Good, I.J. (1967). A Bayesian significance test for multinomial distributions. *Journal of Royal Statistical Society, Series B* **29**, 399–431.

Gordon, N., Salmond, D., and Smith, A.F.M. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation, *IEE Proceedings* **F-140**, 107–113.

Gormley, I.C. and Murphy, T.B. (2007). A latent space model for rank data. In *Statistical Network Analysis: Models, Issues and New Directions* (Eds. E.M. Airoldi, D.M. Blei, S.E. Fienberg, A. Goldenberg, E.P. Xing, and A.X. Zheng). Berlin: Springer, pp. 90–107.

Gourieroux, C., Monfort, A., and Renault, E. (1993). Indirect Inference, Journal of Applied Econometrics. *Supplement: Special Issue on Econometric Inference Using Simulation Techniques* **8**, s85–s118.

Goutis, C. (1994). Ranges of posterior measures for some classes of priors with specifid moments. *International Statistical Review*, **62**, 245–256.

Gramacy, R., Lee, J.H., and Silva, R. (2009). On estimating covariances between many assets with histories of variable length. *Working Paper*, University of Cambridge.

Gramacy, R. and Pantaleo, E. (2009). Shrinkage regression for multivariate inference with missing data and an application to portfolio balancing. *Working Paper*, University of Cambridge.

Green, P.E. (1975). Marketing applications of MDS: Assessment and outlook. *Journal of Marketing* **39**, 24–31.

Green, P.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–722.

Green, P.J. and Richardson, S. (2001). Modelling heterogeneity with and without the Dirichlet process. *Scandinavian Journal of Statistics* **28**, 355–375.

Grendar, M. and Judge, G. (2009). Asymptotic equivalence of empirical likelihood and Bayesian map. *Annals of Statistics* **37**, 2445–2457.

Griffin, J. and Brown, P.J. (2005). Alternative prior distributions for variable selection with very many more variables than observations. *Technical Report*, Department of Statistics, University of Warwick.

Griffiths, T.L. and Ghahramani, Z. (2006). Infinite latent feature models and the Indian buffet process. In *Advances in Neural Information Processing Systems* **18** (Eds. Y. Weiss, B. Schölkopf, and J. Platt). Cambridge, MA: MIT Press, pp. 475–482.

Griffiths, T.L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 5228–5235.

Gron, A., Jorgensen, B., and Polson, N.G. (2004). Optimal portfolio choice and stochastic volatility. *Working Paper*, Booth School of Business, University of Chicago.

Grzebyk, M. and Wackernagel, H. (1994). Multivariate analysis and spatial/temporal scales: Real and complex models. In *Proceedings of the XVIIth International Biometrics Conference*, Hamilton, Ontario, Canada: International Biometric Society, pp. 19–33.

Guerrero, V.M. and Johnson, R.A. (1982). Use of the Box-Cox transformation with binary response models. *Biometrika* **69**, 309–314.

Guindani, M. and Gelfand, A.E. (2006). Smoothness properties and gradient analysis under spatial Dirichlet process models. *Models and Computing in Applied Probability* **8**, 159–189.

Guo, H., Weiss, R.E., Gu, X., and Suchard., M.A. (2007). Time squared: Repeated measures on phylogenies. *Molecular Biology and Evolution* **24**, 352–362.

Gustafson, P. (1994). Local sensitivity of posterior expectations. *Unpublished Ph.D. Dissertation*, Department of Statistics, Carnegie Mellon University.

Gustafson, P. (1996a). Local sensitivity of inferences to prior marginals. *Journal of the American Statistical Association* **91**, 774–781.

Gustafson, P. (1996b). Local sensitivity of posterior expectations. *Annals of Statistics* **24**, 174–195.

Gustafson, P. (2000). Local robustness in Bayesian analysis. In *Robust Bayesian Analysis Lecture Notes in Statistics* **152**. (Eds. D. Ríos Insua and F. Ruggeri). New York: Springer, pp. 71–88.

Gustafson, P. (2005). On model expansion, model contraction, identifiability and prior information: Two illustrative scenarios involving mismeasured variables. *Statistical Science* **20**, 111–140.

Gustafson, P., Le, N.D., and Vallee, M. (2002). A Bayesian approach to case-control studies with errors in covariables. *Biostatistics* **3**, 229–243.

Gustafson, P. and Wasserman, L. (1995). Local sensitivity diagnostics for Bayesian inference. *Annals of Statistics* **23**, 2153–2167.

Haario, H., Saksman, E., and Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernoulli* **2**, 223–242.

Haario, H., Saksman, E., and Tamminen, J. (2005). Componentwise adaptation for high dimensional MCMC. *Computational Statistics* **20**, 265–273.

Haldane, J.B.S. (1954). An exact test for randomness in mating. *Journal of Genetics* **52**, 631–635.

Handcock, M.S. and Stein, M.L. (1993). A Bayesian analysis of kriging. *Technometrics* **35**, 403–410.

Hans, C. (2008). Bayesian Lasso regression. *Technical Report*, Department of Statistics, Ohio State University.

Hans, C., Dobra, A., and West, M. (2007). Shotgun stochastic search in regression with many predictors. *Journal of the American Statistical Association* **102**, 507–516.

Harvey, C.R., Liechty, J.C., Liechty, M.W., and Müller, P. (2004). Portfolio selection with higher moments. *Technical Report*, Department of Statistical Science, Duke University.

Hartigan, J.A. (1998). The maximum likelihood prior. *Annals of Statistics* **26**, 2083–2103.

Hastie, T. and Stuetzle, W. (1989). Principal curves. *Journal of the American Statistical Association* **84**, 502–516.

Hastie, T., Tibshirani, R.J., and Friedman, J.H. (2001). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York: Springer.

Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

Heagerty, P. and Lele, S. (1998). A composite likelihood approach to binary spatial data. *Journal of the American Statistical Association* **93**, 1099–1111.

Hedeker, D., Gibbons, R.D., and Waternaux, C. (1999). Sample size estimation for longitudinal designs with attrition: Comparing time-related contrasts between two groups. *Journal of Educational and Behavioral Statistics* **24**, 70–93.

Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Kallioniemi, O.P., Wilfond, B., Borg, A., Trent, J., Raffeld, M., Yakhini, Z., Ben-Dor, A., Dougherty, E., Kononen, J., Bubendorf, L., Fehrle, W., Pittaluga, S., Gruvberger, S., Loman, N., Johannsson, O., Olsson, H., and Sauter, G. (2001). Gene-expression profiles in hereditary breast cancer. *New England Journal of Medicine* **344**, 539–48.

Herche, J. and Swenson, M.J. (1991). Multidimensional scaling: A marketing research tool to evaluate faculty performance in the classroom. *Journal of Marketing Education* **13**, 14–20.

Hervé, A. (2007). Distance. In *Encyclopedia of Measurement and Statistics* (Ed. N.J. Salkind). Thousand Oaks, CA: Sage, pp. 270–275.

Higdon, D., Gattiker, J., Williams, B., and Rightley, M. (2008). Computer model calibration using high-dimensional output. *Journal of the American Statistical Association* **103**, 570–583.

Hinton, G., Dayan, P., and Revow, M. (1997). Modeling the manifolds of images of handwritten digits. *IEEE Transactions of Neural Networks* **8**, 65–74.

Hjort, N.L. (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. *Annals of Statistics* **18**, 1259–1294.

Hjort, N.L., Dahl, F., and Steinbakk, G. (2006). Post-processing posterior predictive p-values. *Journal of the American Statistical Association* **101**, 1157–1174.

Hobert, J.P. and Casella, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association* **91**, 1461–1473.

Hobert, J.P. and Marchev, D. (2008). A theoretical comparison of the data augmentation, marginal augmentation and PX-DA algorithms. *Annals of Statistics* **36**, 532–554.

Hoeting, J.A., Madigan, D., Raftery, A.E., and Volinsky, C.T. (1999). Bayesian model averaging: A tutorial (with Discussion). *Statistical Science* **14**, 382–417.

Hoeting, J.A., Raftery, A.E., and Madigan, D. (1996). A method for simultaneous variable selection and outlier identification in linear regression. *Computational Statistics and Data Analysis* **22**, 251–270.

Holder, M. and Lewis, P.O. (2003). Phylogeny estimation: Traditional and Bayesian approaches. *Nature Reviews Genetics* **4**, 275–284.

Hore, S., Lopes, H., and McCulloch, R.E. (2009). Put option implied risk-premia in general equilibrium under recursive preferences. *Discussion Paper*, Booth School of Business, University of Chicago.

Hosmer, D.A. and Lemeshow, S. (2000). *Applied Logistic Regression*. Second Edition. New York: Wiley.

Huang, E., Chen, S., Dressman, H., Pittman, J., Tsou, M.H., Horng, C.F., Bild, A., Iversen, E.S., Liao, M., Chen, C.M., West, M., Nevins, J.R., and Huang, A.T. (2003a). Gene expression predictors of breast cancer outcomes. *The Lancet* **361**, 1590–1596.

Huang, E., Ishida, S., Pittman, J., Dressman, H., Bild, A., D'Amico, M., Pestell, R., West, M., and Nevins, J.R. (2003b). Gene expression phenotypic models that predict the activity of oncogenic pathways. *Nature Genetics* **34**, 226–230.

Huber, P. (1973). The use of choquet capacities in statistics. *Bulletin of the International Statistics Institute* **45**, 181–191.

Hubert, L.J. and Arabie P. (1985). Comparing partitions, *Journal of Classification* **2**, 63–76.

Hudson, R.R. (1983). Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, 23, 183–201.

Huelsenbeck, J.P., Larget, B., R.E. Miller, R.E., and Ronquist, F. (2002). Potential applications and pitfalls of Bayesian inference of phylogeny. *Systematic Biology* **51**, 673–688.

Huelsenbeck, J.P., Ronquist, F., Nielsen, R., and J.P. Bollback (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* **294**, 2310–2314.

Huelsenbeck, J.P. and Suchard, M.A. (2007). A nonparametric method for accommodating and testing across-site rate variation. *Systematic Biology* **56**, 975–987.

Huson, D.H. and Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* **23**, 254–267.

Hwang, J.T. (1982). Semi-tail upper bounds on the class of admissible estimators in discrete exponential families, with applications to Poisson and negative binomial distributions. *Annals of Statistics* **10**, 1137–1147.

Hwang, J.T. and Casella, G. (1982). Minimax confidence sets for the mean of a multivariate normal didtribution. *Annals of Statistics* **10**, 868–881.

Ibragimov, I. and Hasminsky, R. (1973). On the information in a sample about a parameter. In *Proceedings of 2nd International Symposium on Information Theory* (Eds. N. Petrov and V. Csaki). Budapest: Akadémiai Kiadó, pp. 295–309.

Ibrahim, J.G. and Chen, M.-H. (2000). Power prior distributions for regression models. *Statistical Science* **15**, 46–60.

Ibrahim, J.G., Chen, M.-H., and Lipsitz, S.R. (1999). Monte Carlo EM for missing covariates in parametric regression models. *Biometrics* **55**, 591–596.

Ibrahim, J., Chen, M.-H., Lipsitz, S.R., and Herring, A. (2005). Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association* **100**, 332–346.

Ibrahim, J.G., Chen, M.-H., and Sinha, D. (2001). Criterion-based methods for Bayesian model assessment. *Statistica Sinica* **11**, 419–443.

Ibrahim, J.G., Zhu, H., and Tang, N. (2009). Bayesian local inflence for survival models. *Techinical Report*, Department of Biostatistics, University of North Carolina at Chapel Hill.

Ioannidis, J.P.A. (2005). Contradicted and initially stronger effects in highly cited clinical research. *Journal of the American Medical Association* **294**, 218–228.

Ishwaran, H. and James, L.F. (2003). Generalized weighted Chinese restaurant processes for species sampling mixture models. *Statistica Sinica* **13**, 1211–1235.

Jackman, S. (2004). What do we learn from graduate admissions committees?: A multiple-rater, latent variable model with incomplete discrete and continuous indicators. *Political Analysis* **12**, 400–424.

Jacquier, E., Johannes, M., and Polson, N.G. (2007). MCMC maximum likelihood for latent state models. *Journal of Econometrics* **137**, 615–640.

Jacquier, E., Kane, A., and Marcus, A. (2005). Optimal forecasts of long-term returns and asset allocation: Arithmetic, geometric or other means. *Journal of Financial Econometrics* **3**, 37–55.

James, L.F. (2008). Large sample asymptotics for the two-parameter Poisson-Dirichlet process. In *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh* (Eds. B. Clarke and S. Ghosal). Beachwood, OH: Institute of Mathematical Statistics, pp. 187–199.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* **1** (Ed. J. Neyman). Berkerly, CA: University of California Press, pp. 361–379.

Jang, G.H., Lee, J., and Lee, S. (2010). Posterior consistency of species sampling priors. *Statistica Sinica*. In press.

Jasra, A., Holmes, C., and Stephens, D. (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science* **20**, 50–67.

Jaynes, E. (2003). *Probability Theory*. Cambridge: Cambridge University Press.

Jeffreys, H. (1961). *Theory of Probability*. Third Edition. Oxford: Clarendon Press.

Ji, C. and Schmidler, S.C. (2009). Adaptive Markov chain Monte Carlo for Bayesian variable selection. *Technical Report*, Department of Statistical Science, Duke University.

Ji, C., Shen, H., and West, M. (2009). Monte carlo variational approximation of marginal likelihoods. *Technical Report*, Department of Statistical Science, Duke University.

Jobson, J. and Korkie, R. (1980). Estimation for Markowitz efficient portfolios. *Journal of the American Statistical Association* **75**, 544–554.

Johannes, M. and Polson, N.G. (2009). MCMC methods for financial econometrics. In *Handbook of Financial Econometrics* **2** (Eds. Y. Ait-Sahalia and L. Hansen). Oxford: Elsevier, pp. 1–72.

Johannes, M., Korteweg, A., and Polson, N.G. (2009). Sequential predictive regressions. *Working Paper*, Booth School of Business, University of Chicago.

Johannes, M., Polson, N.G., and Stroud, J.R. (2009). Optimal filtering of jump diffusions: Extracting latent states from asset prices. *Review of Financial Studies* **22**, 2759–2799.

Johnson, R.M. (1971). Market segmentation: A strategic management tool. *Journal of Marketing Research* **8**, 13–18.

Johnson, V.E. and Albert, J.H. (1999). *Ordinal Data Modeling*. New York: Springer..

Johnson, R.A. and Wichern, D.W. (2007). *Applied Multivariate Statistical Analysis*. Sixth Edition. New York: Prentice Hall.

Johnson, W. (1985). Inflence measures for logistic regression: Another point of view. *Biometrika* **72**, 59–65.

Johnson, W. and Geisser, S. (1983). A predictive view of the detection and characterization of inflential observations in regression analysis. *Journal of the American Statistical Association* **78**, 137-144.

Johnson, W. and Geisser, S. (1985). Estimative inflence measures f or the multivariate general linear model. *Journal of Statistical Planning and Inference* **11**, 33–56.

Jolliffe, I.T. (1986). *Principal Component Analysis*. New York: Springer.

Jones, B. and Kenward, M.G. (1987). Modeling binary data from a three-period cross-over trial. *Statistics in Medicine* **6**, 555–564.

Jones, G.L., Haran, M., Caffo, B.S., and Neath, R. (2006). Fixed-width output analysis for Markov chain Monte Carlo. *Journal of the American Statistical Association* **101**, 1537–1547.

Jordan, M.I., Ghahramani, Z., Jaakkola, T., and Saul, K. (1999). An introduction to variational methods for graphical models. *Machine Learning* **37**, 183–233.

Jörnsten, R. and Keleş, S. (2008). Mixture models with multiple levels, with application to the analysis of multifactor gene expression data. *Biostatistics* **9**, 540–554.

Joutard, C.J., Airoldi, E.M., Fienberg, S.E., and Love, T.M. (2007). Discovery of latent patterns with hierarchical Bayesian mixed-membership models and the issue of model choice. Chapter 11 in *Data Mining Patterns: New Methods and Applications* (Eds. P. Poncelet, F. Masseglia, and M. Teisseire). Hershey, PA: Information Science Reference, pp. 240–275.

Kadane, J.B. and Vlachos, P. (2002). Hybrid methods for calculating optimal few-stage sequential strategies: Data monitoring for a clinical trial. *Statistics and Computing* **12**, 147–152.

Kandel, S., McCulloch, R.E., and Stambaugh, R.F. (1995). Bayesian inference and portfolio efficiency. *Review of Financial Studies* **8**, 1–53.

Kanehisa, M. and Goto, S. (2002). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* **28**, 27–30.

Kass, R.E. (1989). The geometry of asyptotic inference (with Discussion). *Statistical Science* **4**, 188–234.

Kass, R.E. and Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association* **90**, 773–795.

Kass, R.E., Tierney, L., and Kadane, J.B. (1989). Approximate methods for assessing inflence and sensitivity in Bayesian analysis. *Biometrika* **76**, 663–674.

Kass, R.E. and Vos, P. (1997). *Geometrical Foundations of Asymptotic Inference*. New York: Wiley.

Kass, R.E. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association* **90**, 928–934.

Kastellec, J.P., Gelman, A., and Chandler, J.P. (2009). Predicting and dissecting the seats-votes curve in the 2006 U.S. House election. *PS: Political Science and Politics* **41**, 729-732.

Kato, K. (2009). Improved prediction for a multivariate normal distribution with unknown mean and variance. *Annals of the Institute of Statistical Mathematics* **61**, 531–542.

Kennedy, M. and O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society, Series B* **63**, 425–464.

Kent, J.T. (1989). Continuity properties for random fields. *Annals of Probability* **17**, 1432–1440.

Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā, Series A* **62**, 49–66.

Key, R.M., Kozyr, A., Sabine, C.L., Lee, K., Wanninkhof, R., Bullister, J., Feely, R.A., Millero, F., Mordy, C., and Peng, T.-H. (2004). A global ocean carbon climatology: Results from Global Data Analysis Project (GLODAP). *Global Biogeochemical Cycles* **18**, GB4031.

Khoury, M.J., Beaty, T.H., and Cohen, B.H. (1993). *Fundamentals of Genetic Epidemiology*. Oxford: Oxford University Press.

Kim, I., Cohen, N.D., and Carroll, R.J. (2002). A method for graphical representation of effect heterogeneity by a matched covariate in matched case-control studies exemplified using data from a study of colic in horses. *American Journal of Epidemiology* **156**, 463–470.

Kim, S., Chen, M.-H., and Dey, D.K. (2008). Flexible generalized t-link models for binary response data. *Biometrika* **95**, 93–106.

Kim, S., Cohen, A.S., Baker, F.B., Subkoviak, M.J., and Leonard, T. (1994). An investigation of hierarchical Bayes procedures in item response theory. *Psychometrika* **59**, 405–421.

Kim, S.S., Guo, Y., and Agrusa, J. (2005). Preference and positioning analyses of overseas destinations by mainland Chinese outbound pleasure tourists. *Journal of Travel Research* **44**, 212–220.

Kim, Y. (1999). Nonparametric Bayesian estimators for counting processes. *Annals of Statistics* **27**, 562–588.

Kingman, J.F.C. (1967). Completely random measures. *Pacific Journal of Mathematics* **21**, 59-78.

Kingman, J.F.C. (1975). Random discrete distribution. *Journal of the Royal Statistical Society, Series B* **37**, 1–22.

Kingman, J.F.C. (1982). On the genealogy of large populations. *Journal of Applied Probability* **19**, 27–43.

Kitanidis, P.K. (1986). Parameter uncertainty in estimation of spatial functions: Bayesian analysis. *Water Resources Research* **22**, 449–507.

Kitchen, C.M.R., Kroll, J., Kuritzkes, D.R., Bloomquist, E.W., Deeks, S.G., and Suchard, M.A. (2009). Two-way Bayesian hierarchical phylogenetic models: An application to the co-evolution of gp120 and gp41 during and after enfuvirtide treatment. *Computational Statistics & Data Analysis* **53**, 766–775.

Kivinen, J., Sudderth, E.B., and Jordan, M.I. (2007). Learning multiscale representations of natural scenes using Dirichlet processes. In *IEEE International Conference on Computer Vision (ICCV)*, Rio de Janeiro, Brazil.

Kluge, A.G. (1989). A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boidae,Serpentes). *Systematic Zoology* **38**, 7–25.

Kneib, T. and Fahrmeir, L. (2006). Structured additive regression for categorical space-time data: A mixed model approach. *Biometrics* **62**, 109–118.

Kobayashi, K. and Komaki, F. (2008). Bayesian shrinkage prediction for the regression problem. *Journal of Multivariate Analysis* **99**, 1888–1905.

Koehler, A.B. and Murphree, E.S. (1988). A comparison of the Akaike and Schwarz criteria for selecting model order. *Applied Statistics* **37**, 187–195.

Kohonen, T. (1988). Learning vector quantization. *Neural Networks* **1**, 303.

Komaki, F. (1996). On asymptotic properties of predictive distributions. *Biometrika* **83**, 299–313.

Komaki, F. (2001). A shrinkage predictive distribution for multivariate normal observations. *Biometrika* **88**, 859–864.

Komaki, F. (2004). Simultaneous prediction of independent Poisson observables. *Annals of Statistics* **32**, 1744–1769.

Koonin, E.V. (2009). Darwinian evolution in the light of genomics. *Nucleic Acids Research* **37**, 1011–1034.

Kovach, C.R., Logan, B.R., Noonan, P.E., Schlidt, A.M., Smerz, J., Simpson, M., and Wells, T. (2006). Effects of the Serial Trial Intervention on discomfort and behavior of nursing home residents with dementia. *American Journal of Alzheimer's Disease and Other Dementias* **21**, 147–155.

Kuboki, H. (1998). Reference priors for prediction. *Journal of Statistical Planning and Inference* **69**, 295–317.

Kubokawa, T. and Strawderman, W.E. (2007). On Minimaxity and Admissibility of Hierarchical Bayes Estimators. *Journal of Multivariate Analysis* **98**, 829–851.

Lake, J.A. (1991). The order of sequence alignment can bias the selection of tree topology. *Molecular Biology and Evolution* **8**, 378–385.

Lan, K.K.G. and DeMets, D.L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* **70**, 659–663.

Lang, S. (1995). *Diffrential and Riemannian Manifolds*. Third Edition. New York: Springer.

Lartillot, N. and Phillipe, H. (2006). Computing Bayes factors using thermodynamic integration. *Systematic Biology* **55**, 195–207.

Lavine, M. (1991). Sensitivity in Bayesian statistics: The prior and the likelihood. *Journal of the American Statistical Association* **86**, 396–399.

Lavine, M. (1992). Some aspects of Polya tree distributions for statistical modelling. *Annals of Statistics* **20**, 1222–1235.

Lavine, M. (1994). More aspects of Polya tree distributions for statistical modelling. *Annals of Statistics* **22**, 1161–1176.

Lavine, M., Wasserman, L., and Wolpert, R.L. (1991). Bayesian inference with specifid marginals. *Jouranl of the American Statistical Association* **86**, 964–971.

Lax, J.R., and Phillips, J. (2009a). Gay rights in the states: Public opinion and policy responsiveness. *Technical Report*, Department of Political Science, Columbia University.

Lax, J.R., and Phillips, J. (2009b). How should we estimate public opinion in the states? *American Journal of Political Science* **53**, 107-121.

Lazar, N. (2003). Bayesian empirical likelihood. *Biometrika* **90**, 319–326.

Le, N.D. and Zidek, J. (1992). Interpolation with uncertain spatial covariances: A Bayesian alternative to kriging. *Journal of Multivariate Analysis* **43**, 351–374.

Leamer, E.E. (1978). *Specification Searches*. New York: Wiley.

Lee, J. (2007). Sampling methods of neutral to the right processes. *Journal of Computational and Graphical Statistics* **16**, 656–671.

Lee, M.D. (2001). Determining the dimensionality of multidimensional scaling representations for cognitive modeling. *Journal of Mathematical Psychology* **45**, 149–166.

Lee, S. and Zhu, H. (2000). Statistical analysis of nonlinear structural equation models with continuous and polytomous data. *British Journal of Mathematical and Statistical Psychology* **53**, 209–232.

Lefebvrea, G., Steele, R.J., and Vandal, A.C. (2010). A path sampling identity for computing the Kullback-Leibler and J-divergences. *Computational Statistics & Data Analysis*. In press.

Lehrach, W.P. and Husmeier, D. (2009). Segmenting bacterial and viral DNA sequence alignments with a trans-dimensional phylogenetic factorial hidden Markov model. *Journal of the Royal Statistical Society, Series C* **58**, 1–21.

Lenk, P.J. (1988). The logistic normal distribution for Bayesian, nonparametric, predictive densities. *Journal of the American Statistical Association* **83**, 509–516.

Lenk, P.J. (1991). Towards a practicable Bayesian nonparametric density estimator. *Biometrika* **78**, 531–543.

Leonard, T. and Novick, M.R. (1985). Bayesian inference and diagnostics for the three parameter logistic model. *ONR Technical Report 85-5*, CADA Research Group, University of Iowa.

Levy, M.S. and Perng, S.K. (1986). An optimal prediction function for the normal linear model. *Journal of the American Statistical Association* **81**, 196–198.

Lewis, K.F. and Whiteman, C.H. (2006). Empirical Bayesian density forecasting in Iowa and shrinkage for the Monte Carlo era. *Discussion Paper Series 1: Economic Studies 2006, 28*. Deutsche Bundesbank, Research Centre. http://econstor.eu/bitstream/10419/19657/1/200628dkp.pdf.

Lewis, R.J. and Berry, D.A. (1994). Group sequential clinical trials: A classical evaluation of Bayesian decision-theoretic designs. *Journal of the American Statistical Association* **89**, 1528–1534.

Ley, E. and Steel, M. (2007). Jointness in Bayesian variable selection with applications to growth regression. *Journal of Macroeconomics* **29**, 476–493.

Li, C. and Li, H. (2008). Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics* **24**, 1175–1182.

Li, F. and Zhang, N.R. (2008). Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *Technical Report*, Department of Statistics, Stanford University.

Li, H. and Gui, J. (2006). Gradient directed regularization for sparse Gaussian concentration graphs, with applications to inference of genetic networks. *Biostatistics* **7**, 302–317.

Li, H., Wei, Z., and Maris, J. (2009). A hidden Markov model approach for genome-wide association studies. *Biostatistics*. In press.

Li, J. (2005). Clustering based on a multilayer mixture model. *Journal of Computational and Graphical Statistics* **14**, 547–568.

Liang, F. and Barron, A.R. (2004). Exact minimax strategies for predictive density estimation, data compression and model selection. *IEEE Information Theory Transactions* **50**, 2708–2726.

Liang, F., Paulo, R., Molina, G., Clyde, M.A., and Berger, J.O. (2008). Mixtures of g-priors for Bayesian variable selection. *Journal of the American Statistical Association* **103**, 410–423.

Liang, P., Jordan, M.I., and Klein, D. (2010). Probabilistic grammars and hierarchical Dirichlet processes. In The Handbook of Applied Bayesian Analysis (Eds. T. O'Hagan and M. West). Oxford: Oxford University Press. In press.

Liechty, J.C., Liechty, M.W., and Müller, P. (2004). Bayesian correlation estimation. *Biometrika* **91**, 1–14.

Liechty, M.W., Liechty, J.C., and Müller, P. (2009). The shadow prior. *Journal of Computational and Graphical Statistics* **18**, 368–383.

Lijoi, A., Mena, R.H., and Prünster, I. (2005). Hierarchical mixture modeling with normalized inverse-Gaussian priors. *Journal of the American Statistical Association* **100**, 1278–1291.

Lijoi, A., Prünster, I., and Walker, S.G. (2005). On consistency of nonparametric normal mixtures for Bayesian density estimation. *Journal of the American Statistical Association* **100**, 1292–1296.

Lijoi, A., Prünster, I., and Walker, S.G. (2008). Bayesian nonparametric estimators derived from conditional Gibbs structures. *Annals of Applied Probability* **18**, 1519–1547.

Lilien, G.L. and Rangaswamy, A. (2004). *Marketing Engineering*. Revised Second Edition. Reading, MA: Addison-Wesley Publishing.

Lin, P. and Tsai, H. (1973). Generalize Bayes minimax estimators of the multivariate normal mean with unknown covariance matrix. *Annals of Statistics* **1**, 142–145.

Lindley, D.V. (1964). The Bayesian analysis of contingency tables. *Annals of Mathematical Statistics* **35**, 1622–1643.

Lindley, D.V. (1988). Statistical inference concerning Hardy-Weinberg equilibrium (with Discussion). In *Bayesian Statistics* **3** (Eds. J.M. Bernardo, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith). Oxford: University Press, pp. 307–326.

Lindley, D.V. and Smith, A.F.M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B* **34**, 1–41.

Lindsay, B.G. (1995). *Mixture Models: Theory, Geometry, and Applications*. Hayward, CA: Insitute of Mathematical Statistics.

Lipsitz, S.R., Parzen, M., and Ewell, M. (1998). Inference using conditional logistic regression with missing covariates. *Biometrics* **54**, 295–303.

Liseo, B. and Loperfido, N. (2006). A note on the reference prior for the scalar skew normal distribution. *Journal of Statistical Planning and Inference* **136**, 373–389.

Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. Second Edition. New York: Wiley.

Liu, A., Zhang, Y., Gehan, E., and Clarke, R. (2002). Block principal component analysis with application to gene microarray data application. *Statistics in Medicine* **21**, 3465–3474.

Liu, F., Bayarri, M.J., Berger, J.O., Paulo, R., and Sacks, J. (2008). A Bayesian analysis of the thermal challenge problem. *Computer Methods in Applied Mechanics and Engineering* **197**, 2457–2466.

Liu, F., Bayarri, M.J., and Berger, J.O. (2009). Modularization in Bayesian Analysis with emphasis on analysis of computer models. *Bayesian Analysis* **4**, 119–150.

Liu, G. and Liang, K.-Y. (1997). Sample size calculations for studies with correlated observations. *Biometrics* **53**, 937–947.

Liu, J., Gustafson, P., Cherry, N., and Burstyn, I. (2009). Bayesian analysis of a matched case-control study with expert prior information on both the misclassification of exposure and the exposure-disease association. *Statistics in Medicine* **28**, 3411–3423.

Liu, J. and West, M. (2001). Combined parameters and state estimation in simulation-based filtering. In *Sequential Monte Carlo Methods in Practice* (Eds. A. Doucet, N. de Freitas, and N. Gordon). New York: Springer, pp. 197–224.

Liu, J.S. (1994). The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association* **89**, 958–966.

Liu, J.S. (1996). Nonparametric hierarchical Bayes via sequential imputations. *Annals of Statistics* **24**, 911–930.

Liu, R. and Ghosh, M. (2009). Objective priors: A selective review. *Techinical Report*, Department of Statistics, University of Florida.

Lo, A.Y. (1984). On a class of Bayesian nonparametric estimates I: Density estimates. *Annals of Statistics* **12**, 351–357.

Lopes, H., Carvalho, C.M., Johannes, M., and Polson, N.G. (2010). Particle Learning for sequential Bayesian computation. In *Bayesian Statistics* **9** (Eds. J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith, and M. West). Oxford: Oxford University Press. In press.

Lopes, H. and Polson, N.G. (2010). Extracting SP500 and NASDAQ volatility: The credit crisis of 2007-2008. In *Handbook of Applied Bayesian Analysis* (Eds. A. O'Hagan and M. West). Oxford: Oxford University Press. In press.

Lopes, H. and West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica* **14**, 41–67.

Lord, F.M. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement* **13**, 517–548.

Lu, T., Pan, Y., Kao, S.-Y., Li, C., Kohane, I., Chan, J., and Yankner, B.A. (2004). Gene regulation and DNA damage in the aging human brain. *Nature* **429**, 883–891.

Lucas, J.E., Carvalho, C.M., Wang, Q., Bild, A., Nevins, J.R., and West, M. (2006). Sparse statistical modelling in gene expression genomics. In *Bayesian Inference for Gene Expression and Proteomics* (Eds. P.M.K.A. Do, P. Müller, and M. Vannucci). New York: Cambridge University Press, pp. 155–176.

Lucas, J.E., Carvalho, C.M., Chi, J.-T.A., and West, M. (2009). Cross-study projections of genomic biomarkers: An evaluation in cancer genomics. *PLoS One* **4**, e4523.

Lucas, J.E., Carvalho, C.M., and West, M. (2009). A Bayesian analysis strategy for cross-study translation of gene expression biomarkers. *Statistical Applications in Genetics and Molecular Biology* **8**, Article 11.

Lui, K.-J. and Cumberland, W.G. (1992). Sample size requirement for repeated measurements in continuous data. *Statistics in Medicine* **11**, 633–641.

Lunter, G., Miklós, I., Drummond, A., Jensen, J.L., and Hein, J. (2005). Bayesian co-estimation of phylogeny and sequence alignment. *BMC Bioinformatics* **6**, 83.

Luo, Z. and Wahba, G. (2007). Hybrid adaptive splines. *Journal of the American Statistical Association* **92**, 107–116.

MacEachern, S.N. (1998). Estimation of Bayes factors in mixture of Dirichlet process models. *Technical Report*, Department of Statistics, Ohio State University.

MacEachern, S.N. and Müller, P. (1998). Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics* **7**, 223–228.

MacEachern, S.N. and Müller, P. (2000). Efficient MCMC schemes for robust model extensions using encompassing Dirichlet process mixture models. In *Robust Bayesian Analysis Lecture Notes in Statistics* **152** (Eds. D. Ríos Insua and F. Ruggeri). New York: Springer, pp. 295–316.

Mackay, D.J.C. (1992). Bayesian interpolation. *Neural Computation* **4**, 415–447.

Madigan, D. and Raftery, A.E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of American Statistical Association* **89**, 1535–1546.

Madigan, D. and York, J. (1995). Bayesian graphical models for discrete data. *International Statistical Review* **63**, 215–232.

Majumdar, A., Munnekke, H., Gelfand, A.E., Banerjee, S., and Sirmans, C.F. (2006). Gradients in spatial response surfaces with applications to land-value prices. *Journal of Business and Economics Statistics* **24**, 77–90.

Maki, R.G., Wathen, J.K., Patel, S.R., Priebat, D.A., Okuno, S.H., Samuels, B., Fanucchi, M., Harmon, D.D., Schuetze, S.M., Reinke, D., Thall, P.F., Benjamin, R.S., Baker, L.H., and Hensley, M.L. (2007). An adaptively randomized phase ii study of gemcitabine and docetaxel versus gemcitabine alone in patients with metastatic soft-tissue sarcomas. *Journal of Clinical Oncology* **25**, 2755–2763.

Mallows, C.L. (1973). Some comments on C_p. *Technometrics* **15**, 661–676.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* **22**, 719–748.

Marden, J.I. (2000). Hypothesis testing: From p values to Bayes factors. *Journal of the American Statistical Association* **95**, 1316–1320.

Mardia, K.V., Kent, J.T., Goodall, C.R., and Little, J.A. (1996). Kriging and splines with derivative information. *Biometrika* **83**, 207–221.

Marin, J.-M. and Robert, C.P. (2007). *Bayesian Core*. New York: Springer.

Marjoram, P., Molitor, J., Plagnol, V., and Tavare, S. (2003). Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 15324–15328.

Markowitz, H.M. (1959). *Portfolio Selection: Efficient Diversification of Investments*. New York: Wiley.

Markowitz, H.M. (2006). de Finetti scoops Markowitz. *Journal of Investment Management* **4**, 5–18.

Marshall, R.J. (1988). Bayesian analysis of case-control studies. *Statistics in Medicine* **7**, 1223–1230.

Martin, A.D. and Quinn, K.Q. (2002). Dynamic ideal point estimation via Markov chain Monte Carlo for the U.S. Supreme Court, 1953-1999. *Political Analysis* **10**, 134–153.

Maruyama, Y. (2003a). Admissible minimax estimators of a mean vector of scale mixtures of multivariate normal distributions. *Journal of Multivariate Analysis* **84**, 274–283.

Maruyama, Y. (2003b). A robust generalized Bayes estimator improving on the James-Stein estimator for spherically symmetric distributions. *Statistics & Decisions* **21**, 69–77.

Maruyama, Y. and Strawderman, W.E. (2005). A new class of generalized Bayes minimax ridge regression estimators. *Annals of Statistics* **33**, 1753–1770.

Matsusaka, J.G. and McCarty, N.M. (2001). Political resource allocation: Benefits and costs of voter initiatives. *Journal of Law, Economics, and Organization* **17**, 413–448.

Mauldin, R.D., Sudderth, W.D., and Williams, S.C. (1992). Polya trees and random distributions. *Annals of Statistics* **20**, 1203–1221.

Maynard, P.V., Davies, C.J., Blamey, R.W., Elston, C.W., Johnson, J., and Griffiths, K. (1978). Relationship between oestrogen-receptor content and histological grade in human primary breast tumours. *British Journal of Cancer* **38**, 745–748.

McAuliffe, J., Blei, D.M., and Jordan, M.I. (2006). Nonparametric empirical Bayes for the Dirichlet process mixture model nonparametric empirical Bayes for the Dirichlet process mixture model. *Statistics and Computing* **16**, 5–14.

McCarthy, R., Bower, F., and Jesson, J. (1977). The fluorocarbon-ozone theory I. Production and releaseworld production and release of CCl3F and CCl2F2 (fluorocarbons 11 and 12) through 1975. *Atmospheric Environment (1967)* **11**, 491–497.

McCullagh, P. (1990). A note on partially Bayes inference for generalized linear models. *Technical Report 284*, Department of Statistics, The University of Chicago.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*. Second Edition. London: Chapman and Hall.

McCulloch, R.E. (1989). Local model inflence. *Journal of the American Statistical Association* **84**, 473–478.

McCulloch, R.E., Polson, N.G., and Rossi, P. (2000). A Bayesian analysis of the multinomial probit model with fully identified parameters. *Journal of Econometrics* **99**, 173–193.

McGrory, C.A. and Titterington, D.M. (2007). Variational approximations in Bayesian model selection for finite mixture distributions. *Computational Statistics and Data Analysis* **51**, 5352–5367.

McLachlan, G.J. and Basford, K.E. (1987). *Mixture Models: Inference and Applications to Clustering*. New York: Marcel Dekker.

McQuarrie, A.D.R. and Tsai, C.-L. (1998). *Regression and Time Series Model Selection*. Singapore: World Scientific.

McShane, L.M., Midthune, D.N., Dorgan, J.F., Freedman, L.S., and Carroll, R.J. (2001). Covariate measurement error adjustment for matched case-control studies. *Biometrics* **57**, 62–73.

McVinish R., Rousseau, J., and Mengersen, K. (2009). Bayesian goodness of fit testing with mixtures of triangular distributions. *Scandinavian Journal of Statistics* **36**, 337–354.

Ménard, S., Pupa, S.M., Campiglio, M., and Tagliabue, E. (2003). Biologic and therapeutic role of her2 in cancer. *Oncogene* **22**, 6570–6578.

Meng, X.-L. (1994). Posterior predictive p-values. *Annals of Statistics* **22**, 1142–1160.

Meng, X.-L., Lee, T.C.M., and Li, Z. (2009). What can we do when EM is not applicable? Self consistency principle for semi-parametric and non-parametric estimation with incomplete and irregularly spaced data. *Techinical Report*, Department of Statistics, Harvard University.

Meng, X.-L. and Wong, W.H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica* **6**, 831–860.

Merl, D., Chen, J.L.-Y., Chi, J.T., and West, M. (2009a). Integrative analysis of cancer gene expression studies using Bayesian latent factor modelling. *Annals of Applied Statistics*. In press.

Merl, D., Lucas, J.E., Nevins, J.R., Shen, H., and West, M. (2009b). Trans-study projection of genomic biomarkers using sparse factor regression models. In *The Handbook of Applied Bayesian Analysis* (Eds. A. O'Hagan and M. West). Oxford: Oxford University Press.

Merton, R.C. (1971). Optimum consumption and portfolio rules in a continuous-time model. *Journal of Economic Theory* **3**, 373–413.

Merton, R.C. (1972). An analytical derivation of the efficient portfolio frontier. *Journal of Financial and Quantitative Analysis* **7**, 1851–1872.

Miller, A.J. (2002). *Subset Selection in Regression*. Second Edition. Boca Raton, FL: Chapman & Hall / CRC Press.

Minka, T. (2000). Automatic choice of dimensionality for PCA. *Technical Report No. 514*, MIT Media Laboratory.

Minin, V.N., Dorman, K.S., Fang, F., and Suchard, M.A. (2005). Dual multiple change-point model leads to more accurate recombination detection. *Bioinformatics* **21**, 3034–3042.

Mislevy, R.J. and Bock, R.D. (1984). BILOG maximum likelihood item analysis and test scoring: Logistic model. Mooresville, ID: Scientific Software.

Mitas, L. and Mitasova, H. (1999). Spatial interpolation. In *Geographical Information Systems: Principles, Techniques, Management and Applications, GeoInformation International* (Eds. P.Longley, M.F. Goodchild, D.J. Maguire, and D.W. Rhind). New York: Wiley, pp. 481–492.

Moggs, J.G. and Orphanieds, G. (2001). Estrogen receptors: Orchestrators of pleiotropic cellular responses. *EMBO Reports* **2**, 775–781.

Molenberghs, G. and Rousseeuw, P.J. (1994). The shape of correlation matrices. *The American Statistician* **48**, 4.

Molina, G., Bayarri, M.J., and Berger, J.O. (2005). Statistical inverse analysis for a network microsimulator. *Technometrics* **47**, 388–398.

Møller, J. (2003). *Spatial Statistics and Computational Methods*. New York: Springer.

Møller, J., Pettitt, A.N., Reeves, R.W., and Berthelsen, K.K. (2006). An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika* **93**, 451–458.

Monni, S. and Li, H. (2008). Vertex clustering of graphs using reversible jump MCMC. *Journal of Computational and Graphical Statistics* **17**, 388–409.

Moon, H.R. and Schorfheide, F. (2004). Bayesian inference for econometric models using empirical likelihood functions. In *Econometric Society 2004 North American Winter Meetings*. Paper #284.

Moore, D.S. (2006). *The Basic Practice of Statistics*. Third Edition. New York: W. H. Freeman.

Moreno, E. (2000). Global Bayesian robustness for some classes of prior distributions. In *Robust Bayesian Analysis*. *Lecture Notes in Statistics* **152** (Eds. D. Ríos Insua and F. Ruggeri). New York: Springer, pp. 5–37.

Moreno, E. and Gonzalez, A. (1990). Empirical bayes analysis for ε-contaminated priors with shape and quantile constraints. *Brazilian Journal of Probability and Statistics* **4**, 177–200.

Moreno, E. and Pericchi, L. (1993). Bayesian robustness for hierarchical ε-contaminations models. *Journal of Statistical Planning and Inference* **37**, 159–168.

Morgan, B.J.T. (1983). Observations on quantitative analysis. *Biometrics* **39**, 879–886.

Morris, C. (1983). Parametric empirical Bayes confidence intervals. In *Scientific Inference, Data Analysis, and Robustness* (Eds. G.E.P. Box, T. Leonard, and C.F.J. Wu). New York: Academic Press, pp. 25–50.

Morrison, D.A. (2009). Why would phylogeneticists ignore computerized sequence alignment? *Systematic Biology*. In press.

Mossel, E. and Vigoda, E. (2005). Phylogenetic MCMC algorithms are misleading on mixtures of trees. *Science* **309**, 2207–2209.

Mukherjee, B. and Chatterjee, N. (2008). Exploiting gene-environment independence for analysis of case-control studies: An empirical Bayes-type shrinkage estimator to trade-off between bias and efficiency. *Biometrics* **64**, 685–694.

Mukherjee, B., Liu, I., and Sinha, S. (2007). Analysis of matched case-control data with ordinal disease states: Possible choices and comparisons. *Statistics in Medicine* **26**, 3240–3257.

Mukherjee, S. and Speed, T.P. (2008). Network inference using informative priors. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 14313–14318.

Muller, K.E., LaVange, L. M., Ramey, S.L., and Ramey, C.T. (1992). Power calculations for general linear multivariate models including repeated measures applications. *Journal of the American Statistical Association* **87**, 1209–1226.

Müller, P., Erkanli, A., and West, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika* **83**, 67–79.

Müller, P. and Parmigiani, G. (1995). Optimal design via curve fitting of Monte Carlo experiments. *Journal of the American Statistical Association* **90**, 1322–1330.

Müller, P., Parmigiani, G., Schildkraut, J., and Tardella, L. (1999). A Bayesian hierarchical approach for combining case-control and prospective studies. *Biometrics* **55**, 858–866.

Müller, P. and Roeder, K. (1997). A Bayesian semiparametric model for case-control studies with errors in variables. *Biometrika* **84**, 523–537.

Müller, P. and Quintana, F.A. (2004). Nonparametric Bayesian data analysis. *Statistical Science* **19**, 95–110.

Murray, G.D. (1977). A note on the estimation of probability density functions. *Biometrika* **64**, 150–152.

Murray, I. (2007). Advances in Markov chain Monte Carlo methods. *Unpublished Ph.D. Dissertation*, Gatsby Computational Neuroscience Unit, University College London, 17 Queen Square, London WC1N 3AR, United Kingdom.

Nandram, B. and Sedransk, J. (1993). Bayesian predictive inference for longitudinal sample surveys. *Biometrics* **49**, 1045–1055.

Navarrete, C., Quintana, F.A., and Müller, P. (2008). Some issues on nonparametric bayesian modeling using species sampling models. *Statistical Modelling* **8**, 3–21.

Neal, R. (1994). Contribution to the discussion of "approximate bayesian inference with the weighted likelihood bootstrap" by Michael A. Newton and Adrian E. Raftery. *Journal of the Royal Statistical Society, Series B* **56**, 41–42.

Neal, R. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics* **9**, 249–265.

Nelder, J.A. and Wedderburn, R.M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A* **135**, 370–384.

Nesse, R.M. and Stearns, S.C. (2008). The great opportunity: Evolutionary applications to medicine and public health. *Evolutionary Applications* **1**, 28–24.

Neville, J., Simsek, O., Jensen, D., Komoroske, J., Palmer, K., and Goldberg, H. (2005). Using relational knowledge discovery to prevent securities fraud. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* **11**.

Newton, M.A., Quintana, F.A., den Boon, J.A., Sengupta, S., and Ahlquist, P. (2007). Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. *Annals of Applied Statistics* **1**, 85–106.

Newton, M.A. and Raftery, A.E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society, Series B* **56**, 3–48.

Ng, A.Y. (1997). Preventing "overfitting" of cross-validation data. In *Proceedings of the Fourteenth International Conference on Machine Learning*. San Francisco, CA: Morgan Kaufmann Publishers Inc., pp-245–253.

Ng, V.M. (1980). On the estimation of parametric density functions. *Biometrika* **67**, 505–506.

Nobile, A. and Fearnside, A.T. (2007). Bayesian finite mixtures with an unknown number of components: The allocation sampler. *Statistics and Computing* **17**, 147–162.

Nott, D.J. and Green, P.J. (2004). Bayesian variable selection and the Swendsen-Wang algorithm. *Journal of Computational and Graphical Statistics* **13**, 141–157.

Nott, D.J. and Kohn, R. (2005). Adaptive sampling for Bayesian variable selection. *Biometrika* **92**, 747–763.

Nurminen, M. and Mutanen, P. (1987). Exact Bayesian analysis of two proportions. *Scandinavian Journal of Statistics* **14**, 67–77.

Oakes, D. (1986). Semiparametric inference in a model for association in bivariate survival data. *Biometrika* **73**, 353–361.

Oakes, D. (1989). Bivariate survival models induced by frailties. *Journal of the American Statistical Association* **84**, 487–493.

Oakley, J.E. and O'Hagan, A. (2004). Probabilistic sensitivity analysis of complex models: A Bayesian approach. *Journal of the Royal Statistical Society, Series B* **66**, 751–769.

O'Brien, P.C. and Fleming, T.R. (1979). A multiple testing procedure for clinical trials. *Biometrics* **35**, 549–556.

Oh, M. and Raftery, A.E. (2001). Bayesian multidimensional scaling and choice of dimension. *Journal of the American Statistical Association* **96**, 1031–1044.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society, Series B* **57**, 99–138.

O'Hagan, A. (2009). Reification and true parameters: Discussion of Goldstein and Rougier. *Journal of Statistical Planning and Inference* **139**, 1240–1242.

O'Malley, A.J. and Zaslavsky, A.M. (2005). Cluster-level covariance analysis for survey data with structured nonresponse. *Technical Report*, Department of Health Care Policy, Harvard Medical School.

Owen, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75**, 237–249.

Owen, A. (1990). Empirical likelihood for confidence regions. *Annals of Statistics* **18**, 90–120.

Owen, A. (1991). Empirical likelihood for linear models. *Annals of Statistics* **19**, 1725–1747.

Owen, R. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association* **70**, 351–356.

Pagel, M. and Meade, A. (2004). A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Systematic Biology* **53**, 571–581.

Paik, M.C. and Sacco, R. (2000). Matched case-control data analyses with missing covariates. *Applied Statistics* **49**, 145–156.

Pan, D.W, and Baker, J.A.W. (1998). Perceptual mapping of banned substances in athletics: Gender- and sport-defined differences. *Journal of Sports & Social Issues* **22**, 170–182.

Pardoe, I. (2001). A Bayesian sampling approach to regression model checking. *Journal of Computational and Graphical Statistics* **10**, 617–627.

Park, J., DeSarbo, W.S., and Liechty, J.C. (2008). A hierarchical Bayesian multidimensional scaling methodology for accommodating both structural and preference heterogeneity. *Psychometrika* **73**, 451–472.

Park, T. and Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association* **103**, 681–686.

Pasarica, C. and Gelman, A. (2010). Adaptively scaling the Metropolis algorithm using expected squared jumped distance. *Statistica Sinica*. In press.

Patrick, G.F., Blake, B.F., and Whitaker, S.H. (1983). Farmers' goals: Uni- or multidimensional? *American Journal of Agricultural Economics* **65**, 315–320.

Patz, R.J. and Junker, B.W. (1999a). A straightforward approach to Markov chain Monte Carlo methods for item response models. Journal of Educational and Behavioral Statistics **24**, 146–178.

Patz, R.J. and Junker, B.W. (1999b). Applications and extensions of MCMC in IRT: Multiple item types, missing data and rated responses. *Journal of Educational and Behavioral Statistics* **24**, 342–366.

Paulo, R. (2005). Default priors for Gaussian processes. *Annals of Statistics* **33**, 556–582.

Peng, F. and Dey, D.K. (1995). Bayesian analysis of outlier problems using divergence measures. *Canadian Journal of Statistics* **23**, 199–213.

Peng, J., Wang, P., Zhou, N., and Zhu, J. (2009). Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*. In press.

Perez, C., Martín, J., and Rufo, M. (2006). MCMC-based local parametric sensitivity estimations. *Computational Statistics and Data Analysis* **51**, 823–835.

Pericchi, L. (2009). How large should be the training sample in testing hypotheses? *Technical Report*, Department of Mathematics, University of Puerto Rico at Rio.

Pericchi, L. and Smith, A.F.M. (1992). Exact and approximate posterior moments for a normal location parameter. *Journal of the Royal Statistical Society, Series B* **54**, 793–804.

Pericchi, L. and Walley, P. (1991). Robust Bayesian credible intervals and prior ignorance. *International Statistical Review* **58**, 1–23.

Perold, A.F. (1988). Large-Scale portfolio optimization. *Management Science* **30**, 1143–1160.

Perou, C.M., Sorlie, T., Eisen, M.B., van deRijn, M., Jeffrey, S.S., Rees, C.A., Pollack, J.R., Ross, D.T., Johnsen, H., Akslen, L.A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S.X., Lonning, P.E., Borresen-Dale, A., Brown, P.O., and Botstein, D. (2000). Molecular portraits of human breast tumours. *Nature* **406**, 747–752.

Petrone, S. and Wasserman, L. (2002). Consistency of Bernstein polynomial posteriors. *Journal of the Royal Statistical Society, Series B* **64**, 79–100.

Pettit, L. (1986). Diagnostics in Bayesian model choice. *The Statistician: Journal of the Institute of Statisticians 35*, 183–190.

Pettit, L. and Smith, A.F.M. (1985). Outliers and inflential observations in linear models. In *Bayesian Statistics* **2** (Eds. J.M. Bernardo, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith). Amsterdam: Elsevier, pp. 473–494.

Piegorsch, W., Weinberg, C.R., and Taylor, J.A. (1994). Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Statistics in Medicine* **13**, 153–162.

Pincus, M. (1968). Letter to the Editor—A closed form solution of certain programming problems. *Operations Research* **16**, 690–694.

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability Theory Related Fields* **102**, 145–158.

Pitman, J. (1996). Some developments of the Blackwell-MacQueen urn scheme. In *Statistics, Probability and Game Theory*, IMS Lecture Notes-Monograph Series

30 (Eds. T.S. Ferguson, L.S. Shapley, and J.B. MacQueen). Beachwood, OH: Institute of Mathematical Statistics, pp. 245–267.

Pitman, J. (2003). Poisson-Kingman partitions. In *Statistics and Science: A Festschrift for Terry Speed*, IMS Lecture Notes-Monograph Series **40** (Ed. D.R. Goldstein). Beachwood, OH: Institute of Mathematical Statistics, pp. 1–34.

Pitman, J. (2006). *Combinatorial Stochastic Processes*. Berlin: Springer.

Pitman, J. and Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability* **25**, 855–900.

Pitt, M. and Shephard, N. (1999). Filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association* **94**, 590–599.

Plummer, M. (2008). Penalized loss functions for Bayesian model comparison. *Biostatistics* **9**, 523–539.

Pocock, S.J., Assmann, S.F., Enos, L.E., and Kasten, L.E. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: Current practice and problems. *Statististics in Medicine* **21**, 2917–2930.

Polson, N.G. (1996). Convergence of Markov chain Monte Carlo algorithms. In *Bayesian Statistics* **5** (Eds. J.M. Bernado, J.O. Berger, A.P. Dawid, and A.F.M. Smith). Oxford: Oxford University Press, pp. 297–323.

Polson, N.G. and Scott, J.G. (2009). Alternative global–local shrinkage rules using hypergeometric–beta mixtures. *Technical Report*, Department of Statistical Science, Duke University.

Polson, N.G. and Sorensen, M. (2009). Q-Learning via simulation. *Working Paper*, Booth School of Business, University of Chicago.

Polson, N.G. and Stroud, J.R. (2003). Bayesian inference for derivative prices. In *Bayesian Statistics* **7** (Eds. J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith, and M. West). Oxford: Oxford University Press, pp. 641–650.

Polson, N.G. and Tew, B. (2000). Bayesian portfolio selection: An empirical analysis of the S&P500 from 1970-1996. *Journal of Business and Economic Statistics* **18**, 164–173.

Poon, W. and Poon, Y. (1999). Conformal normal cuvature and assessment of local inflence. *Journal of the Royal Statistical Society, Series B* **61**, 51–61.

Postman, M., Huchra, J., and Geller, M. (1986). Probes of large-scale structure in the corona borealis region. *The Astronomical Journal* **92**, 1238–47.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika* **86**, 677–690.

Prasad, N.G.N. and Rao, J.N.K. (1990). The estimation of mean squared errors of small area estimators. *Journal of the American Statistical Association* **85**, 163–171.

Prentice, R.L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* **73**, 1–11.

Prentice, R.L. and Breslow, N.E. (1978). Retrospective studies and failure time model. *Biometrika* **65**, 153–158.

Prentice, R.L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66**, 403–411.

Prescott, G.J. and Garthwaite, P.H. (2005). Bayesian analysis of misclassified binary data from a matched case-control study with a validation sub-study. *Statistics in Medicine* **24**, 379–401.

Pritchard, J., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959.

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *Annals of Statistics* **22**, 300–325.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org.

Rabiner, L.R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of IEEE* **77**, 257–286.

Raftery, A.E. (1995). Bayesian model selection for social research (with Discussion). *Sociological Methodology* **25**, 111–196.

Raftery, A.E. (1996a). Hypothesis testing and model selection. In *Markov Chain Monte Carlo in Practice* (Eds. W.R. Gilks, D.J. Spiegelhalter, and S. Richardson). London: Chapman and Hall, pp. 163–188.

Raftery, A.E. (1996b). Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika* **83**, 251–266.

Raftery, A.E. and Richardson, S. (1996). Model selection for generalized linear models via GLIB: Application to nutrition and breast cancer. In *Bayesian Biostatistics* (Eds. D.A. Berry and D.K. Stangl). New York: Marcel Dekker, pp. 321–353.

Rannala, B. (2002). Identifiability of parameters in MCMC Bayesian inference of phylogeny. *Systematic Biology* **51**, 754–760.

Rannala, B. and Yang, Z. (2003). Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* **164**, 1645–1656.

Rannala, B., and Yang, Z. (2008). Phylogenetic inference using whole genomes. *Annual Review of Genomics and Human Genetics* **9**, 17–31.

Rao, J.N.K. (2003). *Small Area Estimation*. New York: Wiley.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* **4** (Ed. J. Neyman). Berkeley, CA: University of California press, pp. 321–334.

Rathouz, P.J., Satten, G.A., and Carroll, R.J. (2002). Semiparametric inference in matched case-control studies with missing covariate data. *Biometrika* **89**, 905–916.

Ray, S., Berger, J.O., Bayarri, M., and Pericchi, L. (2007). An extended BIC for model selection. *Presentation at the 2007 Joint Statistical Meetings, Salt Lake City*.

Redelings, B.D. and Suchard, M.A. (2005). Joint Bayesian estimation of alignment and phylogeny. *Systematic Biology* **54**, 401–418.

Redelings, B.D. and Suchard, M.A. (2007). Incorporating indel information into phylogeny estimation for rapidly emerging pathogens. *BMC Evolutionary Biology* **7**, 40.

Redelings, B.D. and Suchard, M.A. (2009). Robust inferences from ambiguous alignments. In *Sequence Alignment: Methods, Models, Concepts, and Strategies* (Ed. M.S. Rosenberg). Berkeley, CA: University of California Press, pp. 209–271.

Reed, D.K., Korytko, A.I., Hipkin, R.W., Wehrenberg, W.B., Schonbrunn, A., and Cuttler, L. (1999). Pituitary somatostatin receptor (sst)1-5 expression during rat development: Age-dependent expression of sst2. *Endocrinology* **140**, 4739–4744.

Reeves, R. and Pettitt, A.N. (2005). A theoretical framework for approximate Bayesian computation. In 20^{th} *International Workshop on Statistical Modelling*, Sydney, Australia, July 10-15, 2005, pp. 393–396.

Reich, B.J. and Fuentes, M. (2007). A multivariate nonparametric Bayesian spatial framework for hurricane surface wind fields. *Annals of Applied Statistics* **1**, 249–264.

Reich, B.J., Hodges, J.S., and Zadnik, V. (2006). Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics* **62**, 1197–1206.

Ren, C., Sun, D., and He, Z. (2009). Objective Bayesian analysis for a spatial model with nugget effects. *Technical Report*, Department of Statistics, Univeristy of Missouri-Columbia.

Ren, L., Dunson, D.B., and Carin, L. (2008). The dynamic hierarchical Dirichlet process. In *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland. New York: Association for Computing Machinery, pp. 824–831.

Rice, K.M. (2003). Full-likelihood approaches to misclassification of a binary exposure in matched case-control studies. *Statistics in Medicine* **22**, 3177–3194.

Rice, K.M. (2004). Equivalence between conditional and mixture approaches to the rasch model and matched case-control studies, with application. *Journal of the American Statistical Association* **99**, 510–522.

Rice, K.M. (2006). On Bayesian analysis of misclassified binary data from a matched case-control study with a validation sub-study, by Gordon Prescott and Paul Garthwaite. *Statistics in Medicine* **25**, 537–539.

Rice, K.M. (2008). Equivalence between conditional and random-effects likelihoods for pair-matched case-control studies, *Journal of the American Statistical Association* **103**, 385–396.

Rich, J.N., Hans, C., Jones, B., Iversen, E.S., McLendon, R.E., Rasheed, B.K., Dobra, A., Dressman, H.K., Bigner, D.D., Nevins, J.R., and West, M. (2005). Gene expression profiling and genetic markers in glioblastoma survival. *Cancer Research* **65**, 4051–4058.

Richardson, S. and Green, P.J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with Discussion). *Journal of the Royal Statistical Society, Series B* **59**, 731–792.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *Annals of Mathematical Statistics* **22**, 400–407.

Robert, C.P. (2001). *The Bayesian Choice*. Second Edition. New York: Springer.

Robert, C.P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Second Edition. New York: Springer.

Robert, C.P., Chopin, N., and Rousseau, J. (2009). Theory of Probability revisited (With Discussion). *Statistical Science*. In press.

Robert, C.P., Rydén, T., and Titterington, D.M. (2000). Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method. *Journal of the Royal Statistical Society, Series B* **62**, 57–75.

Robert, C.P. and Wraith, D. (2009). Computational methods for Bayesian model choice. In *MaxEnt 2009 Proceedings*. Albany, NY: MaxEnt Workshops Inc.

Roberts, G.O. and Rosenthal, J.S. (2007). Coupling and ergodicity of adaptive MCMC. *Journal of Applied Probability* **44**, 458–475.

Roberts, G.O. and Rosenthal, J.S. (2009). Examples of adaptive MCMC. *Technical Report*, Department of Statistics, University of Warwick.

Robins, J., van der Vaart, A.W., and Ventura, V. (2000). Asymptotic distribution of p values in composite null models. *Journal of the American Statistical Association* **95**, 1143–1159.

Rodríguez, A., Dunson, D.B., and Gelfand, A.E. (2008). The nested Dirichlet process. *Journal of the American Statistical Association* **103**, 1131–1154.

Roeder, K. (1990). Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *Journal of the American Statistical Association* **85**, 617–624.

Roeder, K., Carroll, R.J., and Lindsay, B.G. (1996). A semiparametric mixture approach to case-control studies with errors in covariables. *Journal of the American Statistical Association* **91**, 722–732.

Roeder, K. and Wasserman, L. (1997). Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association* **92**, 894–902.

Rokas, A., Williams, B.L., King, N., and Carroll, S.B. (2003). Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* **425**, 798–804.

Rosenberg, N.A., Pritchard, J.K., Weber, J.L., Cann, H.M., Kidd, K.K., Zhivotovsky, L.A., and Feldman, M.W. (2002). Genetic structure of human populations. *Science* **298**, 2381–2385.

Rossi, P., Allenby, G., and McCulloch, R.E. (2005). *Bayesian Statistics and Marketing*. New York: Wiley.

Rothwell, P.M. (2005). Subgroup analysis in randomised controlled trials: Importance, indications, and interpretation. *Lancet* **365**, 176–186.

Rougier, J. (2008a). Comment on article by Sansó et al. *Bayesian Analysis* **3**, 45–56.

Rougier, J. (2008b). Efficient emulators for multivariate deterministic functions. *Journal of Computational and Graphical Statistics* **17**, 827–843.

Rowe, D. (2002). *Multivariate Bayesian Statistics: Models for Source Separation and Signal Unmixing*. Boca Raton, FL: Chapman & Hall / CRC Press.

Roy, A., Bhaumik, D.K., Aryal, S., and Gibbons, R.D. (2007). Sample size determination for hierarchical longitudinal designs with differential attrition rates. *Biometrics* **63**, 699–707.

Rubin, D.B. (1978). Multiple imputation in sample surveys: A phenomenological Bayesian approach to nonresponse (with Discussion). *Proceedings of the American Statistical Association, Survey Research Methods Section*, 20–34.

Rubin, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association* **91**, 473–489.

Rubin, H. and Sethuraman, H. (1965). Bayes risk efficiency. *Sankhyā, Series A*, 347–356.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Boca Raton, FL: Chapman & Hall / CRC Press.

Ruggeri, F., Ríos Insua, D., and Martín, J. (2005). Robust Bayesian analysis. In *Bayesian Thinking: Modeling and Computation. Handbook of Statistics* **25** (Eds. D.K. Dey and C.R. Rao). Amsterdam: North-Holland, pp.623–667.

Ruggeri, F. and Sivaganesan, S. (2000). On a global sensitivity measure for Bayesian inference. *Sankhyā Series A* **62**, 110–127.

Ruggeri, F. and Wasserman, L. (1993). Infiitesimal s ensitivity of posterior distributions. *Canadian Journal of Statistics* **21**, 195–203.

Ruggeri, F. and Wasserman, L. (1995). Density based classes of priors: infiitesimal properties and approximations. *Journal of Statistical Planning and Inference* **46**, 311–324.

Sacks, J., Welch, W.J., Mitchell, T.J., and Wynn, H.P. (1989). Design and analysis of computer experiments (C/R: P423-435). *Statistical Science* **4**, 409–423.

Saito, T., Iwata, N., Tsubuki, S., Takaki, Y., Takano, J., Huang, S.M., Suemoto, T., Higuchi, M., and Saido, T.C. (2005). Somatostatin regulates brain amyloid beta-peptide A-beta-42 through modulation of proteolytic degradation. *Nature Medicine* **11**, 434–439.

Sala-i Martin, X., Doppelhofer, G., and Miller, R.I. (2004). Determinants of long-term growth: A Bayesian averaging of classical estimates (BACE) approach. *American Economic Review* **94**, 813–835.

Samuelson, P.A. (1969). Lifetime portfolio selection by dynamic stochastic programming. *Review of Economics and Statistics* **51**, 239–246.

Sansó, B., Forest, C., and Zantedeschi, D. (2008). Inferring climate system properties using a computer model. *Bayesian Analysis* **3**, 1–38.

Satten, G.A. and Carroll, R.J. (2000). Conditional and unconditional categorical regression models with missing covariates. *Biometrics* **56**, 384–388.

Satten, G.A. and Kupper, L. (1993). Inferences about exposure-disease association using Probabiliy-of-Exposure Information. *Journal of the American Statistical Association* **88**, 200–208.

Satten, G.A., Flanders, W.D., and Yang, Q. (2001). Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. *Ameriacan Journal of Human Genetics* **68**, 466–477.

Schabenberger, O. and Gotway, C.A. (2004). *Statistical Methods for Spatial Data Analysis*. Boca Raton, FL: Chapman & Hall / CRC Press.

Schafer, J.L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research* **8**, 3–15.

Schmidt, A.M. and Gelfand, A.E. (2003). A Bayesian coregionalization approach for multivariate pollutant data. *Journal of Geophysical Research - Atmosphere* **108**, D24, 8783.

Schmidt, S. and Schaid, D.J. (1999). Potential misinterpretation of the case-only study to assess gene-environment interaction. *American Journal of Epidemiology* **150**, 878–885.

Schmittner, A., Urban, N.M., Keller, K., and Matthews, D. (2009). Using tracer observations to reduce the uncertainty of ocean diapycnal mixing and climate carbon-cycle projections. *Global Biogeochemical Cycles* **23**, GB4009.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464.

Schwartz, L. (1965). On Bayes procedures. *Zeitschrift fir Wahrscheinlichkeitstheorie und Verwandte Gebiete* **4**, 10–26.

Scott, A.J. and Wild, C.J. (1986). Fitting logistic models under case-control or choice-based sampling. *Journal of the Royal Statistical Society, Series B* **48**, 170–182.

Scott, J.G. (2009). Flexible learning on the sphere via adaptive needlet shrinkage and selection. *Technical Report*, Department of Statistical Science, Duke University.

Scott, J.G. and Berger, J.O. (2006). An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference* **136**, 2144–2162.

Scott, J.G. and Berger, J.O. (2008). Bayes and Empirical-Bayes multiplicity adjustment in the variable-selection problem. *Technical Report*, Department of Statistical Science, Duke University.

Scott, J.G. and Carvalho, C.M. (2008). Feature-inclusion stochastic search for Gaussian graphical models. *Journal of Computational and Graphical Statistics*, **17**, 790–808.

Seaman, S.R. and Richardson, S. (2001). Bayesian analysis of case-control studies with categorical covariates. *Biometrika* **88**, 1073–1088.

Seaman, S.R. and Richardson, S. (2004). Equivalence of prospective and retrospective models in the Bayesian analysis of case-control studies. *Biometrika* **91**, 15–25.

Seo, D.M., Goldschmidt-Clermont, P.J., and West, M. (2007). Of mice and men: Sparse statistical modelling in cardiovascular genomics. *Annals of Applied Statistics* **1**, 152–178.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica* **4**, 639–650.

Severini, T.A., Mukerjee, R., and Ghosh, M. (2002). On an exact probability matching property of right-invariant priors. *Biometrika* **89**, 952–957.

Shannon, C. (1948a). A mathematical theory of communication, Part I. *Bell Labs Technical Journal 27*, 379–423.

Shannon, C. (1948b). A mathematical theory of communication, Part II. *Bell Labs Technical Journal* **27**, 623–656.

Shao, J. (1996). Bootstrap model selection. *Journal of American Statistical Association* **91**, 655–665.

Shen, X. and Wasserman, L. (2001). Rates of convergence of posterior distributions. *Annals of Statistics* **29**, 666–686.

Shepard, R.N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science* **210**, 390–398.

Shi, J. and Wei, B. (1995). Bayesian local inflence. *Chinese Mathematica Applicata* **8**, 237–245.

Shih, J.H. and Chatterjee, N. (2002). Analysis of survival data from case-control family studies. *Biometrics* **58**, 502–509.

Shin, J.S., Fong, D.K.H., and Kim, K.J. (1998). Complexity reduction of a house of quality chart using correspondence analysis. *Quality Management Journal* **5**, 46–58.

Shinozaki, N. (1984). Simultaneous estimation of location parameters under quadratic loss. *Annals of Statistics* **12**, 322–335.

Sinha, D. and Dey, D.K. (1997). Semiparametric Bayesian analysis of survival data. *Journal of the American Statistical Association* **92**, 1195–1212.

Sinha, S., Mukherjee, B., and Ghosh, M. (2004). Bayesian semiparametric modeling for matched case-control studies with multiple disease states. *Biometrics* **60**, 41–49.

Sinha, S., Mukherjee, B., Ghosh, M., Mallick, B.K., and Carroll, R. (2005). Semiparametric Bayesian modeling of matched case-control studies with with missing exposure. *Journal of the American Statistical Association* **100**, 591–601.

Sinha, S., Mukherjee, B., and Ghosh, M. (2007). Modeling association among multivariate exposures in matched case-control study. *Sankhyā* **64**, 379–404.

Sinharay, S., Johnson, M.S., and Stern, H.S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement* **30**, 298–321.

Sinsheimer, J.S., Lake, J.A., and Little, R.J.A. (1996). Bayesian hypothesis testing of four-taxon topologies using molecular sequence data. *Biometrics* **52**, 193–210.

Sivaganesan, S. (1988). Range of posterior measures for priors wiht arbitrary contaminations. *Communications in Statistics* **17**, 1591–1612.

Sivaganesan, S. (1989). Sensitivity of posterior mean to unimodality preserving contaminations. *Statistics & Decisions* **7**, 77–93.

Sivaganesan, S. (1991). Sensitivity of some standard Bayesian estimates to prior uncertainty: A comparison. *Journal of Statistical Planning and Inference* **27**, 85–103.

Sivaganesan, S. (1993). Range of the posterior probability of an interval for priors with unimodality preserving contaminations. *Annals of the Institute of Mathematical Statistics* **45**, 171–188.

Sivaganesan, S. (2000). Global and local robustness approaches: Uses and limitations. In *Robust Bayesian Analysis*. *Lecture Notes in Statistics* **152** (Eds. D. Ríos Insua and F. Ruggeri). New York: Springer, pp. 89-108.

Sivaganesan, S. and Berger, J.O. (1989). Ranges of posterior measures for priors with unimodal contaminations. *Annals of Statistics* **17**, 868–889.

Sivaganesan, S. and Berger, J.O. (1993). Robust Bayesian analysis of the binomial empirical Bayes problem. *Canadian Journal of Statistics* **21**, 107–119.

Sivaganesan, S., Laud, P.W., and Müller, P. (2009). A Bayesian subgroup analysis with a zero-enriched Polya urn scheme. *Technical Report*, Department of Mathematical Sciences, University of Cincinnati.

Skilling, J. (2006). Nested sampling for general Bayesian computation. *Bayesian Analysis* **1**, 833–860.

Smith, M. and Kohn, R. (1996). Non-parametric regression using Bayesian variable selection. *Journal of Econometrics* **75**, 317–344.

Smith, R.L. (2003). Statistics of extremes, with applications in environment, insurance and finance. In *Extreme Values in Finance, Telecommunications and the Environment* (Eds. B. Finkenstadt and H. Rootzen). Boca Raton, FL: Chapman & Hall / CRC Press, pp. 1–78.

Sørlie, T., Perou, C.M., Tibshirani, R.J., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Thorsen, T., Quist, H., Matese, J.C., Brown, P.O., Botstein, D., Lønning, P., and Børresen-Dale, A. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 10869–10874.

Spezia, L. (2009). Reversible jump and the label switching problem in hidden Markov models. *Journal of Statistical Planning and Inference* **139**, 2305–2315.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with Discussion). *Journal of the Royal Statistical Society, Series B: Statistical Methodology* **64**, 583–639.

Spiegelhalter, D.J. and Freedman, L.S. (1986). A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. *Statistics in Medicine* **5**, 1–13.

Spiegelhalter, D.J., Thomas, A., Best, N.G., Gilks, W.R., and Lunn, D. (1994, 2002). *BUGS: Bayesian Inference Using Gibbs Sampling*. MRC Biostatistics Unit, Cambridge, England.

Srinivasan, C. and Truszaczynska, A. (1990). Approximations to the range of a ratio linear posterior quantity. *Technical Report*, Department of Statistics, University of Kentucky.

Stallard, N. (2003). Decision-theoretic designs for phase ii clinical trials allowing for competing studies. *Biometrics* **59**, 402–409.

Stallard, N., Thall, P.F., and Whitehead, J. (1999). Decision theoretic designs for phase II clinical trials with multiple outcomes. *Biometrics* **55**, 971–977.

Stambaugh, R.F. (1999). Predictive regressions. *Journal of Financial Economics* **54**, 375–421.

Stamey, T., Kabalin, J., McNeal, J., Johnstone, I., Freiha, F., Redwine, E., and Yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate II: Radical prostatectomy treated patients. *Journal of Urology* **16**, 1076–1083.

Steele, R.J. (2002). *Importance sampling methods for inference in mixture models and missing data*. *Unpublished Ph.D. Dissertation*, Department of Statistics, University of Washington.

Steele, R.J. and Raftery, A.E. (2009). Performance of Bayesian model selection criteria for Gaussian mixture models. *Technical Report 559*, Department of Statistics, University of Washington.

Steele, R.J., Raftery, A.E., and Emond, M.J. (2006). Computing normalizing constants for finite mixture models via incremental mixture importance sampling (IMIS). *Journal of Computational and Graphical Statistics* **15**, 712–734.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* **1**. Berkerly, CA: University of California Press, pp. 197–206.

Stein, C. (1974). Estimation of the mean of a multivariate normal distribution. In *Proceedings of the Prague Symposium on Asymptotic Statistics* (Ed. J. Hájek). Prague: Universita Karlova, pp. 345–381.

Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *Annals of Statistics* **9**, 1135–1151.

Stein, M.L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. New York: Springer.

Steinbakk, G. and Storvik, G. (2009). Posterior predictive p-values in Bayesian hierarchical models. *Scandinavian Journal of Statistics*, **36**, 320–336.

Stephens, M. (2000a). Bayesian analysis of mixture models with an unknown number of components – An alternative to reversible jump methods. *The Annals of Statistics* **28**, 40–74.

Stephens, M. (2000b). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society, Series B* **62**, 795–809.

Stewart, F.J., Young, C.R., and Cavanaugh, C.M. (2009). Evidence for homologous recombination in intracellular chemosynthetic clam symbionts. *Molecular Biology and Evolution* **26**, 1391–1404.

Storvik, G. (2002). Particle filters in state space models with the presence of unknown static parameters. *IEEE Transactions of Signal Processing* **50**, 281–289.

Strawderman, W.E. (1971). Proper Bayes minimax estimators of multivariate normal mean. *Annals of Mathematical Statistics* **42**, 385–388.

Strawderman, W. E. (1972). On the existence of proper Bayes minimax estimators of the mean of a multivariate normal distribution. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*. Berkerly, CA: University of California Press, pp. 51–55.

Strawderman, W.E. (1973). Proper Bayes minimax estimators of multivariate normal mean for the case of common unknown variances. *Annals of Statistics* **1**, 1189–1194.

Strawderman, W.E. (1974). Minimax estimation of location parameters for certain spherically symmetric distributions. *Journal of Multivariate Analysis* **4**, 255–264.

Strawderman, W.E. (1992). The James-Stein estimator as an empirical Bayes estimator for an arbitrary location family. In *Bayesian Statistics* **4** (Eds. J.M. Bernado, J.O. Berger, A.P. Dawid, and A.F.M. Smith). Oxford: Oxford University Press, pp. 821–824.

Strawderman, W.E. (2003). On minimax estimation of a normal mean vector for general quadratic loss. In *Mathematical Statistics & Applications: A Festschrift for Constance van Eeden*, IMS Monograph Series **42** (Eds. M. Moore, S. Froda, and C. Leger). Beachwood, OH: Institute of Mathematical Statistics, pp. 3–14.

Strawderman, W.E. and Cohen, A. (1971). Admissibility of estimators of the mean vector of a multivariate normal distribution with quadratic loss. *Ann. Math. Statist.* **42**, 270-296.

Stukel, T. (1988). Generalized logistic models. *Journal of the American Statistical Association* **83**, 426–431.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L. Golub, T.R., Lauder, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 15545–15550.

Suchard, M.A. (2005). Stochastic models for horizontal gene transfer: Taking a random walk through tree space. *Genetics*, **170**, 419–431.

Suchard, M.A. and Rambaut, A. (2009). Many-core algorithms for statistical phylogenetics. *Bioinformatics* **25**, 1370–1376.

Suchard, M.A., Weiss, R.E., and Sinsheimer, J.S. (2001). Bayesian selection of continuous-time Markov chain evolutionary models. *Molecular Biology and Evolution* **18**, 1001–1013.

Suchard, M.A., Kitchen, C.M.R., Sinsheimer, J.S., and Weiss, R.E. (2003a). Hierarchical phylogenetic models for analyzing multipartite sequence data. *Systematic Biology* **52**, 649–664.

Suchard, M.A., Weiss, R.E., Dorman, K.S., and Sinsheimer, J.S. (2003b). Inferring spatial phylogenetic variation along nucleotide sequences: A multiple change-point model. *Journal of American Statistical Association* **98**, 427–437.

Suchard, M.A., Weiss, R.E., and Sinsheimer, J.S. (2005). Models for estimating Bayes factors with applications to phylogeny and tests of monophyly. *Biometrics* **61**, 665–673.

Sudderth, E.B. (2006). Graphical models for visual object recognition and tracking. *Unpublished Ph.D. Dissertation*, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.

Sun, D. and Berger, J.O. (1998). Reference priors with partial information. *Biometrika* **85**, 55–71.

Sun, D. and Berger, J.O. (2007). Objective Bayesian analysis for the multivariate normal model (with Discussion). In *Bayesian Statistics* **8** (Eds. J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith, and M. West). Oxford: Oxford University Press, pp. 525–547.

Suzuki, Y., Glazko, G.V., and Nei, M. (2002). Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 16138–16143.

Swaminathan, H. and Gifford, J.A. (1982). Bayesian estimation in the Rasch model. *Journal of Educational Statistics* **7**, 175–192.

Swaminathan, H. and Gifford, J.A. (1985). Bayesian estimation in the two-parameter logistic model. *Psychometrika* **50**, 349–364.

Sweeting, T.J., Datta, G.S., and Ghosh, M. (2006). Nonsubjective priors via predictive relative entropy regret. *Annals of Statistics* **34**, 441–468.

Swendsen, R.H. and Wang, J. (1987). Nonuniversal critical dynamics in Monte Carlo simulations. *Physical Review Letter* **58**, 86–88.

Tai, F. and Pan, W. (2009). Bayesian variable selection in regression with networked predictors. *Technical Report*, Division of Biostatistics, University of Minnesota.

Takane, Y., Young, F.W., and de Leeuw, J. (1977). Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features. *Psychometrika* **42**, 7–67.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewen, S., Dmitrovsky, E., Lander, E.S., and Golub, T.R., (1999). Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 2907–2912.

Tanaka, F. (2006). Generalized Bayesian predictive density operators. *The 14th Quantum Information Technology Symposium*, 107-110.

Tanaka, F. and Komaki, F. (2005). Bayesian predictive density operators for exchangeable quantum-statistical models. *Physical Review A, American Institute of Physics* **71**, 052323.

Tanner, M.A. and Wong, W.H. (1987). The calculation of posterior distributions by data augmentation (with Discussion). *Journal of the American Statistical Association* **82**, 528–549.

Tantrum, J., Murua, A., and Stuetzle, W. (2003). Assessment and pruning of hierarchical model based clustering. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: Association for Computing Machinery, pp. 197–205.

Tavazoie, S., Hughes, J.D., Cambell, M.J., Cho, R.J., and Church, G.M. (1999). Systematic determination of genetic network architecture. *Nature Genetics* **22**, 281–285.

Taylor, J.W. and Buizza, R. (2004). A comparison of temperature density forecasts from GARCH and atmospheric models. *Journal of Forecasting* **23**, 337–355.

Teh, Y.W., Jordan, M.I., Beal, M.J., and Blei, D.M. (2006). Hierarchical Dirichlet processes. *Journal of American Statistical Association* **101**, 1566–1581.

Thibaux, R. and Jordan, M.I. (2007). Hierarchical beta processes and the Indian buffet process. In *Proceedings of the International Workshop on Artificial Intelligence and Statistics*, volume 11.

Thorne, J.L., Kishino, H., and Felsenstein, J. (1992). Inching towards reality: An improved likelihood model of sequence evolution. *Journal of Molecular Evolution* **34**, 3–16.

Thorne, J.L., Kishino, H., and Painter, I.S. (1998). Estimating the rate of evolution of the rate of molecular evolution. *Molecular Biology and Evolution* **15**, 1647–1657.

Tibshirani, R.J. (1992). Principal curves revisited. *Statistics and Computing* **2**, 183–190.

Tibshirani, R.J. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58**, 267–288.

Tibshirani, R.J., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society, Series B* **67**, 91–108.

Tierney, L. and Kadane, J.B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association* **81**, 82–86.

Tipping, M.E. (2001). Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research* **1**, 211–244.

Tipping, M.E. and Bishop, C.M. (1999a). Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B* **61**, 611–622.

Tipping, M.E. and Bishop, C.M. (1999b). Mixtures of principal component analyzers. *Neural Computation* **11**, 443–482.

Tipping, M.E. and Faul, A.C. (2003). Fast marginal likelihood maximisation for sparse Bayesian models. In *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics, Key West, FL, Jan 3-6* (Eds. C.M. Bishop and B.J. Frey).

Titterington, D.M., Smith, A.F.M., and Makov, U.E. (1985). *Statistical Analysis of Finite Mixture Distributions*. New York: Wiley.

Tokdar, S.T. (2006). Posterior consistency of Dirichlet location-scale mixture of normals in density estimation and regression. *Sankhyā* **68**, 90–110.

Tokdar, S.T. (2007). Towards a faster implementation of density estimation with logistic Gaussian process priors. *Journal of Computational and Graphical Statistics* **16**, 633–655.

Tokdar, S.T. (2009). Testing normality with Dirichlet mixture alternatives. *Technical Report*, Department of Statistical Science, Duke University.

Tokdar, S.T. and Ghosh, J.K. (2007). Posterior consistency of logistic Gaussian process priors in density estimation. *Journal of Statistical Planning and Inference* **137**, 34–42.

Torrie, G. and Valleau, J. (1977). Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics* **23**, 187–199.

Troxel, A., Ma, G., and Heitjan, D. (2004). An index of local sensitivity to nonignorability. *Statistica Sinica* **14**, 1221–1237.

Tsai, C. and Wu, X. (1992). Transformation-model diagnostics. *Technometrics* **34**,197–202.

Tsui, K.W. (1979). Multiparameter estimation of discrete exponential distributions. *Canadian Journal of Statistics* **7**, 193-200.

Tsutakawa, R.K. and Lin, H.Y. (1986). Bayesian estimation of item response curves. *Psychometrika* **51**, 251–267.

Tukey, J.W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

van den Hout, A. and van der Heijden, P.V. (2002). Randomized response, statistical disclosure control and misclassification: A review. *International Statistical Review* **70**, 269–288.

van der Linde, A. (2000). Reference priors for shrinkage and smoothing parameters. *Journal of Statistical Planning and Inference* **90**, 245–274.

van der Vaart, A.W. and van Zanten, H. (2008). Rates of contraction of posterior distributions based on Gaussian process priors. *Annals of Statistics* **36**, 1435–1463.

Van't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A.M., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T., Schreiber, G.J., Kerkhoven, R.M., Roberts, C., Linsley, P.S., Bernards, R., and Friend, S.H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536.

Verdinelli, I. and Wasserman, L. (1998). Bayesian goodness of fit testing using infinite dimensional exponential families. *Annals of Statistics* **20**, 1203–1221.

Ver Hoef, J.M. and Barry, R.P. (1998). Constructing and fitting models for cokriging and multivariable spatial prediction. *Journal of Statistical Planning and Inference* **69**, 275–294.

Vidakovic, B. (1998). Nonlinear wavelet shrinkage with Bayes rules and Bayes factors. *Journal of the American Statistical Association* **93**, 173–179.

Wacholder, S. (1996). The case-control study as data missing by design: Estimating risk difference. *Epidemiology* **7**, 144–150.

Wackernagel, H. (2003). *Multivariate Geostatistics: An Introduction with Applications*, Third Edition. New York: Springer.

Walker, S.G. (2003). On sufficient conditions for Bayesian consistency. *Biometrika* **90**, 482–490.

Walker, S.G. (2004). New approaches to Bayesian consistency. *Annals of Statistics* **32**, 2028–2043.

Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities*. London: Chapman and Hall.

Wang, F. and Gelfand, A.E. (2002). A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models. *Statistical Science* **17**, 193–208.

Wang, N. and Raftery, A.E. (2002). Nearest neighbor variance estimation (NNVE): Robust covariance estimation via nearest neighbor cleaning (with Discussion). *Journal of the American Statistical Association* **97**, 994–1019.

Wang, Q., Carvalho, C.M., Lucas, J.E., and West, M. (2007). BFRM: Bayesian factor regression modelling. *Bulletin of the International Society for Bayesian Analysis* **14**, 4–5.

Wang, Q., Carvalho, C.M., Lucas, J.E., and West, M. (2007-present). BFRM: Software for sparse Bayesian factor regression modelling. http://www.stat.duke.edu/research/software/west/bfrm/index.html.

Wang, X. and Dey, D.K. (2009). Generalized extreme value regression for binary response data: An application to B2B electronic payments system adoption. *Technical Report*, Department of Statistics, University of Connecticut.

Wang, X., Mohanty, N., and McCallum, A.K. (2005). Group and topic discovery from relations and text. In *Proceedings of the 3rd International Workshop on Link Discovery*. New York: Association for Computing Machinery, pp. 28–35.

Warner, S. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association* **60**, 63–70.

Wasserman, L. (1992). Recent methodological advances in robust Bayesian inference. In *Bayesian Statistics* **4** (Eds. J.M. Bernado, J.O. Berger, A.P. Dawid, and A.F.M. Smith). Oxford: Oxford University Press, pp. 483–502.

Wasserman, L. (2000). Asymptotic inference for mixture models using data-dependent priors. *Journal of the Royal Statistical Society, Series B* **62**, 159–180.

Wathen, J.K. and Christen, J.A. (2006). Implementation of backward induction for sequentially adaptive clinical trials. *Journal of Computational and Graphical Statistics* **15**, 398–413.

Wathen, J.K. and Thall, P.F. (2008). Bayesian adaptive model selection for optimizing group sequential clinical trials. *Statistics in Medicine* **27**, 5586–5604.

Weaver, A., Eby, M., Wiebe, E., Bitz, C., Duffy, P., Ewen, T., Fanning, A., Holland, M., MacFadyen, A., Matthews, H., et al. (2001). The UVic Earth System Climate Model: Model description, climatology, and applications to past, present and future climates. *Atmosphere-Ocean* **39**, 361–428.

Wei, Z. and Li, H. (2007). A Markov random field model for network-based analysis of genomic data. *Bioinformatics* **23**, 1537–1544.

Wei, Z. and Li, H. (2008). A hidden spatial-temporal Markov random field model for network-based analysis of time course gene expression data. *Annals of Applied Statistics* **2**, 408–429.

Weinberg, J., Brown, L.D., and Stroud, J.R. (2007). Bayesian forecasting of an inhomogeneous Poisson process with applications to call center data. *Journal of the American Statistical Association* **102**, 1185–1198.

Weiss, R.E. (1994). Pediatric pain, predictive inference and sensitivity analysis. *Evaluation Review* **18**, 651–678.

Weiss, R. (1996). An approach to Bayesian sensitivity analysis. *Journal of the Royal Statistical Society, Series B* **58**, 739–750.

Weiss, R.E. (1997). Bayesian sample size calculations for hypothesis testing. *The Statistician: Journal of the Institute of Statisticians* **46**, 185–191.

Weiss, R.E. (2005). *Modeling Longitudinal Data*. New York: Springer.

Weiss, R.E. and Cho, M. (1998). Bayesian marginal inflence assessment. *Journal of Statistical Planning and Inference* **71**,163–177.

Weiss, R.E. and Cook, R. (1992). A graphical case statistic for assessing posterior inflence. *Biometrika* **79**, 51–55.

Weiss, R.E., Wang, Y., and Ibrahim, J.G. (1997). Predictive model selection for repeated measures random effects models using Bayes factors. *Biometrics* **53**, 592–602.

Welch, B. and Peers, H.W. (1963). On formulae for confidence points based on integrals of weighted likelihoods. *Journal of the Royal Statistical Society, Series B* **25**, 318–329.

West, M. (1992). Modelling with mixtures (with Discussion). In *Bayesian Statistics* **7** (Eds. J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith). Oxford: Oxford University Press, pp. 503–524.

West, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. In *Bayesian Statistics* **7** (Eds. J.M. Bernardo, M. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith, and M. West). Oxford: Oxford University Press, pp. 723–732.

West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Marks, J.R., and Nevins, J.R. (2001). Predicting the clinical status of human breast cancer utilizing gene expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 11462–11467.

West, M. and Harrison, P. (1997). *Bayesian Forecasting and Dynamic Models*. Second Edition. New York: Springer.

Whittmore, A.S. (1983). Transformations to linearity in binary regression. *SIAM Journal on Applied Mathematics* **43**, 703–710.

Wipf, D.P. and Nagarajan, S. (2008). A new view of automatic relevance determination. In *Advances in Neural Information Processing Systems* **20** (Eds. J.C. Platt, D. Koller, Y. Singer, and S. Roweis). Cambridge, MA: MIT Press.

Wolff, U. (1989). Collective Monte Carlo updating for spin systems. *Physical Review Letters* **62**, 361.

Womble, W.H. (1951). Differential systematics. *Science* **114**, 315–322.

Wong, K.M., Suchard, M.A., and Huelsenbeck, J.P. (2008). Alignment uncertainty and genomic analysis. *Science* **319**, 473–475.

Wu, X. and Luo, Z. (1993). Second-order approach to local inflence. *Journal of the Royal Statistical Society, Series B* **55**, 929–936.

Wu, Y. and Ghosal, S. (2008a). Posterior consistency for some semiparametric problems. *Sankhyā, Series A* **70**, 0-46.

Wu, Y. and Ghosal, S. (2008b). Kullback Leibler property of kernel mixture priors in Bayesian density estimation. *Electronic Journal of Statistics* **2**, 298–331.

Xie, W., Lewis, P.O., Fan, Y., Kuo, L., and Chen, M.-H. (2009). Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Technical Report*, Department of Statistics, University of Connecticut.

Xing, E.P., Jordan, M.I., and Sharan, R. (2007). Bayesian haplotype inference via the Dirichlet process. *Journal of Computational Biology* **14**, 267–284.

Xu, X. and Liang, F. (2010). Asymptotic minimax risk of predictive density estimation for nonparametric regression. *Bernoulli*. In press.

Yacubova, E. and Komuro, H. (2002). Stage-specific control of neuronal migration by somatostatin. *Nature* **415**, 77–81.

Yeung, K.Y., Fraley, C., Murua, A., Raftery, A.E., and Ruzzo, W.L. (2001). Model based clustering and data transformation for gene expression data. *Bioinformatics* **17**, 977–987.

You, Y. and Chapman, B. (2006). Small area estimation using area level models and estimated sampling variances. *Survey Methodology* **32**, 97–103.

Yuan, A., Zheng, G., and Xu, J. (2009). On empirical likelihood statistical functions. *Technical Report*, Human Genome Project, Howard University.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B* **68**, 49–67.

Zelen, M. and Parker, R.A. (1986). Case-control studies and Bayesian inference. *Statistics in Medicine* **5**, 261–269.

Zellner, A. (1975). Bayesian analysis of regression error terms. *Journal of the American Statistical Association* **70**, 138–144.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distribution. In *Bayesian Inference and Decision Techniques. Essays in Honor of Bruno De Finetti* (Eds. P.K. Goel and A. Zellner). Amsterdam: Elsevier, pp. 233–243.

Zellner, A. (2002). Information processing and Bayesian analysis. *Journal of Econometrics* **107**, 41–50.

Zellner, A. and Moulton, B.R. (1985). Bayesian regression diagnostics with applications to international consumption and income data. *Journal of Econometrics* **29**, 187–211.

Zellner, A. and Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In *Bayesian Statistics: Proceedings of the First International Meeting held in Valencia* (Eds. J.M. Bernado, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith). Valencia, Spain: University Press, pp. 585–603.

Zhang, L., Mukherjee, B., Ghosh, M., and Wu, R.L. (2006). A Bayesian framework for genetic association in case-control studies accounting for unknown population substructure. *Statistical Modeling* **6**, 352–372.

Zhang, L., Mukherjee, B., Hu, B., Cooney, K., and Moreno, V. (2009). Semiparametric Bayesian modeling of random genetic effects in family-based association studies. *Statistics in Medicine*, **28**, 113–139.

Zhang, Z. (1994). Discrete noninformative priors. *Unpublished Ph.D. Dissertation*, Department of Statistics, Yale University.

Zhu, H., Ibrahim, J.G., and Cho, N. (2010). Bayesian case-deletion measures. *Technical Report*, Department of Biostatistics, University of North Carolina at Chapel Hill.

Zhu, H., Ibrahim, J.G., Lee, S., and Zhang, H. (2007). Perturbation selection and inflence measures in local inflence analysis. *Annals of Statistics* **35**, 2565–2588.

Zhu, H., Ibrahim, J.G., and Tang, N. (2009). Bayesian local inflence analysis: A geometric approach. *Technical Report*, Department of Biostatistics, University of North Carolina at Chapel Hill.

Zhu, H. and Lee, S. (2001). Local inflence for incomplete-data models. *Journal of the Royal Statistical Society, Series B* **63**, 111–126.

Zhu, H. and Lee, S. (2003). Local inflence for generalized linear mixed models. *Canadian Journal of Statistics* **31**, 293–309.

Zhu, Z., He, X., and Fung, W. (2003). Local inflence analysis for penalized gGussian likelihood estimators in partially linear models. *Scandinavian Journal of Statistics* **30**, 767–780.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, **101**, 1418-1-429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B* **67**, 301–320.

Author Index

Subject Index