Institute of Mathematical Statistics
LECTURE NOTES–MONOGRAPH SERIES

Science and Statistics: A Festschrift for Terry Speed

Darlene R. Goldstein, Editor

Volume 40

Institute of Mathematical Statistics

LECTURE NOTES–MONOGRAPH SERIES

Volume 40

Statistics and Science: A Festschrift for Terry Speed

Darlene R. Goldstein, Editor

Institute of Mathematical Statistics
Beachwood, Ohio

Institute of Mathematical Statistics
Lecture Notes-Monograph Series

Series Editor:
Joel Greenhouse

The production of the *Institute of Mathematical Statistics Lecture Notes-Monograph Series* is managed by the IMS Societal Office: Julia A. Norton, Treasurer and Elyse Gustafson, Executive Director.

Library of Congress Control Number: 2002117748

International Standard Book Number 0-940600-56-0

Printed in the United States of America

Preface

As a token of appreciation and in celebration of the 60th birthday of Professor Terence P. Speed, this volume has been compiled from contributions offered by his colleagues and former students.

Terry's intellectual curiosity and enthusiasm, along with deep and very broad subject matter knowledge and expertise, have been highly influential to many individuals. His influence extends to the wider statistical community as well. Terry has a solid record of service to the statistical profession, including serving as reviewer or editor for numerous journals and also for the U. S. National Science Foundation and National Institutes of Health, and is a member (or Fellow) of several professional societies, including the ASA, ISI, WNAR, and IMS. Terry has also accepted officer responsibilities (for example, past President of WNAR), and indeed is serving as the current President of the IMS. In addition, he has received a number of professional accolades; for example, he has recently given the prestigious Wald lectures at the Joint Statistical Meetings (2001) and the Forum Lecture at the European Meeting of Statisticians (2002).

After completing the Ph. D., Terry became a Lecturer at the University of Sheffield. He was later Associate Professor, then Professor, at the University of Western Australia. He was Chief of the Division of Mathematics and Statistics at the CSIRO in Australia before coming to the University of California at Berkeley as a Professor in 1987. Since 1997, he has split his time, roughly evenly, between Berkeley and Melbourne, Australia, where he is Division Head of Bioinformatics at the Walter and Eliza Hall Institute of Medical Research.

His career thus far has encompassed a number of remarkably distinct areas, beginning with algebra, then probability, followed by statistics and its applications in a number of fields, particularly in genetics and, most recently, bioinformatics. This broad background has allowed him to find connections between apparently vastly different fields: for example, between algebraic group theory and genetic linkage analysis. Terry also maintains an active role in statistical education at all levels, promoting learning both through his own teaching and his book *Stat Labs* and papers on the subject.

To present a concrete picture of Terry's research activities, we considered including a list of his publications here. However, we have not done so for two practical reasons. First, space: his publications number in the hundreds; and second, timeliness: Terry is even now publishing at so rapid a pace that any such list would be immediately out of date. We refer interested readers to electronic databases such as Current Index to Statistics, MathSciNet, PubMed of the United States National Library of Medicine, and Current Contents, for example. However, to give an impression of the variety of Terry's work, we note that his papers have been published in such diverse journals as:

> *Mathematics, Probability, and Information Theory*: Advances in Applied Probability, Annals of Applied Probability, Canadian Mathematical Bulletin, IEEE Transactions on Information Theory, Journal of Applied Probability, Journal of the Australian Mathematical Society, Journal of the Lon-

don Mathematical Society, Probability Theory and Related Fields, Proceedings of the Cambridge Philosophical Society, SIAM Journal on Applied Mathematics, Stochastic Processes and their Applications;

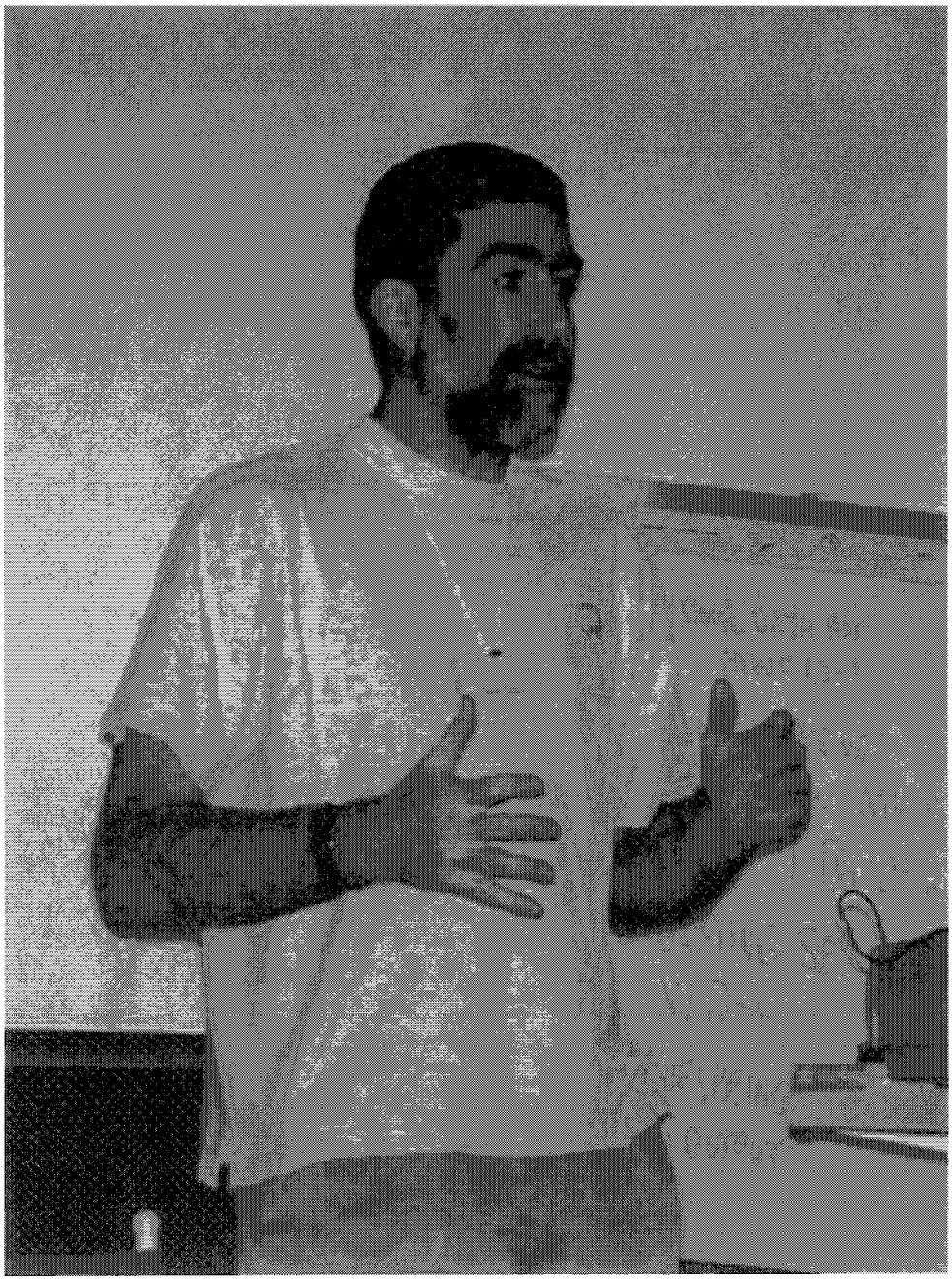
Statistics: American Statistician, Annals of Statistics, Applied Statistics, Australian Journal of Statistics, Biometrics, Indian Journal of Statistics, International Statistical Review, Journal of Computational and Graphical Statistics, Journal of Educational Statistics, Journal of the American Statistical Association, Journal of the Royal Statistical Society Scandinavian Journal of Statistics, Statistica Sinica, Statistical Science;

Genetics: American Journal of Human Genetics, Annals of Human Genetics, Bioinformatics, Cytogenetics and Cell Genetics, Genetic Epidemiology, Genetics, Genome Research, Genomics, Human Heredity, Journal of Molecular Evolution, Nature Reviews Genetics, Nucleic Acids Research, Theoretical and Applied Genetics;

Others: Canadian Journal of Fisheries and Aquatic Sciences, Electrophoresis, European Journal of Immunology, Infection and Immunity, Investigative Ophthalmology, Journal of Applied Bacteriology, Journal of Computational Biology, Journal of Immunological Methods, Journal of Solid State Chemistry, Molecular and Biochemical Parasitology, Molecular Vision, Neuron, Proceedings of the National Academy of Sciences USA, Sociological Methodology, Statistics for the Environment, Survey Methodology, Theoretical Population Biology.

The response to invitations to submit contributions was tremendous, and has resulted in this very diverse collection of papers. They address topics in many of the areas in which Terry has had an interest during some part of his highly varied career. He has had great impact in the development and progress of several of these fields, most recently in experimental design and statistical analysis of microarray studies of gene expression. This volume contains refereed papers, roughly organized by topic, on probability, algebraic experimental design, generalized linear models, statistical education, and assorted applications, including the US census, fire risk assessment, genetics and other biological applications. We all hope that every reader will find something of interest and will benefit from this access to the broad spectrum of scientific and mathematical/statistical ideas represented here.

Happy Birthday, Terry!



List of Contributors

Roxana Alexandridis, Ohio State University
Norman Arnheim, University of Southern California
R. A. Bailey, Queen Mary, University of London, U. K.
John W. Benoit, USDA, Forest Service
David R. Brillinger, University of California, Berkeley
Karl W. Broman, Johns Hopkins University
Sandrine Dudoit, University of California, Berkeley
Steven N. Evans, University of California, Berkeley
David A. Freedman, University of California, Berkeley
Jane Fridlyand, University of California, San Francisco
Sabrina Giglio, Diagnostica e Ricerca, Milan, Italy
Darlene R. Goldstein, Institut Suisse de Recherche Expérimentale sur le Cancer and École Polytechnique Fédérale de Lausanne, Switzerland
Rudy Guerra, Rice University
Jian Han, PPD, Inc.
Mark H. Hansen, Lucent Technologies
Rafael A. Irizarry, Johns Hopkins University
Harri T. Kiiveri, CSIRO, Australia
David H. Ledbetter, University of Chicago
Lei Li, University of Southern California
Shili Lin, Ohio State University
John H. Maindonald, Australian National University
Christa Lese Martin, University of Chicago
Naomichi Matsumoto, Nagasaki University School of Medicine, Japan
Mary Sara McPeek, University of Chicago
William C. Navidi, Colorado School of Mines
David O. Nelson, Lawrence Livermore National Laboratory
Deborah A. Nolan, University of California,Berkeley
Anthony G. Pakes, University of Western Australia
Jim Pitman, University of California, Berkeley
Yvonne Pittelkow, Australian National University
Cheryl E. Praeger, University of Western Australia
Haiganoush K. Preisler, USDA, Forest Service
Jessica A. Roseberry, University of Chicago
Csaba Schneider, The University of Western Australia
Gordon K. Smyth, Walter and Eliza Hall Institute of Medical Research, Australia
Andrew L. Strahs, Harvard School of Public Health
Ning Sun, Yale University School of Medicine
Simon Tavaré, University of Southern California
Natalie P. Thorne, Walter and Eliza Hall Institute of Medical Research, Australia
Kenneth W. Wachter, University of California, Berkeley

James L. Weber, Marshfield Medical Research Foundation
Susan Wilson, Australian National University
Baolin Wu, Yale University School of Medicine
Yee Hwa Yang, University of California, San Francisco
Bin Yu, University of California, Berkeley
Hongyu Zhao, Yale University School of Medicine
Orsetta Zuffardi, Università di Pavia, Italy

Acknowledgements

First, thanks to the IMS for support of this undertaking. Particular thanks are due to Elyse Gustafson, Patrick Kelly, and, most significantly, Joel Greenhouse, who provided guidance, advice and encouragement in the transformation of the original idea into the book you are reading now. The compilation of the volume was made much easier than it had any right to be – your efforts are very greatly valued.

Deepest thanks are due the referees, who gave careful, thoughtful input in reviews of the submitted papers. In addition, the cooperation of authors in producing the final volume is gratefully acknowledged. Thanks also to the LaTeX Development Team, whose website `http://www.latex-project.org/` provided invaluable aid during the finalization of the manuscripts.

A note of appreciation goes to Sandrine Dudoit, colleague, friend, and co-conspirator from the word go. Without her ideas and assistance, this project would not have been realized. *Je n'ai pas encore rendu la monnaie de ta pièce!*

The last words of recognition belong to the Mountford boys (Tom, Aneurin, Tavis, and Ian) for their forbearance and sustenance throughout this project, and most especially during the final push.

Table of Contents

Poisson-Kingman Partitions

Jim Pitman

Abstract

This paper presents some general formulas for random partitions of a finite set derived by Kingman's model of random sampling from an interval partition generated by subintervals whose lengths are the points of a Poisson point process. These lengths can be also interpreted as the jumps of a subordinator, that is an increasing process with stationary independent increments. Examples include the two-parameter family of Poisson-Dirichlet models derived from the Poisson process of jumps of a stable subordinator. Applications are made to the random partition generated by the lengths of excursions of a Brownian motion or Brownian bridge conditioned on its local time at zero.

Keywords: exchangeable; stable; subordinator; Poisson-Dirichlet; distribution

1 Introduction

This paper presents some general formulas for random partitions of a finite set derived by Kingman's model of random sampling from an interval partition generated by subintervals whose lengths are the points of a Poisson point process. Instances and variants of this model have found applications in the diverse fields of population genetics [17, 19], combinatorics [4, 48], Bayesian statistics [23], ecology [15, 37], statistical physics [11, 12, 13, 53, 55], and computer science [25].

Section 2 recalls some general results for partitions obtained by sampling from a random discrete distribution. These results are then applied in Section 3 to the Poisson-Kingman model. Section 4 discusses three basic operations on Poisson-Kingman models: scaling, exponential tilting, and deletion of classes. Section 5 then develops formulas for specific examples of Poisson-Kingman models. Section 6 recalls the two-parameter family of Poisson-Dirichlet models derived in [50] from the Poisson process of jumps of a stable(α) subordinator for $0 < \alpha < 1$. Section 7 reviews some results of [41, 46, 49, 50] relating the two-parameter family to the lengths of excursions of a Markov process whose zero set is the range of a stable subordinator of index α. Section 8 provides further detail in the case $\alpha = \frac{1}{2}$ which corresponds to partitioning a time interval by the lengths of excursions of a Brownian motion. As shown in [2, 3], it is this stable($\frac{1}{2}$) model which governs the asymptotic distribution of partitions derived in various ways from random forests, random mappings, and the additive coalescent. See also [5, 9] for further developments in terms of Brownian paths, and [10, 25] for

applications to hashing and parking algorithms. This paper is a revision of the earlier preprint [42]. See [48] for a broader context and further developments.

2 Preliminaries

This section recalls some basic ideas from Kingman's theory of exchangeable random partitions [30, 31], as further developed in [43]. See [45, 48] for more extensive reviews of these ideas and their applications. Except where otherwise specified, all random variables are assumed to be defined on some background probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and $\mathbb{E}$ denotes expectation with respect to $\mathbb{P}$. Let $\mathbb{N} := \{1, 2, \ldots\}$, let F denote a random probability distribution on the line, and let Π be a random partition of $\mathbb{N}$ *generated by sampling from* F. That is to say, two positive integers i and j are in the same block of Π iff $X_i = X_j$, where conditionally given F the X_i are independent and identically distributed according to F. Formally, Π is identified with the sequence (Π_n), where Π_n is the restriction of Π to the finite set $\mathbb{N}_n := \{1, \ldots, n\}$. The distribution of Π_n is such that for each particular partition $\{A_1, \cdots, A_k\}$ of $\mathbb{N}_n$ with $\#(A_i) = n_i$ for $1 \leq i \leq k$, where $n_i \geq 1$ and $\sum_{i=1}^{k} n_i = n$,

$$\mathbb{P}(\Pi_n = \{A_1, \cdots, A_k\}) = p(n_1, \cdots, n_k) \tag{1}$$

for some symmetric function p of sequences of positive integers, called the *exchangeable partition probability function (EPPF)* of Π. Conversely, Kingman [30, 31] showed that if Π is an exchangeable random partition of $\mathbb{N}$, meaning that the distribution of its restrictions Π_n is of the form (1) for every n, for some symmetric function p, then Π has the same distribution as if generated by sampling from some random probability distribution F. Let P_i denote the size of the ith largest atom of F. If F is a random discrete distribution, then $\sum_i P_i = 1$ almost surely, and Π is said to have *proper frequencies* (P_i). In that case, let $\tilde{P}_j$ denote the size of the jth atom discovered in the process of random sampling. Put another way, $\tilde{P}_j$ is the asymptotic frequency of the jth class of Π when the classes are put in order of their least elements. It is assumed now for simplicity that $P_i > 0$ for all i almost surely, and hence $\tilde{P}_j > 0$ for all j almost surely. The sequence $(\tilde{P}_j)$ is a *size-biased permutation* of (P_i). That is to say, $\tilde{P}_j = P_{\pi_j}$ where for all finite sequences $(i_j, 1 \leq j \leq k)$ of distinct positive integers, the conditional probability of the event $(\pi_j = i_j$ for all $1 \leq j \leq k)$ given $(P_1, P_2, \ldots)$ is

$$P_{i_1} \frac{P_{i_2}}{1 - P_{i_1}} \cdots \frac{P_{i_k}}{1 - P_{i_1} - \ldots - P_{i_{k-1}}}. \tag{2}$$

The distribution of Π_n is determined by the distribution of the sequence of ranked frequencies (P_i) through the distribution of the size-biased permutation $(\tilde{P}_j)$. To be precise, the EPPF p in (1) is given by the formula [43]

$$p(n_1, \cdots, n_k) = \mathbb{E}\left[\prod_{i=1}^{k} \tilde{P}_i^{n_i - 1} \prod_{i=1}^{k-1} \left(1 - \sum_{j=1}^{i} \tilde{P}_j\right)\right]. \tag{3}$$

Alternatively [45]

$$p(n_1,\cdots,n_k) = \sum_{(j_1,\ldots,j_k)} \mathbb{E}\prod_{i=1}^{k} P_{j_i}^{n_i} \tag{4}$$

where $(j_1,\ldots,j_k)$ ranges over all permutations of k positive integers, and the same formula holds with P_{j_i} replaced by $\tilde{P}_{j_i}$. For each $n = 1,2,\cdots$ the EPPF p, when restricted to $(n_1,\cdots,n_k)$ with $\sum_i n_i = n$, determines the distribution of Π_n. Since Π_n is the restriction of Π_{n+1} to $\mathbb{N}_n$, the EPPF is subject to the following sequence of *addition rules* [43]: for $k = 1,2,\ldots$

$$p(n_1,\cdots,n_k) = \sum_{j=1}^{k} p(\ldots,n_j+1,\ldots) + p(n_1,\ldots,n_k,1) \tag{5}$$

where $(\ldots,n_j+1,\ldots)$ is derived from $(n_1,\ldots,n_k)$ by substituting n_j+1 for n_j. The first few rules are

$$1 = p(1) = p(2) + p(1,1) \tag{6}$$

$$p(2) = p(3) + p(2,1); \quad p(1,1) = 2p(2,1) + p(1,1,1) \tag{7}$$

where $p(2,1) = p(1,2)$ by symmetry of p. Let $\mu(q)$ denote the qth moment of $\tilde{P}_1$:

$$\mu(q) := \mathbb{E}[\tilde{P}_1^q] = \int_0^1 p^q \,\tilde{\nu}(dp), \tag{8}$$

where $\tilde{\nu}$ denotes the distribution of $\tilde{P}_1$ on $(0,1]$. Following Engen [15], call $\tilde{\nu}$ the *structural distribution* associated with an random discrete distribution whose size-biased permutation is $(\tilde{P}_j)$, or with an exchangeable random partition Π whose sequence of class frequencies is $(\tilde{P}_j)$. The special case of (3) for $k = 1$ and $n_1 = n$ is

$$p(n) = \mathbb{E}[\tilde{P}_1^{n-1}] = \mu(n-1) \quad (n = 1,2,\cdots). \tag{9}$$

From (6), (7), and (9) the following values of the EPPF are also determined by the first two moments of the structural distribution:

$$p(1,1) = 1-\mu(1); \;\; p(2,1) = \mu(1)-\mu(2); \;\; p(1,1,1) = 1-3\mu(1)+2\mu(2). \tag{10}$$

So the distribution of the random partition of $\{1,2,3\}$ induced by Π with class frequencies $(\tilde{P}_i)$ is determined by the first two moments of the structural distribution of $\tilde{P}_1$. It is not true in general that the EPPF is determined for all $(n_1,\cdots,n_k)$ by the structural distribution, because it is possible to construct different distributions for a sequence of ranked frequencies which have the same structural distribution.

Continuing to suppose that (P_i) is the sequence of ranked atoms of a random discrete probability distribution, and that $(\tilde{P}_j)$ is a size-biased permutation of (P_i), for an arbitrary non-negative measurable function f, there is the well known formula

$$\mathbb{E}\left[\sum_i f(P_i)\right] = \mathbb{E}\left[\sum_j f(\tilde{P}_j)\right] = \mathbb{E}\left[\frac{f(\tilde{P}_1)}{\tilde{P}_1}\right] = \int_0^1 \frac{f(p)}{p}\tilde{\nu}(dp). \tag{11}$$

This formula shows that the structural distribution $\tilde{\nu}$ encodes much information about the entire sequence of random frequencies. Taking f in (11) to be the indicator of a subset B of $(0,1]$, the quantity in (11) is $\nu(B) = \int_B p^{-1}\tilde{\nu}(dp)$. This measure ν is the mean intensity measure of the point process with a point at each $P_i \in (0,1]$. For $x > \frac{1}{2}$ there can be at most one $P_i > x$, so the structural distribution $\tilde{\nu}$ determines the distribution of $P_1 = \max_j \tilde{P}_j$ on $(\frac{1}{2}, 1]$ via the formula

$$\mathbb{P}(P_1 > x) = \nu(x,1] = \int_{(x,1]} p^{-1}\tilde{\nu}(dp) \qquad (x > \tfrac{1}{2}). \tag{12}$$

Typically, formulas for $\mathbb{P}(P_1 > x)$ get progressively more complicated on the intervals $(\frac{1}{3}, \frac{1}{2}]$, $(\frac{1}{4}, \frac{1}{3}]$, $\cdots$. See for instance [40, 50].

A random variable of interest in many applications is the sum of mth powers of frequencies

$$S_m := \sum_{i=1}^{\infty} P_i^m = \sum_{j=1}^{\infty} \tilde{P}_j^m \qquad (m = 1,2,\ldots)$$

where it is still assumed that $S_1 = 1$ almost surely. Let $\pi := \{A_1, \cdots, A_k\}$ be some particular partition of $\mathbb{N}_n$ with $\#(A_i) = n_i$ for $1 \le i \le k$, and consider the event $(\Pi_n \ge \pi)$, meaning that each block of Π_n is some union of blocks of π. Then it is easily shown that

$$\mathbb{P}(\Pi_n \ge \pi) = \mathbb{E}\left[\prod_{i=1}^{k} S_{n_i}\right] = \sum_{j=1}^{k} \sum_{\{B_1,\ldots,B_j\}} p(n_{B_1}, \ldots, n_{B_j}) \tag{13}$$

where the second sum is over partitions $\{B_1, \ldots, B_j\}$ of $\mathbb{N}_k$, and $n_B := \sum_{i \in B} n_i$. In particular, for $n_i \equiv m$ this gives an expression for the kth moment of S_m for each $k = 1,2,\ldots$:

$$\mathbb{E}\left[S_m^k\right] = \sum_{j=1}^{k} \frac{1}{j!} \sum_{(k_1,\ldots,k_j)} \frac{k!}{k_1! \cdots k_j!}\, p(mk_1, \ldots, mk_j) \tag{14}$$

where the second sum is over all sequences of j positive integers $(k_1, \ldots, k_j)$ with $k_1 + \cdots + k_j = k$. Thus the EPPF associated with a random discrete distribution directly determines the positive integer moments of the power sums S_m, hence the distribution of S_m, for each m.

3 The Poisson-Kingman Model

Following McCloskey [37], Kingman [29], Engen [15], Perman-Pitman-Yor [40, 41, 50], consider the ranked random discrete distribution $(P_i) := (J_i/T)$ derived from an inhomogeneous Poisson point process of random lengths $J_1 \ge J_2 \ge \cdots \ge 0$ by normalizing these lengths by their sum $T := \sum_{i=1}^{\infty} J_i$. So it is assumed that the number N_I of J_i that fall in an interval I is a Poisson variable with mean $\Lambda(I)$, for some Lévy measure

Λ on $(0,\infty)$, and the counts $N_{I_1},\cdots,N_{I_k}$ are independent for every finite collection of disjoint intervals $I_1,\cdots,I_k$. It is also assumed that

$$\int_0^1 x\Lambda(dx) < \infty \text{ and } \Lambda[1,\infty) < \infty$$

to ensure that $\mathbb{P}(T<\infty)=1$. The sequence (P_i) may be regarded as a random element of the space $\mathcal{P}^{\downarrow}$ of decreasing sequences of positive real numbers with sum 1. Throughout this section, the following further assumption is made to ensure that various conditional probabilities can be defined without quibbling about null sets:

Regularity assumption. *The Lévy measure Λ has a density $\rho(x)$ such that the distribution of T is absolutely continuous with density*

$$f(t) := \mathbb{P}(T \in dt)/dt$$

which is strictly positive and continuous on $(0,\infty)$.

Note that the regularity assumption implies the total mass of the Lévy measure is infinite:

$$\int_0^\infty \rho(x)dx = \infty. \tag{15}$$

The results described below also have weaker forms for a Lévy density $\rho(x)$ just subject to (15), with appropriate caveats about almost everywhere defined conditional probabilities.

It is well known that f is uniquely determined by ρ via the Laplace transform

$$\mathbb{E}(e^{-\lambda T}) = \int_0^\infty e^{-\lambda x} f(x)dx = \exp[-\psi(\lambda)] \qquad (\lambda \geq 0) \tag{16}$$

where, according to the Lévy-Khintchine formula,

$$\psi(\lambda) = \int_0^\infty (1-e^{-\lambda x})\rho(x)dx. \tag{17}$$

Alternatively, f is the unique solution of the following integral equation, which can be derived from (16) and (17) by differentiation with respect to λ:

$$f(t) = \int_0^t \rho(v) f(t-v)\frac{v}{t}dv. \tag{18}$$

Let $(\tilde{P}_j)$ be a size-biased permutation of the normalized lengths $(P_i) := (J_i/T)$ and let $(\tilde{J}_j) = (T\tilde{P}_j)$ be the corresponding size-biased permutation of the ranked lengths (J_i). Then (18) admits the following probabilistic interpretation [37, 41]:

$$\mathbb{P}(\tilde{J}_1 \in dv, T \in dt) = \rho(v)dv f(t-v)dt\frac{v}{t}. \tag{19}$$

This can be understood as follows. The left side of (19) is the probability that among the Poisson lengths there is some length in dv near v, and the sum of the rest of the

lengths falls in an interval of length dt near $t-v$, and finally that the interval of length about v is the one picked by length-biased sampling. Formally, (19) is justified by the description of a Poisson process in terms of its Palm measures [41].

The following two Lemmas are read from [41, Theorem 2.1]. The first Lemma is immediate from (19), and the second is obtained by a similar Palm calculation.

Lemma 1
[41] *For each $t > 0$ the formula*

$$\tilde{f}(p\,|\,t) := pt\,\rho(pt)\,\frac{f(\overline{p}t)}{f(t)} \qquad (0 < p < 1;\ \overline{p} := 1-p), \tag{20}$$

where ρ is the density of the Lévy measure of T and f is the probability density of T, defines a function of p which is a probability density on $(0,1)$. This is the density of the structural distribution of $\tilde{P}_1 := \tilde{J}_1/T$ given $T = t$:

$$\mathbb{P}(\tilde{P}_1 \in dp\,|\,t) = \tilde{f}(p\,|\,t)dp \qquad (0 < p < 1). \tag{21}$$

Lemma 2
[41] *For $j = 0,1,2,\cdots$ let*

$$T_j := T - \sum_{k=1}^{j} \tilde{J}_k = \sum_{k=j+1}^{\infty} \tilde{J}_k \tag{22}$$

which is the total length remaining after removal of the first j Poisson lengths $\tilde{J}_1,\ldots,\tilde{J}_j$ chosen by length-biased sampling. Then the family of densities (20) on $(0,1)$, parameterized by $t > 0$, provides the conditional density of the random variable

$$G_{j+1} := \frac{\tilde{J}_{j+1}}{T_j} = \frac{\tilde{P}_{j+1}}{\tilde{P}_{j+1} + \tilde{P}_{j+2} + \cdots}$$

given $T_0,\cdots,T_j$ via the formula

$$\mathbb{P}\left(G_{j+1} \in dp\,\middle|\,T_0,\cdots,T_j\right) = \tilde{f}(p|T_j)\,dp \qquad (0 < p < 1). \tag{23}$$

Lemma 2 provides an explicit construction of a regular conditional distribution for $(\tilde{P}_j)$ given $T = t$ for arbitrary $t > 0$. This conditional distribution of $(\tilde{P}_j)$ given $T = t$ determines corresponding conditional distributions for the $\mathcal{P}^{\downarrow}$-valued ranked sequence (P_i) and for an associated random partition Π of $\mathbb{N}$.

Definition 3
The distribution of $(P_i) := (J_i/T)$ on $\mathcal{P}^{\downarrow}$ determined by the ranked points J_i of a Poisson process with Lévy density ρ will be called the *Poisson-Kingman distribution with Lévy density* ρ, and denoted $\mathrm{PK}(\rho)$. Denote by $\mathrm{PK}(\rho\,|\,t)$ the regular conditional distribution of (P_i) given $(T = t)$ constructed above. For a probability distribution γ on $(0,\infty)$, let

$$\mathrm{PK}(\rho,\gamma) := \int_0^{\infty} \mathrm{PK}(\rho\,|\,t)\gamma(dt) \tag{24}$$

be the distribution on $\mathcal{P}^{\downarrow}$ obtained by mixing the $\mathrm{PK}(\rho\,|\,t)$ with respect to $\gamma(dt)$. Call $\mathrm{PK}(\rho,\gamma)$ the *Poisson-Kingman distribution with Lévy density* ρ *and mixing distribution* γ.

Note that $\mathrm{PK}(\rho\,|\,t) = \mathrm{PK}(\rho,\delta_t)$, where δ_t is a unit mass at t, and that $\mathrm{PK}(\rho) = \mathrm{PK}(\rho,\gamma)$ for $\gamma(dt) = f(t)dt$. A formula for the joint density of $(P_1,\cdots,P_n)$ for (P_i) with $\mathrm{PK}(\rho\,|\,t)$ distribution was obtained by Perman [40] in terms of the joint density $p_1(t,x)$ of T and J_1. This function can be described in terms of ρ and f as the solution of an integral equation [40], or as a series of repeated integrals [50]. But this formula will not be used here.

For a probability distribution Q on $\mathcal{P}^{\downarrow}$, such as $Q = \mathrm{PK}(\rho,\gamma)$, a random partition Π of $\mathbb{N}$ will be called a *Q-partition* if Π is an exchangeable random partition of $\mathbb{N}$ whose ranked class frequencies are distributed according to Q. Immediately from Definition 3, the structural distribution of a $\mathrm{PK}(\rho,\gamma)$-partition Π of $\mathbb{N}$, that is the distribution on $(0,1)$ of the frequency $\tilde{P}_1$ of the class of Π containing 1, has density

$$\mathbb{P}(\tilde{P}_1 \in dp)/dp = \int_0^\infty \tilde{f}(p\,|\,t)\gamma(dt) \quad (0 < p < 1) \tag{25}$$

where $\tilde{f}(p\,|\,t)$ given by (20) is the density of the structural distribution of $\tilde{P}_1$ given $T = t$ in the basic Poisson construction. Similarly, the EPPF of Π is

$$p(n_1,\cdots,n_k) = \int_0^\infty p(n_1,\cdots,n_k\,|\,t)\gamma(dt) \tag{26}$$

where $p(n_1,\cdots,n_k\,|\,t)$, the EPPF of a $\mathrm{PK}(\rho\,|\,t)$-partition, is determined as follows:

Theorem 4
The EPPF of a $\mathrm{PK}(\rho\,|\,t)$*-partition is given by the formula*

$$p(n_1,\cdots,n_k\,|\,t) = t^{k-1}\int_0^1 p^{n+k-2}I(n_1,\cdots,n_k;tp)\tilde{f}(p\,|\,t)dp \tag{27}$$

where $n = \sum_1^k n_i$, $I(n;v) = 1$ *if* $k = 1$ *and* $n_1 = n$, *and for* $k = 2,3,\ldots$

$$I(n_1,\cdots,n_k;v) := \frac{1}{\rho(v)}\int_{S_k}\left[\prod_{i=1}^k \rho(vu_i)u_i^{n_i}\right]du_1\cdots du_{k-1} \tag{28}$$

where S_k *is the simplex* $\{(u_1,\ldots,u_k) : u_i \geq 0 \text{ and } u_1+\cdots+u_k = 1\}$.

Proof. In view of the formula (20) for $\tilde{f}(p\,|\,t)$, the formula (27) is obtained from formula (31) in the following Lemma by dividing by $f(t)dt$, letting $p = \sum_i x_i/t$, and integrating out with respect to p and to $u_i = x_i/(pt)$ for $1 \leq i \leq k-1$. □

A change of variables gives the following variant of formula (27), whose connection to the next lemma is a bit more obvious:

$$p(n_1,\cdots,n_k\,|\,t) = \int_0^t dv\frac{f(t-v)}{t^n f(t)}v^{n+k-1}I(n_1,\ldots,n_k;v)\rho(v). \tag{29}$$

Lemma 5
Let Π_n be the restriction to $\mathbb{N}_n$ of a PK(ρ) *partition Π whose class frequencies (in order of least elements) are $\tilde{P}_j = \tilde{J}_j/T$, where $T = \sum_j \tilde{J}_j$ has density f, and the lengths $\tilde{J}_j$ are the points of a Poisson process of lengths with intensity ρ, in length-biased random order. Then for each partition $\{A_1, \cdots, A_k\}$ of $\mathbb{N}_n$ such that $\#(A_i) = n_i$ for $1 \le i \le k$,*

$$\mathbb{P}(\Pi_n = \{A_1, \cdots, A_k\}, \tilde{J}_i \in dx_i \text{ for } 1 \le i \le k, T \in dt) \tag{30}$$

$$= t^{-n} f(t - \textstyle\sum_{i=1}^k x_i)\, dt \prod_{i=1}^k \rho(x_i) x_i^{n_i}\, dx_i. \tag{31}$$

Proof. This can be derived by evaluation of the expectation (3) for the joint distribution of $\tilde{P}_1, \ldots, \tilde{P}_k$ given $T = t$ determined by Lemma 2. Alternatively, there is the following more intuitive argument, which can be made rigorous using the characterization of Poisson process by its a Palm measures, as in [49, 41]. Let Π be constructed as in [46] using random intervals I_i laid down on $[0, T]$ in some arbitrary random order, where the lengths $J_i := |I_i|$ are the ranked points of the Poisson process with intensity $\rho(x)$, and $T = \sum_i J_i$. Let $U_1, U_2, \cdots$ be i.i.d. uniform on $(0, 1)$ independent of this construction. Let Π be the partition of $\mathbb{N}$ generated by the random equivalence relation $n \sim m$ iff either $n = m$ or TU_n and TU_m fall in the same interval I_i for some i. Then by construction, Π is a PK(ρ) partition. For the event in (30) to occur,

(i) there must be some Poisson point in dx_i for each $1 \le i \le k$, and

(ii) given (i), the sum of the rest of the Poisson points must fall in an interval of length dt near $t - \sum_{i=1}^k x_i$, and

(iii) given (i) and (ii), for each $1 \le i \le k$ and each $m \in A_i$ the sample point TU_m must fall in the interval of length x_i.

The infinitesimal probability in (30) therefore equals

$$\left(\prod_{i=1}^k \rho(x_i)\, dx_i \right) f(t - \textstyle\sum_{i=1}^k x_i)\, dt \prod_{i=1}^k \left(\frac{x_i}{t} \right)^{n_i} \tag{32}$$

which rearranges as (31). □

The formula (27) expresses $p(n_1, \cdots, n_k | t)$ as the expectation of a function of $\tilde{P}_1$ given $T = t$, where the function depends on t and $n_1, \cdots, n_k$. Because some values of an EPPF can always be expressed as moments of $\tilde{P}_1$, as in (8) and (10), it seems natural to try to express an EPPF similarly whenever possible. This idea serves as a guide to simplifying calculations in a number of particular cases treated later. The integrations in (27) and (28) are essentially convolutions, which can be expressed or evaluated in various ways. Consider for instance the length $T_k := T - \sum_{i=1}^k \tilde{J}_i$ which remains after removal of the first k lengths discovered by the sampling process. Then the formula of Lemma 5 can be recast as

$$\mathbb{P}(\Pi_n = \{A_1, \cdots, A_k\}, \tilde{J}_i \in dx_i \text{ for } 1 \le i \le k, T_k \in dv) \tag{33}$$

$$= (v+\textstyle\sum_{i=1}^{k} x_i)^{-n} f(v)dv \prod_{i=1}^{k} \rho(x_i) x_i^{n_i} dx_i \tag{34}$$

which yields the following integrated forms of (27):

Corollary 6
The EPPF of a PK(ρ)*-partition is given by the formula*

$$p(n_1,\cdots,n_k) = \int_0^\infty \cdots \int_0^\infty \frac{f(v)dv \prod_{i=1}^{k} \rho(x_i) x_i^{n_i} dx_i}{(v+\sum_{i=1}^{k} x_i)^n} \tag{35}$$

where $n := \sum_{i=1}^{k} n_i$, *or again by*

$$p(n_1,\cdots,n_k) = \frac{(-1)^{n-k}}{\Gamma(n)} \int_0^\infty \lambda^{n-1} d\lambda e^{-\psi(\lambda)} \prod_{i=1}^{k} \psi_{n_i}(\lambda) \tag{36}$$

where $\psi(\lambda) := \int_0^\infty (1-e^{-\lambda x})\rho(x)dx$ *is the Laplace exponent as in* (17), *and*

$$\psi_m(\lambda) := \frac{d^m}{d\lambda^m}\psi(\lambda) = (-1)^{m-1} \int_0^\infty x^m e^{-\lambda x} \rho(x)dx \quad (m = 1,2,\ldots). \tag{37}$$

Proof. Formula (34) yields (35) by integration, and (36) follows after applying the formula $b^{-n} = \Gamma(n)^{-1} \int_0^\infty \lambda^{n-1} e^{-\lambda b} d\lambda$ to $b = v + \sum_{i=1}^{k} x_i$. □

These integrated forms (35) and (36) also hold more generally, with $f(v)dv$ replaced by $\mathbb{P}(T \in dv)$, and $\rho(x)dx$ replaced by the corresponding Lévy measure on $(0,\infty)$, assuming only that the Lévy measure has infinite total mass.

Provided $\mathbb{E}(e^{\varepsilon T}) < \infty$ for some $\varepsilon > 0$, the Laplace exponent ψ can be expanded in a neighbourhood of 0 as

$$\psi(\lambda) = -\sum_{m=1}^{\infty} \frac{\kappa_m}{m!}(-\lambda)^m$$

where the *cumulants* κ_m of T are the moments of the Lévy measure

$$\kappa_m = (-1)^{m-1}\psi_m(0) = \int_0^\infty x^m \rho(x)dx.$$

Then for each partition $\{A_1,\cdots,A_k\}$ of $\mathbb{N}_n$ such that $\#(A_i) = n_i$ for $1 \le i \le k$, Lemma 5 yields the formula

$$\mathbb{P}(\Pi_n = \{A_1,\cdots,A_k\}, T \in dt) = t^{-n}\mathbb{P}(T + \Sigma_{i=1}^{k} J_{i,n_i} \in dt) \prod_{i=1}^{k} \kappa_{n_i} \tag{38}$$

where J_{i,n_i} denotes a random length distributed according to the Lévy density tilted by x^{n_i}:

$$\mathbb{P}(J_{i,n_i} \in dx) = \kappa_{n_i}^{-1} \rho(x) x^{n_i} dx$$

and T and the J_{i,n_i} for $1 \leq i \leq k$ are assumed to be independent. If $f_{n_1,\ldots,n_k}(t)$ denotes the probability density of $T + \Sigma_{i=1}^{k} J_{i,n_i}$, then formula (27) for the EPPF of a $\mathrm{PK}(\rho \,|\, t)$-partition can be rewritten

$$p(n_1, \cdots, n_k \,|\, t) = \frac{f_{n_1,\ldots,n_k}(t)}{t^n f(t)} \prod_{i=1}^{k} \kappa_{n_i} \tag{39}$$

and formula (35) for the EPPF of a $\mathrm{PK}(\rho)$-partition becomes

$$p(n_1, \cdots, n_k) = \mathbb{E}\left[(T + \Sigma_{i=1}^{k} J_{i,n_i})^{-n} \right] \prod_{i=1}^{k} \kappa_{n_i}. \tag{40}$$

See also James [23] for closely related formulas, with applications to Bayesian nonparametric inference.

4 Operations

Later discussion of specific examples of Poisson-Kingman partitions will be guided by a number of basic operations on Lévy densities ρ and their associated families of partitions.

4.1 Scaling

By an obvious scaling argument, the $\mathrm{PK}(\rho)$ and $\mathrm{PK}(\rho')$ distributions are identical whenever $\rho'(x) = b\rho(bx)$ is a rescaling of ρ for some $b > 0$. The converse is less obvious, but true [49, Lemma 7.5].

4.2 Exponential tilting

It is elementary that if ρ is a Lévy density, corresponding to a density f for T, and b is a real number such that $\psi(b)$ defined by (17) is finite, then

$$\rho^{(b)}(x) = \rho(x) e^{-bx} \tag{41}$$

is also a Lévy density, and the corresponding density of T is

$$f^{(b)}(t) = f(t)\, e^{\psi(b) - bt} \tag{42}$$

It is also well known [34, Proposition 2.1.3] that if $\mathbb{P}^{(b)}$ denotes the probability distribution governing the Poisson set up with Lévy density $\rho^{(b)}$ then (42) extends to the absolute continuity relation

$$\frac{d\mathbb{P}^{(b)}}{d\mathbb{P}^{(0)}} = e^{\psi(b) - bT}. \tag{43}$$

This relation is equivalent to a combination of (42) and the following identity, which can also be verified using the construction of Lemma 2:

$$\mathrm{PK}(\rho^{(b)}\,|\,t) = \mathrm{PK}(\rho\,|\,t) \text{ for all } t > 0. \tag{44}$$

Consequently

$$\mathrm{PK}(\rho^{(b)}, \gamma) = \mathrm{PK}(\rho, \gamma) \tag{45}$$

for every γ. In particular, the distribution on $\mathcal{P}^{\downarrow}$ derived from the unconditioned Poisson model with Lévy density $\rho^{(b)}$ is

$$\mathrm{PK}(\rho^{(b)}) = \mathrm{PK}(\rho, \gamma^{(b)}) \tag{46}$$

where $\gamma^{(b)}$ is the $\mathbb{P}^{(b)}$ distribution of T, that is $\gamma^{(b)}(dt) = f^{(b)}(t)dt$ for $f^{(b)}$ as in (42). It can also be shown that if ρ' and ρ are two regular Lévy densities such that $\mathrm{PK}(\rho') = \mathrm{PK}(\rho, \gamma)$ for some γ, then $\rho' = \rho^{(b)}$ and $\gamma = \gamma^{(b)}$ for some b.

4.3 Deletion of Classes

The following proposition, which generalizes a result of [41], provides motivation for study of $\mathrm{PK}(\rho, \gamma)$-partitions for other distributions γ besides $\gamma(dt) = f(t)dt$ corresponding to the unconditioned Poisson set up, and $\gamma = \delta_t$ corresponding to conditioning on $T = t$. Given a random partition Π of $\mathbb{N}$ with infinitely many classes, for each $k = 0, 1, \cdots$ let Π_k be the partition of $\mathbb{N}$ derived from Π by *deletion of the first k classes*, an operation made precise as follows. First let Π'_k be the restriction of Π to $H_k := \mathbb{N} - G_1 - \cdots - G_k$ where $G_1, \cdots G_k$ are the first k classes of Π in order of least elements, then derive Π_k on $\mathbb{N}$ from Π'_k on H_k by renumbering the points of H_k in increasing order.

Proposition 7
Let Π be a $\mathrm{PK}(\rho, \gamma)$-partition of $\mathbb{N}$, and let Π_k be derived from Π be deletion of its first k classes. Then Π_k is a $\mathrm{PK}(\rho, \gamma_k)$-partition of $\mathbb{N}$, where $\gamma_k = \gamma Q^k$ for Q the Markov transition operator on $(0, \infty)$

$$Q(t, dv) = \rho(t - v)(t - v)t^{-1} f(v) 1(0 < v < t)dv.$$

In particular, if Π is a $\mathrm{PK}(\rho)$ partition of $\mathbb{N}$, then Π_k is $\mathrm{PK}(\rho, \gamma_k)$-partition of $\mathbb{N}$, where γ_k is the distribution of T_k, the total sum of Poisson lengths T minus the sum of the first k lengths discovered by a process of length-biased sampling, as in (22).

Proof. According to a result of [41] which is implicit in Lemma 2, the sequence (T_k) is Markov with stationary transition probabilities given by Q. The conclusion follows from this observation, the construction of $\mathrm{PK}(\rho; \gamma)$, and the general construction of an exchangeable partition of $\mathbb{N}$ conditionally given its class frequencies [43].

5 Examples

5.1 The one-parameter Poisson-Dirichlet distribution

Following Kingman [29], for the particular choice

$$\rho(x) = \theta x^{-1} e^{-bx} \tag{47}$$

where $\theta > 0$ and $b > 0$, corresponding to T with the gamma(θ, b) density

$$f(t) = \frac{b^\theta}{\Gamma(\theta)} t^{\theta-1} e^{-bt}, \tag{48}$$

the PK(ρ) distribution is the *Poisson-Dirichlet distribution with parameter* θ, abbreviated PD(θ). Note the lack of dependence on the inverse scale parameter b. The well known fact the structural distribution of PD(θ) is beta$(1, \theta)$ follows immediately from (20). It follows easily from any one of the previous general formulas (27), (35), (36) or (40), that the EPPF of a PD(θ)-partition $\Pi = (\Pi_n)$ is given by the formula

$$p_\theta(n_1, \cdots, n_k) = \frac{\theta^k \Gamma(\theta)}{\Gamma(\theta + n)} \prod_{i=1}^{k} (n_i - 1)! \qquad (n = \sum_{i=1}^{k} n_i). \tag{49}$$

This is a known equivalent [32, 43] of the Ewens sampling formula [18, 17] for the joint distribution of the number of blocks of Π_n of various sizes. It is also known [41, 49] that the following conditions on ρ are equivalent:

(i) ρ is of the form (47), for some $b > 0, \theta > 0$;
(ii) PK$(\rho \mid t)$ =PK(ρ) for all $t > 0$;
(iii) PK(ρ) =PD(θ) for some $\theta > 0$.
(iv) a PK(ρ)-partition has EPPF of the form (49) for some $\theta > 0$.

See also [4, 33] for further properties and applications of PD(θ).

5.2 Generalized gamma

After the one-parameter Poisson-Dirichlet family, the next simplest Lévy density ρ to consider is

$$\rho_{\alpha,c,b}(x) = c x^{-\alpha-1} e^{-bx} \tag{50}$$

for positive constants c and b, and α which is restricted to $0 \leq \alpha < 1$ by the constraints on a Lévy density and (15). The corresponding distributions of T are known as *generalized gamma distributions* [8]. Note that the usual family of gamma distributions is recovered for $\alpha = 0$, and that a stable distribution with index α is obtained for $b = 0$ and $0 < \alpha < 1$. One can also take $\alpha = -\kappa$ for arbitrary $\kappa > 0$, except that in this model the Lévy measure has a total mass $\psi(\infty) < \infty$ so

$$\mathbb{P}(T = 0) = \exp(-\psi(\infty)) > 0,$$

contrary to the present assumption that the distribution of T has a density. Such models can be analyzed by first conditioning on the Poisson total number of lengths, which reduces the model to one with say m i.i.d. lengths with probability density proportional to ρ. In the case (50) for $\alpha = -\kappa$, that is to say that the lengths are i.i.d. gamma(κ, b) variables. This model for random partitions has been extensively studied. It is well known that features of the PD(θ) model can be derived by taking limits of this more elementary model with m i.i.d. gamma(κ, b) lengths as $\kappa \to 0$ and $m \to \infty$ with $\kappa m \to \theta$. See [45] for a review of this circle of ideas and its applications to species sampling models.

The PK$(\rho_{\alpha,c,b})$ model for a random partition defined by $\rho_{\alpha,c,b}$ in (50) for $0 \leq \alpha < 1$ was proposed by McCloskey [37], who first exploited the key idea of size-biased sampling in the setting of species sampling problems. Due to the remarks in Section 4 about scaling and exponential tilting, for $0 < \alpha < 1$ the family of PK$(\rho_{\alpha,c,b}, \gamma)$ distributions, as γ varies over all distributions on $(0, \infty)$, depends only on α and not on c or b. So in studying this family of distributions on $\mathcal{P}^{\downarrow}$ and their associated exchangeable partitions of $\mathbb{N}$, the choice of c and b is entirely a matter of convenience. This study is taken up in the next section, with the choice of $b = 0$ and $c = \alpha/\Gamma(1-\alpha)$ which leads to the simplest form of most results. See also [8, 24, 23] regarding generalized gamma random measures and further developments.

5.3 The stable (α) model

Suppose now that $\mathbb{P}_\alpha$ governs the Poisson model for T with stable (α) distribution with Laplace transform

$$\mathbb{E}_\alpha[\exp(-\lambda T)] = \int_0^\infty e^{-\lambda x} f_\alpha(x)\, dx = \exp(-\lambda^\alpha) \tag{51}$$

for some $0 < \alpha < 1$, where $f_\alpha(x)$ is the stable(α) density of T, that is [52]

$$f_\alpha(t) = \frac{-1}{\pi} \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} \sin(\pi\alpha k) \frac{\Gamma(\alpha k + 1)}{t^{\alpha k + 1}}. \tag{52}$$

For $\alpha = \frac{1}{2}$ this reduces to the following formula of Doetsch [14, pp. 401-402] and Lévy [36]:

$$\mathbb{P}_{\frac{1}{2}}(2T \in dx)/dx = \tfrac{1}{2} f_{\frac{1}{2}}(\tfrac{1}{2}x) = \frac{1}{\sqrt{2\pi}} x^{-\frac{3}{2}} e^{-\frac{1}{2x}}. \tag{53}$$

Special results for $\alpha = \frac{1}{2}$, discussed in Section 8, involve cancellations due to simplification of $f_\alpha(pt)/f_\alpha(t)$ for $0 < p < 1$, which does not appear to be possible for general α. The Lévy density corresponding to the Laplace transform (51) is well known to be

$$\rho_\alpha(x) = \frac{\alpha x^{-\alpha-1}}{\Gamma(1-\alpha)} \quad (x > 0). \tag{54}$$

Write $\mathbb{P}_\alpha(\cdot\,|\,t)$ for $\mathbb{P}_\alpha(\cdot\,|\,T=t)$. So the $\mathbb{P}_\alpha$ distribution of (P_i) on $\mathcal{P}^\downarrow$ is $\mathrm{PK}(\rho_\alpha)$, and the $\mathbb{P}_\alpha(\cdot\,|\,t)$ distribution of (P_i) is $\mathrm{PK}(\rho_\alpha\,|\,t)$. Note from (51) that if T_c is the total length in the model governed by $c\rho_\alpha$ for a constant $c>0$, then T_c has the same distribution as $c^{1/\alpha}T_1$ for $T_1 = T$ as in (51). Together with similar scaling properties of the lengths J_i, this implies that for all $0<\alpha<1$ and $t>0$ there is the formula

$$\mathrm{PK}(c\rho_\alpha\,|\,t) = \mathrm{PK}(\rho_\alpha\,|\,c^{-1/\alpha}t). \tag{55}$$

Formulas for the $\mathrm{PK}(\rho_\alpha\,|\,t)$ distribution are described in Section 5.4. These formulas can be understood as disintegrations of simpler formulas obtained in [43], and recalled in Section 6, for a particular subfamily of the class of $\mathrm{PK}(\rho_\alpha,\gamma)$ distributions.

One reason for special interest in the Kingman family associated with the stable Lévy densities ρ_α is the following result which will be proved elsewhere.

Theorem 8
The EPPF of an exchangeable random partition Π of $\mathbb{N}$ with an infinite number of classes with proper frequencies has an EPPF of the Gibbs form

$$p(n_1,\cdots,n_k) = c_{n,k}\prod_{i=1}^{k} w_{n_i} \text{ where } n=\textstyle\sum_{i=1}^k n_i \tag{56}$$

for some positive weights $w_1 = 1, w_2, w_3, \ldots$ and some $c_{n,k}$ if and only if

$$w_m = \prod_{j=1}^{m-1}(j-\alpha) \quad (m=1,2,\ldots)$$

for some $0\le\alpha<1$. If $\alpha=0$ then the distribution of Π corresponds to $\int_0^\infty \mathrm{PD}(\theta)\gamma(d\theta)$ for some probability distribution γ on $(0,\infty)$, whereas if $0<\alpha<1$ then the distribution of Π corresponds to $\mathrm{PK}(\rho_\alpha,\gamma) := \int_0^\infty \mathrm{PK}(\rho_\alpha\,|\,t)\gamma(dt)$ for some γ.

See also Kerov [28] and Zabell [57] for related characterizations of the two-parameter family discussed in Section 6. This family is characterized by an EPPF of the form (56) with $c_{n,k}$ a product of a function of n and a function of k.

5.4 Conditioning on T

Assume throughout this section that $0<\alpha<1$. Immediately from (20) and (54), in the $\mathrm{PK}(\rho_\alpha\,|\,t)$ model, the distribution of $\tilde{P}_1$ has density

$$\tilde{f}_\alpha(p\,|\,t) = \frac{\alpha(pt)^{-\alpha}}{\Gamma(1-\alpha)}\,\frac{f_\alpha((1-p)t)}{f_\alpha(t)} \qquad (0<p<1). \tag{57}$$

Let h be a non-negative measurable function with $\mathbb{E}_\alpha h(T) = \int_0^\infty h(t)f_\alpha(t)dt = 1$, and let $h\cdot f_\alpha$ denote the distribution on $(0,\infty)$ with density $h(t)f_\alpha(t)$. Then by integration

from (57), under the probability $\mathbb{P}_{\alpha,h}$ governing the $\mathrm{PK}(\rho_\alpha, h \cdot f_\alpha)$ model, the structural distribution of $\tilde{P}_1$ has density

$$\mathbb{P}_{\alpha,h}(\tilde{P}_1 \in dp)/dp = \frac{\alpha}{\Gamma(1-\alpha)} p^{-\alpha}(1-p)^{\alpha-1}\eta_{\alpha,h}(1-p) \quad (0<p<1) \tag{58}$$

where

$$\eta_{\alpha,h}(u) := \int_0^\infty v^{-\alpha}h(v/u)f_\alpha(v)dv = \mathbb{E}_\alpha[T^{-\alpha}h(T/u)]. \tag{59}$$

For instance, it is known [41] that

$$C_{\alpha,\theta} := \mathbb{E}_\alpha(T^{-\theta}) = \frac{\Gamma(\frac{\theta}{\alpha}+1)}{\Gamma(\theta+1)} \qquad (\theta > -\alpha). \tag{60}$$

So for $\theta > -\alpha$, (58) and (59) imply:

$$\text{if } h(t) = C_{\alpha,\theta}^{-1}t^{-\theta} \text{ then } \tilde{P}_1 \text{ has beta}(1-\alpha, \alpha+\theta) \text{ distribution.} \tag{61}$$

This example is discussed further in the next section. As another example, if $h(t) = \exp(b^\alpha - bt)$ for some $b > 0$, then according to (46) the model $\mathrm{PK}(\rho_\alpha, h \cdot f_\alpha)$ is identical to the unconditioned generalized gamma model $\mathrm{PK}(\rho_{\alpha,b})$ with

$$\rho_{\alpha,b}(x) := \rho_\alpha(x)e^{-bx} = \frac{\alpha}{\Gamma(1-\alpha)}\frac{e^{-bx}}{x^{\alpha+1}} \quad (x>0).$$

So the structural density of the $\mathrm{PK}(\rho_{\alpha,b})$ model is given by formula (58) with

$$\eta_{\alpha,h}(u) = \exp(b^\alpha)\mathbb{E}_\alpha[T^{-\alpha}\exp(-bT/u)]. \tag{62}$$

For $\alpha = \frac{1}{2}$ the expectation in (62) can be evaluated by using (53) to write for $\xi > 0$

$$\mathbb{E}_{\frac{1}{2}}[T^{-\frac{1}{2}}\exp(-\xi T)] = \frac{1}{\sqrt{\pi}}\int_0^\infty \frac{dx}{x^2}e^{-(\xi x + 1/x)/2} = 2\sqrt{\frac{\xi}{\pi}}K_1(\sqrt{\xi}) \tag{63}$$

where K_1 is the usual modified Bessel function. Thus for $b > 0$ the $\mathrm{PK}(\rho_{\frac{1}{2},b})$ model associated with the inverse Gaussian distribution [54] has structural distribution with density $\tilde{f}_{\frac{1}{2},b}$ given by the formula

$$\tilde{f}_{\frac{1}{2},b}(p) = \frac{\sqrt{b}e^{\sqrt{b}}}{\pi\sqrt{p}(1-p)}K_1\left(\sqrt{\frac{b}{(1-p)}}\right) \qquad (0<p<1). \tag{64}$$

Proposition 9

For $0 < \alpha < 1, q > 0$ let $\mu_\alpha(q\,|\,t)$ denote the qth moment of the structural density (57) *of the* $\mathrm{PK}(\rho_\alpha\,|\,t)$ *distribution:*

$$\mu_\alpha(q\,|\,t) := \int_0^1 p^q \tilde{f}_\alpha(p\,|\,t)\,dp = \mathbb{E}_\alpha(\tilde{P}_1^q\,|\,t). \tag{65}$$

Then for each $t > 0$ the EPPF of a $\mathrm{PK}(\rho_\alpha \mid t)$ *partition of* $\mathbb{N}$ *is*

$$p_\alpha(n_1, \cdots, n_k \mid t) = \frac{\Gamma(1-\alpha)}{\Gamma(n-k\alpha)} \left(\frac{\alpha}{t^\alpha}\right)^{k-1} \mu_\alpha(n-1-k\alpha+\alpha \mid t) \prod_{i=1}^{k} [1-\alpha]_{n_i-1} \tag{66}$$

where

$$[1-\alpha]_{n_i-1} := \prod_{j=1}^{n_i-1} (j-\alpha) = \frac{\Gamma(n_i-\alpha)}{\Gamma(1-\alpha)}.$$

Alternatively,

$$p_\alpha(n_1, \cdots, n_k \mid t) = \frac{\alpha^k}{t^n} g_\alpha(n-k\alpha \mid t) \prod_{i=1}^{k} [1-\alpha]_{n_i-1} \tag{67}$$

where $g_\alpha(q \mid t) := (\Gamma(q) f_\alpha(t))^{-1} \int_0^t f_\alpha(t-v) v^{q-1} dv$.

Proof. This is read from Theorem 4, since the integral (28) reduces to a standard Dirichlet integral. □

As checks on (66), the symmetry in $(n_1, \ldots, n_k)$ is still evident, and $p_\alpha(n \mid t) = \mu_\alpha(n-1 \mid t)$ as required by (8). However, the addition rules (5) for this EPPF are not at all obvious. Rather, they amount to the following identity involving moments of the structural distribution:

Corollary 10

The moments $\mu_\alpha(q|t)$ *of the structural distribution on* $(0,1)$ *associated with the* $\mathrm{PK}(\rho_\alpha \mid t)$ *distribution on* $\mathcal{P}^\downarrow$ *satisfy the following identity: for all* $1 \le k \le n$ *and* $t > 0$

$$\mu_\alpha(n-1-k\alpha+\alpha \mid t) = \mu_\alpha(n-k\alpha+\alpha \mid t) + \frac{\Gamma(n-k\alpha)\,\alpha t^{-\alpha}}{\Gamma(n+1-k\alpha-\alpha)} \mu_\alpha(n-k\alpha \mid t). \tag{68}$$

To illustrate, according to the simplest addition rule (6),

$$1 = p_\alpha(2 \mid t) + p_\alpha(1,1 \mid t),$$

which amounts to (68) for $n = k = 1$, that is

$$1 = \mu_\alpha(1 \mid t) + \frac{\Gamma(1-\alpha)}{\Gamma(2-2\alpha)} \frac{\alpha}{t^\alpha} \mu_\alpha(1-\alpha \mid t). \tag{69}$$

The addition rule underlying (68) can be checked for general α by an argument described in Section 6. In the case $\alpha = \frac{1}{2}$, the later formulae (99) and (93) show that (68) reduces to a known recursion (106) for the Hermite function.

Repeated application of (68) shows that for each $1 \le k \le n$ the moment on the left side of (66) can be expressed as a linear combination of integer moments $\mu_\alpha(j \mid t)$ for $j = 0, \cdots, n-1$, with coefficients depending on n, k, α, t which could easily be computed recursively. But except in the special case $\alpha = \frac{1}{2}$ discussed in Section 8, even the integer moments seem difficult to evaluate.

6 The two-parameter Poisson-Dirichlet family

For $0 < \alpha < 1, \theta > -\alpha$, let $\gamma_{\alpha,\theta}$ denote the distribution on $(0,\infty)$ with density $C_{\alpha,\theta}^{-1} t^{-\theta}$ at t relative to the stable(α) distribution of T defined by (51), that is

$$\gamma_{\alpha,\theta}(dt) = C_{\alpha,\theta}^{-1} t^{-\theta} f_\alpha(t)\, dt \tag{70}$$

where $C_{\alpha,\theta} := \mathbb{E}_\alpha(T^{-\theta}) = \Gamma(\frac{\theta}{\alpha}+1)/\Gamma(\theta+1)$ as in (60) and (61).

Definition 11
[41, 50] The *Poisson-Dirichlet distribution with two parameters* (α,θ), denoted $\mathrm{PD}(\alpha,\theta)$, is the distribution on $\mathcal{P}^\downarrow$ defined for $0 \le \alpha < 1, \theta > -\alpha$ by

$$\mathrm{PD}(\alpha,\theta) = \begin{cases} \mathrm{PD}(\theta) & \text{for } \alpha = 0, \theta > 0 \\ \mathrm{PK}(\rho_\alpha, \gamma_{\alpha,\theta}) & \text{for } 0 < \alpha < 1, \theta > -\alpha \end{cases} \tag{71}$$

This family of distributions on $\mathcal{P}^\downarrow$ has some remarkable properties and applications. As shown in [41], it follows from Lemma 2 that if (P_i) has $\mathrm{PD}(\alpha,\theta)$distribution then the corresponding size-biased sequence $(\tilde{P}_j)$ can be represented as

$$\tilde{P}_j = W_j \prod_{i=1}^{j-1} (1 - W_i) \tag{72}$$

where the W_j are independent with beta$(1-\alpha, \theta + j\alpha)$ distributions. (73)

So the $\mathrm{PD}(\alpha,\theta)$ distribution can just as well be defined, without reference to the Poisson-Kingman construction, as the distribution of (P_i) defined by ranking $(\tilde{P}_j)$ constructed by (72) from independent W_j as in (72). The sequence $(\tilde{P}_j)$ defined by (72) and (73) for $0 \le \alpha < 1$ and $\theta > 0$ was considered by Engen [15] as a model for species abundances. See [50] for further study of the $\mathrm{PD}(\alpha,\theta)$ family. It was shown in [44] that if (P_i) is a random element of $\mathcal{P}^\downarrow$ with $P_i > 0$ a.s. for all i and the corresponding size-biased sequence $(\tilde{P}_j)$ admits the representation (72) with independent residual fractions W_j, then the W_j must have beta distributions as described in (73), and hence the distribution of (P_i) must be $\mathrm{PD}(\alpha,\theta)$ for some $0 \le \alpha < 1$ and $\theta > -\alpha$. Reformulated in terms of random partitions, and combined with Proposition 7, this yields the following:

Proposition 12
Let Π be the exchangable random partition of $\mathbb{N}$ derived by sampling from a random element (P_i) of $\mathcal{P}^\downarrow$ with $P_i > 0$ for all i. Let Π_k be derived from Π by deletion of the first k classes of Π, with classes in order of appearance, as defined above Proposition 7. Then the following are equivalent
(i) *for each k, Π_k is independent of the frequencies $(\tilde{P}_1, \cdots, \tilde{P}_k)$ of the first k classes of Π;*
(ii) *Π is a $\mathrm{PD}(\alpha,\theta)$-partition for some $0 \le \alpha < 1$ and $\theta > -\alpha$, in which case Π_k is a $\mathrm{PD}(\alpha, \theta + k\alpha)$-partition.*

As shown in [43], the independence property (72) of the residual fractions W_j of a PD(α,θ)-partition allows the corresponding EPPF $p_{\alpha,\theta}(n_1,\ldots,n_k)$ to be evaluated using (3). The result is as follows. For all $0 \le \alpha < 1$ and $\theta > -\alpha$,

$$p_{\alpha,\theta}(n_1,\ldots,n_k) = \frac{[\theta+\alpha]_{k-1;\alpha}}{[\theta+1]_{n-1}} \prod_{i=1}^{k} [1-\alpha]_{n_i-1} \tag{74}$$

where $n = \sum_{i=1}^{k} n_i$ and for real x and a and non-negative integer m

$$[x]_{m;a} = \begin{cases} 1 \text{ for } m = 0 \\ x(x+a)\cdots(x+(m-1)a) \text{ for } m = 1,2,\ldots \end{cases}$$

and $[x]_m = [x]_{m;1}$. The previous formula (49) is the special case of (74) for $\alpha = 0$. Both this case of (74), and the case when $0 < \alpha < 1$ and $\theta = 0$, follow easily from (36). Formula (74) shows that a PD(α,θ) partition Π of $\mathbb{N}$ to be constructed sequentially as follows [43, 45]. Starting from $\Pi_1 = \{\{1\}\}$, given that Π_n has been constructed as a partition of $\mathbb{N}_n$ with say k blocks of sizes $(n_1,\cdots,n_k)$, define Π_{n+1} by assigning the new element $n+1$ to the jth class whose current size is n_j with probability

$$\mathbb{P}(j\uparrow | n_1,\cdots,n_k) = \frac{n_j - \alpha}{n+\theta} \tag{75}$$

for $1 \le j \le k$, and assigning $n+1$ to a new class numbered $k+1$ with the remaining probability

$$\mathbb{P}(k+1\uparrow | n_1,\cdots,n_k) = \frac{k\alpha}{n+\theta} \tag{76}$$

For $\alpha = 0$ and $\theta > 0$ this is generalization of Polya's urn scheme developed by Blackwell-McQueen [7] and Hoppe [21]. See [43, 45, 20] for consideration of more general prediction rules for exchangeable random partitions.

The following calculation shows how to derive either of the two EPPF's (74) and (66) from the other. The argument also shows that the function $p_\alpha(n_1,\cdots,n_k\,|\,t)$ defined by (66) satisfies the addition rules of an EPPF as a consequence of the corresponding addition rules for $p_{\alpha,\theta}(n_1,\ldots,n_k)$, which are much more obvious.

The kernel $\gamma_{\alpha,\theta}(dt)$ introduced in (70), is now viewed for a fixed α as a family of probability distributions on $(0,\infty)$ indexed by $\theta \in (-\alpha,\infty)$, that is a Markov kernel γ_α from $(-\alpha,\infty)$ to $(0,\infty)$. For a non-negative measurable function $h = h(t)$ with domain $(0,\infty)$, define a function $\gamma_\alpha h = (\gamma_\alpha h)(\theta)$ with domain $(-\alpha,\infty)$ by the usual action of this Markov kernel as an integral operator:

$$(\gamma_\alpha h)(\theta) = \int_0^\infty \gamma_{\alpha,\theta}(dt) h(t) \tag{77}$$

Then say $(\gamma_\alpha h)(\theta)$ is the γ_α-*transform* of $h(t)$. Let $\mathbb{E}_{\alpha,\theta}$ denote expectation with respect to the probability distribution

$$\mathbb{P}_{\alpha,\theta}(\cdot) := \int_0^\infty \mathbb{P}_\alpha(\cdot\,|\,t)\gamma_{\alpha,\theta}(dt).$$

By definition, for each non-negative random variable X governed by the family of conditional laws $(\mathbb{P}_\alpha(\cdot\,|\,t), t > 0)$,

$$\text{the } \gamma_\alpha\text{-transform of } \mathbb{E}_\alpha(X\,|\,t) \text{ is } \mathbb{E}_{\alpha,\theta}(X). \tag{78}$$

In particular, for each $(n_1,\ldots,n_k)$,

$$\text{the } \gamma_\alpha\text{-transform of } p_\alpha(n_1,\cdots,n_k\,|\,t) \text{ is } p_{\alpha,\theta}(n_1,\ldots,n_k). \tag{79}$$

An obvious change of variable allows uniqueness and inversion results for the γ_α-transform to be deduced from standard results for Mellin or bilateral exponential transforms. So the problem is just to show that the γ_α-transform of the right side of (66) is the right side of (74). To see this, observe first that for each $q > 0$, because $\mu_\alpha(q|t) := \mathbb{E}_\alpha(\tilde{P}_1^q\,|\,t)$,

$$\text{the } \gamma_\alpha\text{-transform of } \mu_\alpha(q|t) \text{ is } \mathbb{E}_{\alpha,\theta}(\tilde{P}_1^q) = \frac{\Gamma(1-\alpha+q)\Gamma(1+\theta)}{\Gamma(1+\theta+q)\Gamma(1-\alpha)} \tag{80}$$

where $\mathbb{E}_{\alpha,\theta}(\tilde{P}_1^q)$ is evaluated using (61). To deal with the factor of $t^{-(k-1)\alpha}$ in (66), note from (60) that for each $\beta > 0$, and any $h(t)$,

$$\text{the } \gamma_\alpha\text{-transform of } t^{-\beta}h(t) \text{ is } \frac{\Gamma(\frac{\theta}{\alpha}+\frac{\beta}{\alpha}+1)\Gamma(\theta+1)}{\Gamma(\frac{\theta}{\alpha}+1)\Gamma(\theta+\beta+1)}\,(\gamma_\alpha h)(\theta+\beta). \tag{81}$$

By (80) for $q = n-1-k\alpha+\alpha$ and (81) for $\beta = \alpha k-\alpha$ and $h(t) = \mu_\alpha(q\,|\,t)$ the right side of (66) has for its γ_α-transform the following function of θ:

$$\frac{\alpha^{k-1}\Gamma(1-\alpha)}{\Gamma(n-k\alpha)}\,\frac{\Gamma(\frac{\theta}{\alpha}+k)\Gamma(\theta+1)}{\Gamma(\frac{\theta}{\alpha}+1)\Gamma(\theta+k\alpha-\alpha+1)}\,\frac{\Gamma(n-k\alpha)\Gamma(1+\theta+k\alpha-\alpha)}{\Gamma(n+\theta)\Gamma(1-\alpha)}\prod_{i=1}^{k}[1-\alpha]_{n_i-1}$$

which reduces by cancellation to the right side of (74).

6.1 The α-diversity

Let Π be an exchangeable random partition of $\mathbb{N}$ with ranked frequencies (P_i). Let K_n denote the number of classes of Π_n, the partition of $\mathbb{N}_n$ induced by Π. Say that Π has *α-diversity* S and write $\alpha\text{-DIVERSITY}(\Pi) = S$ iff there exists a random variable S with $0 < S < \infty$ a.s. and

$$K_n \sim Sn^\alpha \text{ as } n \to \infty \tag{82}$$

where for two sequences of random variables A_n and B_n, the notation $A_n \sim B_n$ will now be used to indicate that $A_n/B_n \to 1$ almost surely as $n \to \infty$. According to a result of Karlin [27], applied conditionally given (P_i), if these ranked frequencies are such that

$$P_i \sim \left(\frac{S}{\Gamma(1-\alpha)i}\right)^{\frac{1}{\alpha}} \tag{83}$$

for some $0 < S < \infty$ then Π has α-diversity S.

Proposition 13
Suppose Π *is a* PK(ρ_α, γ) *partition of* $\mathbb{N}$ *for some* $0 < \alpha < 1$ *and some probability distribution* γ *on* $(0, \infty)$. *Then*

(i) α-DIVERSITY$(\Pi) = S$ for a random variable S with $S = T^{-\alpha}$ where $T = S^{-1/\alpha}$ has distribution γ. In particular, $S = t^{-\alpha}$ is constant if Π is a PK$(\rho_\alpha \,|\, t)$ partition.

(ii) A regular conditional distribution for Π given $S = s$ is defined by the EPPF $p_\alpha(n_1, \cdots, n_k | s^{-1/\alpha})$ obtained by setting $t = s^{-1/\alpha}$ in (66).

(iii) In particular, both (i) and (ii) hold if Π is a PD(α, θ) partition for some $\theta > -\alpha$. Then the α-diversity S of Π is $S = T^{-\alpha}$ for T with the distribution $\gamma_{\alpha,\theta}$ defined by (70).

Proof. Suppose that (P_i) has PK(ρ_α, γ) distribution. The fact that (83) holds for $S = T^{-\alpha}$ in the unconditioned case where T has stable(α) distribution is due to Kingman [29]. Kingman's argument, using the law of large numbers for small jumps of the Poisson process, applies just as well for T conditioned to be a constant t. So (83) follows in general by mixing over t. □

See [50] and papers cited there for further information about the Mittag-Leffler distribution of $S = T^{-\alpha}$ derived from a PD$(\alpha, 0)$ partition. The corresponding distribution of S for PD(α, θ) has density at s proportional to $s^{\theta/\alpha}$ relative to this Mittag-Leffler distribution.

As shown in [50, Proposition 10], if Π is a partition of $\mathbb{N}$ whose ranked frequencies (P_i) have the PD$(\alpha, 0)$ distribution, then $S = \alpha$-DIVERSITY(Π) can be recovered from Π or (P_i) via either (81) or (83). Then $T = S^{-1/\alpha}$ has stable(α) distribution as in (51), and (TP_i) is then sequence of points of a Poisson process with Lévy density ρ_α. See also [47, 48] for more about the distribution of K_n derived from a PD(α, θ) partition.

7 Application to lengths of excursions

This section reviews some results of [41, 49, 46, 50]. Let $\mathbb{P}^0_\alpha$ govern a strong Markov process B starting at a recurrent point 0 of its statespace, such that the inverse $(\tau_\ell, \ell \geq 0)$ of the local time process $(L_t, t \geq 0)$ of B at zero is a stable subordinator of index α for some $0 < \alpha < 1$. That is to say, $\mathbb{E}^0_\alpha \exp(-\lambda \tau_1) = \exp(-c\lambda^\alpha)$ for some constant $c > 0$. So the $\mathbb{P}^0_\alpha$ distribution of τ_1 is the $\mathbb{P}_\alpha$ distribution of $c^{1/\alpha} T$ for T as in (51). For example, B could be a one-dimensional Brownian motion ($\alpha = \frac{1}{2}$) or Bessel process of dimension $2 - 2\alpha$. In the Brownian case, take $c = \sqrt{2}$ to obtain the usual normalization of local time as occupation density relative to Lebesgue measure, which makes $L_1 \stackrel{d}{=} |B_1|$. Let $M = \{t : 0 \leq t \leq 1, B_t = 0\}$ denote the random closed subset of $[0, 1]$ defined by the zero set of B. Component intervals of the complement of M relative to $[0, 1]$ are called *excursion intervals*. For $0 \leq t \leq 1$ let $G_t = \sup\{M \cap [0, t]\}$, the last zero of B before time t. Note that with probability one, $G_1 < 1$, so one of the excursion intervals is the *meander interval* $(G_1, 1]$, whose length $1 - G_1$ is one of the lengths appearing in the list

(P_i) say of ranked lengths of excursion intervals. According to the main result of [49],

$$\text{the sequence } (P_i) \text{ of ranked lengths has } \text{PD}(\alpha,0) \text{ distribution} \tag{84}$$

Let $U_1, U_2, \cdots$ be a sequence of i.i.d. uniform $[0,1]$ random variables, independent of B, called the sequence of *sample points*. Let $\Pi = (\Pi_n)$ be the random partition of $\mathbb{N}$ generated by the random equivalence relation $i \sim j$ iff $G_{U_i} = G_{U_j}$. That is to say $i \sim j$ iff U_i and U_j fall in the same excursion interval. So for example $\Pi_5 = \{\{1,2,5\},\{3\},\{4\}\}$ iff U_1, U_2 and U_5 fall in one excursion interval, U_3 in another, and U_4 in a third. By translation of results of [49, 50] into present notation

$$\Pi \text{ is a } \text{PD}(\alpha,0) \text{ partition and } \alpha\text{-DIVERSITY}(\Pi) = cL_1 \tag{85}$$

where L_1 is the local time of B at zero up to time 1. By construction, the sequence $(\tilde{P}_j)$ of class frequencies of Π is the sequence of lengths of excursion intervals in the order they are discovered by the sample points, and (P_i) is recovered from $(\tilde{P}_j)$ by ranking. To illustrate formula (74), U_1 and U_2 fall in different excursion intervals with probability $p_{\alpha,0}(1,1) = \alpha$, and in the same one with probability $p_{\alpha,0}(2) = 1-\alpha$. Similarly, given that the local time is $L_1 = \ell$, two sample points fall in the same excursion interval with probability $p_\alpha(2\,|\,(c\ell)^{-1/\alpha})$, and in different excursion intervals with probability $p_\alpha(1,1\,|\,(c\ell)^{-1/\alpha})$, for $p_\alpha(\cdots|t)$ defined by (66). See Section 8 for evaluation of these functions in the case $\alpha = \frac{1}{2}$ corresponding to a Brownian motion B.

Let $R_n = 1 - \tilde{P}_1 - \cdots - \tilde{P}_n$, which is the total length of excursions which remain undiscovered after the sampling process has found n distinct excursion intervals. The result of Proposition 12 in this setting, due to [41], is that for each $n = 0,1,2,\cdots$ a $\text{PD}(\alpha, n\alpha)$ distributed sequence is obtained by ranking the sequence

$$\frac{1}{R_n}(\tilde{P}_{n+1}, \tilde{P}_{n+2}, \cdots) \tag{86}$$

of relative excursion lengths which remain after discovery of the first n intervals. For $n = 1$ the same $\text{PD}(\alpha,\alpha)$ distribution is obtained more simply by deleting the meander of length $1 - G_1$, renormalizing and reranking. This is due to the result of [49] that the length $1 - G_1$ of the meander interval is a size-biased choice from (P_i). As the excursion lengths in this case are just the excursion lengths of a standard bridge, equivalent to conditioning on $B_1 = 0$, the ranked excursion lengths of such a bridge have $\text{PD}(\alpha,\alpha)$ distribution. As first shown in [49], this implies that both the unconditioned process B and the bridge B given $B_1 = 0$ share a common conditional distribution for the ranked excursion lengths (P_i) given the local time L_1. In present notation, this conditional distribution of (P_i) given $L_1 = \ell$, with or without conditioning on $B_1 = 0$, is $\text{PK}(\rho_\alpha|(c\ell)^{-1/\alpha})$.

One final identity is worth noting. As a consequence of the above discussion, for the process B, the conditional distribution of the meander length $1 - G_1$ given $L_1 = \ell$ is given by

$$\mathbb{P}^0_\alpha(1 - G_1 \in dp | L_1 = \ell) = \mathbb{P}^0_\alpha(\tilde{P}_1 \in dp | L_1 = \ell) = \tilde{f}_\alpha(p|(c\ell)^{-1/\alpha})dp \tag{87}$$

where $\tilde{f}_\alpha(p|t)$ as in (57) is the structural density of the Poisson model for stable (α) distributed T conditioned on $T = t$. So the moment function $\mu_\alpha(q|t)$ appearing in the EPPF (66) of this model can be interpreted in the present setting as

$$\mu_\alpha(q|t) = \mathbb{E}^0_\alpha[(1-G_1)^q \,|\, L_1 = c^{-1}t^{-\alpha}]. \tag{88}$$

8 The Brownian excursion partition

In this section let Π be the *Brownian excursion partition*, that is the random partition of $\mathbb{N}$ generated by uniform random sampling of points from the interval $[0,1]$ partitioned by the excursion intervals of a standard Brownian motion B. According to the result of [49] recalled in (84),

$$\Pi \text{ is a } \mathrm{PK}(\rho_{\frac{1}{2}}) = \mathrm{PD}(\tfrac{1}{2}, 0) \text{ partition.} \tag{89}$$

With conditioning on $B_1 = 0$, the process B becomes a standard Brownian bridge. So Π given $B_1 = 0$ is a $\mathrm{PD}(\frac{1}{2}, \frac{1}{2})$ partition, as discussed in the previous subsection. Features of the distribution of Π and the conditional distribution of Π given $B_1 = 0$ were described in [46]. This section presents refinements of these results obtained by conditioning on L_1, the local time of B at 0 up to time 1, with the usual normalization of Brownian local time as occupation density relative to Lebesgue measure. Unconditionally, L_1 has the same distribution as $|B_1|$, that is

$$\mathbb{P}(L_1 \in d\lambda) = \mathbb{P}(|B_1| \in d\lambda) = 2\varphi(\lambda)d\lambda \qquad (\lambda > 0)$$

where $\varphi(z) := (1/\sqrt{2\pi})\exp(-\frac{1}{2}z^2)$ is the standard Gaussian density of B_1. Whereas the conditional distribution of L_1 given $B_1 = 0$ is the Rayleigh distribution

$$\mathbb{P}(L_1 \in d\lambda \,|\, B_1 = 0) = \sqrt{2\pi}\lambda\varphi(\lambda)d\lambda \qquad (\lambda > 0).$$

Note from (85) that the $\frac{1}{2}$-diversity of Π is the random variable $\sqrt{2}L_1$. So the number K_n of blocks of Π grows almost surely like $\sqrt{2n}L_1$ as $n \to \infty$. For $\lambda \geq 0$ let $\Pi(\lambda)$ denote a random partition with

$$\Pi(\lambda) \stackrel{d}{=} (\Pi \,|\, L_1 = \lambda) \stackrel{d}{=} (\Pi \,|\, L_1 = \lambda, B_1 = 0) \tag{90}$$

where $\stackrel{d}{=}$ denotes equality in distribution. So according to the previous discussion,

$$\Pi(\lambda) \text{ is a } \mathrm{PK}(\rho_{\frac{1}{2}} \,|\, \tfrac{1}{2}\lambda^{-2}) \text{ partition} \tag{91}$$

whose $\frac{1}{2}$-diversity is $\sqrt{2}\lambda$. Let $\mathrm{PD}(\frac{1}{2}\,||\,\lambda)$ denote the probability distribution on $\mathcal{P}^\downarrow$ associated with $\Pi(\lambda)$, that is the common distribution of ranked lengths of excursions

of a Brownian motion or Brownian bridge over $[0,1]$ given $L_1 = \lambda$. Then by Definition 11 and (53), for $\theta > -\frac{1}{2}$ there is the identity of probability laws on $\mathcal{P}^{\downarrow}$

$$\mathrm{PD}(\tfrac{1}{2},\theta) = \frac{2}{\mathbb{E}[|B_1|^{2\theta}]} \int_0^\infty \mathrm{PD}(\tfrac{1}{2}\,||\,\lambda)\lambda^{2\theta}\varphi(\lambda)d\lambda \tag{92}$$

where, according to the gamma$(\frac{1}{2})$ distribution of $\frac{1}{2}B_1^2$ and the duplication formula for the gamma function,

$$\mathbb{E}(|B_1|^{2\theta}) = 2^{\theta}\frac{\Gamma(\theta+\frac{1}{2})}{\Gamma(\frac{1}{2})} = 2^{-\theta}\frac{\Gamma(2\theta+1)}{\Gamma(\theta+1)} \qquad (\theta > -\tfrac{1}{2}). \tag{93}$$

It was shown in [3] (see also [5, 48]) that it is possible to construct the Brownian excursion partitions as a partition valued *fragmentation process* $(\Pi(\lambda), \lambda \geq 0)$, meaning that $\Pi(\lambda)$ is constructed for each λ on the same probability space, in such a way that $\Pi(\lambda)$ is a coarser partition than $\Pi(\mu)$ whenever $\lambda < \mu$. The question of whether a similar construction is possible for index α instead of index $\frac{1}{2}$ remains open. A natural guess is that such a construction might be made with one of the self-similar fragmentation processes of Bertoin [6], but Miermont and Schweinsberg [38] have recently shown that a construction of this form is possible only for $\alpha = \frac{1}{2}$.

8.1 Length biased sampling

Let $\tilde{P}_j(\lambda)$ denote the frequency of the jth class of $\Pi(\lambda)$. So $(\tilde{P}_j(\lambda), j = 1,2\ldots)$ is distributed like the lengths of excursions of B over $[0,1]$ given $L_1 = \lambda$, as discovered by a process of length-biased sampling. In view of Lévy's formula (53) for the stable$(\frac{1}{2})$ density, the formula (57) reduces for $\alpha = \frac{1}{2}$ to the following more explicit formula for the structural density of $\Pi(\lambda)$:

$$\mathbb{P}(\tilde{P}_1(\lambda) \in dp) = \frac{\lambda}{\sqrt{2\pi}} p^{-\frac{1}{2}}(1-p)^{-\frac{3}{2}} \exp\left(-\frac{\lambda^2}{2}\frac{p}{(1-p)}\right) dp \quad (0 < p < 1) \tag{94}$$

or equivalently

$$\mathbb{P}(\tilde{P}_1 \leq y) = 2\Phi\left(\lambda\sqrt{\frac{y}{1-y}}\right) - 1 \quad (0 \leq y < 1) \tag{95}$$

where $\Phi(z) := \mathbb{P}(B_1 \leq z)$ is the standard Gaussian distribution function. Put another way, there is the equality in distribution

$$\tilde{P}_1(\lambda) \stackrel{d}{=} \frac{B_1^2}{\lambda^2 + B_1^2}. \tag{96}$$

Furthermore, by a similar analysis using Lemma 1, there is the following result which shows how to construct the whole sequence $(\tilde{P}_j(\lambda), j \geq 1)$ for any $\lambda > 0$ from a single sequence of independent standard Gaussian variables. Then $\Pi(\lambda)$ can be constructed by sampling from $(\tilde{P}_j(\lambda), j \geq 1)$ as discussed in Section 2.

Proposition 14
[3, Corollary 5] *Fix $\lambda > 0$. A sequence $(\tilde{P}_j(\lambda), j \geq 1)$ is distributed like a length-biased random permutation of the lengths of excursions of a Brownian motion or standard Brownian bridge over $[0,1]$ conditioned on $L_1 = \lambda$ if and only if*

$$\tilde{P}_j(\lambda) = \frac{\lambda^2}{\lambda^2 + S_{j-1}} - \frac{\lambda^2}{\lambda^2 + S_j} \tag{97}$$

where $S_j := \sum_{i=1}^{j} X_i$ for X_i which are independent and identically distributed like B_1^2 for a standard Gaussian variable B_1.

Let $\mu(q||\lambda)$ denote the qth moment of the distribution of $\tilde{P}_1(\lambda)$. So in the notation of (65) and (68)

$$\mu(q||\lambda) := \mathbb{E}[(\tilde{P}_1(\lambda))^q] = \mu_{\frac{1}{2}}(q|\tfrac{1}{2}\lambda^{-2}). \tag{98}$$

Lemma 15
For each $\lambda > 0$

$$\mu(q||\lambda) = \mathbb{E}\left[\left(\frac{B_1^2}{\lambda^2 + B_1^2}\right)^q\right] = \mathbb{E}(|B_1|^{2q})\, h_{-2q}(\lambda) \quad (q > -\tfrac{1}{2}) \tag{99}$$

where $\mathbb{E}(|B_1|^{2q})$ is given by (93) and h_{-2q} is the Hermite function of index $-2q$, that is $h_0(\lambda) = 1$ and for $q \notin \{0, 1, 2 \ldots\}$

$$h_{-2q}(\lambda) := \frac{1}{2\Gamma(2q)} \sum_{j=0}^{\infty} \Gamma(q + j/2) 2^{q+j/2} \frac{(-\lambda)^j}{j!}. \tag{100}$$

Also,

$$\mu(q||\lambda) = \mathbb{E}[\exp(-\lambda\sqrt{2\Gamma_q}] \quad (q > 0) \tag{101}$$

where Γ_q denotes a Gamma random variable with parameter q:

$$\mathbb{P}(\Gamma_q \in dt) = \Gamma(q)^{-1} t^{q-1} e^{-t} dt \quad (t > 0).$$

Proof. The first equality in (99) is read from (96). The second equality in (99) is the integral representation of the Hermite function provided by Lebedev [35, Problem 10.8.1], and (100) is read from [35, (10.4.3)]. According to another well known integral representation of the Hermite function [35, (10.5.2)], [16, 8.3 (3)], for $q > 0$

$$h_{-2q}(x) = \frac{1}{\Gamma(2q)} \int_0^\infty t^{2q-1} e^{-\frac{1}{2}t^2 - xt} dt = \frac{2^{q-1}}{\Gamma(2q)} \int_0^\infty v^{q-1} e^{-v - x\sqrt{2v}} dv. \tag{102}$$

Formula (101) follows easily from this and (99). □

The identity

$$\mathbb{E}\left[\left(\frac{B_1^2}{\lambda^2 + B_1^2}\right)^q\right] = \mathbb{E}[\exp(-\lambda\sqrt{2\Gamma_q}] \quad (q > 0), \tag{103}$$

which is implied by the previous proposition, can also be checked by the following argument suggested by Marc Yor. Let X be a positive random variable independent of Γ_q, and let ε with $\varepsilon \stackrel{d}{=} \Gamma_1$ be a standard exponential variable independent of both X and Γ_q. Then by elementary conditioning arguments, for $\theta \geq 0$

$$\mathbb{E}\left[\left(\frac{X}{\theta+X}\right)^q\right] = \mathbb{E}\left[e^{-\theta\Gamma_q/X}\right] = \mathbb{P}(\varepsilon X/\Gamma_q > \theta). \tag{104}$$

Take $X = B_1^2$ and $\theta = \lambda^2$, and use the identity $\varepsilon B_1^2 \stackrel{d}{=} \varepsilon^2/2$, which is a well known probabilistic expression of the gamma duplication formula, to deduce (103) from (104).

The following display identifies $h_\nu(z)$ in the notation of various authors:

$$\begin{aligned} h_\nu(z) &= 2^{-\nu/2}H_\nu(z/\sqrt{2}) = 2^{\nu/2}\Psi(-\nu/2, 1/2, z^2/2) && \text{(Lebedev[35])} \\ &= 2^{\nu/2}U(-\nu/2, 1/2, z^2/2) && \text{(Abramowitz and Stegun) [1]} \\ &= e^{\frac{1}{4}z^2}U(-\nu-\tfrac{1}{2}, z) && \text{(Miller[39])} \\ &= e^{\frac{1}{4}z^2}D_\nu(z) && \text{(Erdelyi [16], Toscano [56])} \end{aligned}$$

The functions $U(a,z)$ and $D_\nu(z)$ are known as *parabolic cylinder functions*, *Weber functions* or *Whittaker functions*. The function $U(a,b,z)$, which is available in *Mathematica* as `HypergeometricU[a,b,z]`, is a *confluent hypergeometric function of the second kind*. Note that $h_n(z)$ defined for $n = 0,1,2,\ldots$ by continuous extension of (100) is the sequence of Hermite polynomials orthogonal with respect to the standard Gaussian density $\varphi(x)$. Also, the function $h_{-1}(x)$ for real x is identified as *Mill's ratio* [26, 33.7]:

$$h_{-1}(x) = \frac{\mathbb{P}(B_1 > x)}{\varphi(x)} = e^{\frac{1}{2}x^2}\int_x^\infty e^{-\frac{1}{2}z^2}dz. \tag{105}$$

For all complex ν and z, the Hermite function satisfies the recursion

$$h_{\nu+1}(z) = zh_\nu(z) - \nu h_{\nu-1}(z), \tag{106}$$

which combined with (105) and $h_0(x) = 1$ yields

$$h_{-2}(x) = 1 - xh_{-1}(x) \tag{107}$$

$$2!h_{-3}(x) = -x + (1+x^2)h_{-1}(x) \tag{108}$$

$$3!h_{-4}(x) = 2 + x^2 - (3x+x^3)h_{-1}(x) \tag{109}$$

and so on. See [51] for further interpretations of the Hermite function in terms of Brownian motion and related stochastic processes.

8.2 Partition probabilities

Recall the notation

$$[\tfrac{1}{2}]_n := \prod_{j=1}^{n}(j-\tfrac{1}{2}) = \frac{\Gamma(\frac{1}{2}+n)}{\Gamma(\frac{1}{2})} = \frac{(2n)!}{2^{2n}n!}.$$

Corollary 16
The distribution of $\Pi(\lambda)$, a Brownian excursion partition conditioned on $L_1 = \lambda$, is determined by the following EPPF: for $n_1,\ldots,n_k$ with $\sum_{i=1}^{k} n_i = n$

$$p_{\frac{1}{2}}(n_1,\ldots,n_k||\lambda) = 2^{n-k}\lambda^{k-1}h_{k+1-2n}(\lambda)\prod_{i=1}^{k}[\tfrac{1}{2}]_{n_i-1}. \tag{110}$$

Proof. This is read from (66), (99) and (93). □

Formula (110) combined with (14) gives an expression in terms of the Hermite function for the positive integer moments of the sum $S_m(\lambda)$ of mth powers of lengths of excursions of Brownian motion on $[0,1]$ given $L_1 = \lambda$. This formula for $m = 2$ was derived in another way by Janson [25, Theorem 7.4]. There the distribution of $S_2(\lambda)$ appears as the asymptotic distribution, in a suitable limit regime, of the cost of linear probing hashing.

According to (91) and Definition 11, for each $\theta > -\frac{1}{2}$, the EPPF (110) describes the conditional distribution of a $\mathrm{PD}(\frac{1}{2},\theta)$ partition (Π_n) given $\lim_n K_n/\sqrt{2n} = \lambda$, where K_n is the number of blocks of Π_n. Easily from (110), for each fixed $\lambda > 0$, a sequential description of $(\Pi_n(\lambda), n = 1,2,\ldots)$ is obtained by replacing the prediction rules (75) and (76) by

$$\mathbb{P}(j\uparrow | n_1,\cdots,n_k) = (2n_j - 1)\frac{h_{k-1-2n}(\lambda)}{h_{k+1-2n}(\lambda)} \quad (1 \le j \le k) \tag{111}$$

$$\mathbb{P}(k+1\uparrow | n_1,\cdots,n_k) = \frac{\lambda h_{k-2n}(\lambda)}{h_{k+1-2n}(\lambda)}. \tag{112}$$

The addition rule for the EPPF (110) is equivalent to the fact that these transition probabilities sum to 1. As a check, this is implied the recurrence formula (106) for the Hermite function.

Corollary 17
Let $K_n(\lambda)$ be the number of blocks of $\Pi_n(\lambda)$, where $(\Pi_n(\lambda), n = 1,2,\ldots)$ is the Brownian excursion partition conditioned on $L_1 = \lambda$. Then $(K_n(\lambda), n = 1,2,\ldots)$ is a Markov chain with the following inhomogeneous transition probabilities: for $1 \le k \le n$

$$\mathbb{P}(K_{n+1}(\lambda) = k | K_n(\lambda) = k) = (2n-k)\frac{h_{k-1-2n}(\lambda)}{h_{k+1-2n}(\lambda)} \tag{113}$$

$$\mathbb{P}(K_{n+1}(\lambda) = k+1 \,|\, K_n(\lambda) = k) = \frac{\lambda h_{k-2n}(\lambda)}{h_{k+1-2n}(\lambda)}. \tag{114}$$

Moreover, the distribution of $K_n(\lambda)$ *is given by the formula*

$$\mathbb{P}(K_n(\lambda) = k) = \frac{(2n-k-1)!\,\lambda^{k-1} h_{k+1-2n}(\lambda)}{(n-k)!(k-1)!2^{n-k}} \quad (1 \le k \le n). \tag{115}$$

Proof. The Markov property of $(K_n(\lambda), n = 1,2,\ldots)$ and the transition probabilities (113)–(114) follow easily from (111)–(112). Then (115) follows by induction on n, using the forwards equations implied by the transition probabilities. □

Let K_n denote the number of blocks of Π_n, where (Π_n) is the unconditioned Brownian excursion partition. Then, from the discussion around (90),

$$(K_n(\lambda), n \ge 1) \stackrel{d}{=} (K_n, n \ge 1 \,|\, \lim_n K_n/\sqrt{2n} = \lambda). \tag{116}$$

According to (89), (75) and (76), the sequence $(K_n, n \ge 1)$ is an inhomogeneous Markov chain with transition probabilities

$$\mathbb{P}(K_{n+1} = k \,|\, K_n = k) = \frac{2n-k}{2n} \tag{117}$$

$$\mathbb{P}(K_{n+1} = k+1 \,|\, K_n = k) = \frac{k}{2n} \tag{118}$$

which imply that the unconditional distribution of K_n is given by the formula [46, Corollary 3]

$$\mathbb{P}(K_n = k) = \binom{2n-k-1}{n-1} 2^{k+1-2n} \quad (1 \le k \le n). \tag{119}$$

Due to (116), for each $\lambda > 0$ the inhomogeneous Markov chain $(K_n(\lambda), n \ge 1)$ has the same co-transition probabilities as $(K_n, n \ge 1)$. From (117), (118) and (119), the co-transition probabilities of $(K_n, n \ge 1)$ are

$$\mathbb{P}(K_n = k \,|\, K_{n+1} = k) = \frac{2(n-k+1)}{2n-k+1} \tag{120}$$

$$\mathbb{P}(K_n = k-1 \,|\, K_{n+1} = k) = \frac{k-1}{2n-k+1}. \tag{121}$$

As a check, the fact that $(K_n(\lambda), n \ge 1)$ has the same co-transition probabilities can be read from (113), (114) and (115). It can be shown that the Markov chains $(K_n(\lambda), n \ge 1)$ for $\lambda \in [0,\infty]$, with definition by weak continuity for $\lambda = 0$ or ∞, are the extreme points of the convex set of all laws of Markov chains with these co-transition probabilities. A generalization of this fact, to $\alpha \in (0,1)$ instead of $\alpha = \frac{1}{2}$, and similar considerations for $\alpha = 0$, yield the second sentence of Theorem 8.

To illustrate the formulas above, according to (9) and (99), or (110) for $n = 2$, given $L_1 = \lambda$, two independent uniform $[0,1]$ variables fall in the same excursion interval of the Brownian motion with probability

$$p_{\frac{1}{2}}(2\,||\,\lambda) = \mu(1\,||\,\lambda) = h_{-2}(\lambda) = 1 - \lambda h_{-1}(\lambda) \tag{122}$$

and in different excursion intervals with probability $\lambda h_{-1}(\lambda)$. According to (110) for $n = 3$, given $L_1 = \lambda$, three independent uniform random points U_1, U_2, U_3 with uniform distribution on $[0,1]$ fall in the same excursion interval of a Brownian motion or Brownian bridge with probability

$$\mathbb{P}(K_3(\lambda) = 1) = p_{\frac{1}{2}}(3\,||\,\lambda) = 3h_{-4}(\lambda) = 1 + \tfrac{1}{2}\lambda^2 - (\tfrac{3}{2}\lambda + \tfrac{1}{2}\lambda^3)h_{-1}(\lambda) \tag{123}$$

while U_1 and U_2 fall in one excursion interval and U_3 in another with probability

$$\tfrac{1}{3}\,\mathbb{P}(K_3(\lambda) = 2) = p_{\frac{1}{2}}(2,1\,||\,\lambda) = \lambda h_{-3}(\lambda) = -\tfrac{1}{2}\lambda^2 + (\tfrac{1}{2}\lambda + \tfrac{1}{2}\lambda^3)h_{-1}(\lambda) \tag{124}$$

and the three points fall in three different excursion intervals with probability

$$\mathbb{P}(K_3(\lambda) = 3) = p_{\frac{1}{2}}(1,1,1\,||\,\lambda) = \lambda^2 h_{-2}(\lambda) = \lambda^2 - \lambda^3 h_{-1}(\lambda). \tag{125}$$

As a check, the sum of expressions for $\mathbb{P}(K_3(\lambda) = k)$ over $k = 1,2,3$ reduces to 1. Since

$$\mathbb{P}(K_n(\lambda) = k) = \sum_{n_1 \geq \cdots \geq n_k} \#(n_1, \cdots, n_k) p_{\frac{1}{2}}(n_1, \cdots, n_k\,||\,\lambda) \tag{126}$$

where the sum is over all decreasing sequences of positive integers $(n_1, \cdots, n_k)$ with sum n, and $\#(n_1, \cdots, n_k)$ is the number of distinct partitions of $\mathbb{N}_n$ into k subsets of sizes $(n_1, \cdots, n_k)$, formula (115) amounts to

$$\sum_{n_1 \geq \cdots \geq n_k} \#(n_1, \cdots, n_k) \prod_{i=1}^{k} [\tfrac{1}{2}]_{n_i - 1} = \binom{2n-k-1}{n-1} \frac{\Gamma(n)}{\Gamma(k)} 2^{2k-2n} \tag{127}$$

which can be checked as follows. According to (74) and (89), the unconditional EPPF of the Brownian excursion partition Π is

$$p_{\frac{1}{2},0}(n_1, \cdots, n_k) = \frac{\Gamma(k)}{2^{k-1}\Gamma(n)} \prod_{i=1}^{k} [\tfrac{1}{2}]_{n_i - 1} \tag{128}$$

so (127) can be deduced from (128), (119), and the unconditioned form of (126).

8.3 Some identities

As a consequence of (92) and (99), for all $q > -\frac{1}{2}$ and $\theta > -\frac{1}{2}$ there is the identity

$$\frac{2}{\mathbb{E}(|B_1|^{2\theta})} \int_0^\infty \lambda^{2\theta} \mu(q\,||\,\lambda)\phi(\lambda)d\lambda = \frac{\Gamma(\theta+1)\Gamma(q+\frac{1}{2})}{\Gamma(\frac{1}{2})\Gamma(q+\theta+1)} \tag{129}$$

where the right side is the qth moment of the beta$(\frac{1}{2}, \frac{1}{2}+\theta)$ structural distribution of PD$(\frac{1}{2},\theta)$, and on the left side this moment is computed by conditioning on L_1. As in (80), for each fixed q this formula provides a Mellin transform which uniquely determines $\mu(q\,||\,\lambda)$ as a function of λ. In view of (129) and (93), the formula (99) for $\mu(q\,||\,\lambda)$ in terms of the Hermite function amounts to the identity

$$2\int_0^\infty \lambda^{2\theta} h_{-2q}(\lambda)\phi(\lambda)d\lambda = 2^{-\theta-q}\frac{\Gamma(2\theta+1)}{\Gamma(q+\theta+1)}. \tag{130}$$

As checks, since $h_0(x) = 1$ and $h_{-1}(x) = \Phi(x)/\varphi(x)$, the case $q = 0$ is obvious, and the case $q = \frac{1}{2}$ is easily verified since then the left side of (129) equals $(2\theta+1)^{-1}\mathbb{E}(|B_1|^{2\theta+1})$ by integration by parts. Formula (130) can then be verified for $q = m/2$ for all $m = 0, 1, 2, \ldots$, using the recursion (106). Formula (130) was just derived for $q > -\frac{1}{2}$, but both sides are entire functions of q, so the identity holds for all $q \in \mathbb{C}$. Using the series formula (100) and integrating term by term, the substitution $r = \theta + \frac{1}{2}$ allows the identity (130) to be rewritten in the symmetric form

$$\sum_{j=0}^{\infty} \Gamma\left(q+\frac{j}{2}\right)\Gamma\left(r+\frac{j}{2}\right)\frac{(-2)^j}{j!} = \frac{4\sqrt{\pi}\Gamma(2q)\Gamma(2r)}{\Gamma(q+r+1/2)} \tag{131}$$

where the series is absolutely convergent for real q and r with $q+r+\frac{1}{2} < -1$, and can otherwise be summed by Abel's method provided neither $2q$ nor $2r$ is a non-positive integer. This version of the identity is easily verified using standard identities involving Gauss's hypergeometric function and the gamma function. For $-2q = n$ a positive integer, when h_n is the nth Hermite polynomial

$$h_n(x) = \sum_{k=0}^{\lfloor n/2 \rfloor} h_{n,k}x^{n-2k} \text{ with } h_{n,k} = (-1)^k\binom{n}{2k}\frac{(2k)!}{2^k k!}$$

the identity (130) reduces easily to the following pair of identities of polynomials in θ, which relate the rising and falling factorials $[x]_n := x(x+1)\cdots(x+n-1)$ and $(x)_n := x(x-1)\cdots(x-n+1)$, and which are easily verified directly: for $m = 0, 1, 2\ldots$

$$\sum_{k=0}^{m} h_{2m,k}2^{-k}[\theta+\tfrac{1}{2}]_{m-k} = (\theta)_m$$

and

$$\sum_{k=0}^{m} h_{2m+1,k}2^{-k}[\theta+1]_{m-k} = (\theta-\tfrac{1}{2})_m.$$

Thus the coefficients of the Hermite polynomials are related to some instances of generalized Stirling numbers [22, 48].

Acknowledgments

Thanks to Grégory Miermont for careful reading of a draft of this paper. This research is supported in part by NSF grants MCS–9404345 and DMS–0071448.

Jim Pitman, Department of Statistics, University of California, Berkeley,
pitman@stat.berkeley.edu

References

[1] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions*. Dover, New York, 1965.

[2] D. Aldous and J. Pitman. Brownian bridge asymptotics for random mappings. *Random Structures and Algorithms*, 5:487–512, 1994.

[3] D.J. Aldous and J. Pitman. The standard additive coalescent. *Annals of Probability*, 26:1703–1726, 1998.

[4] R. A. Arratia, A. D. Barbour, and S. Tavaré. Logarithmic combinatorial structures: a probabilistic approach. Book in preparation. Available via http://www-hto.usc.edu/books/tavare/ABT/index.html, 2001.

[5] J. Bertoin. A fragmentation process connected to Brownian motion. *Probability Theory and Related Fields*, 117(2):289–301, 2000.

[6] J. Bertoin. Self-similar fragmentations. *Annales de l'Institut Henri Poincaré. Probabilités et Statististiques*, 38(3):319–340, 2002.

[7] D. Blackwell and J.B. MacQueen. Ferguson distributions via Pólya urn schemes. *Annals of Statistics*, 1:353–355, 1973.

[8] A. Brix. Generalized gamma measures and shot-noise Cox processes. *Advances in Applied Probability*, 31(4):929–953, 1999.

[9] P. Chassaing and S. Janson. A Vervaat-like path transformation for the reflected Brownian bridge conditioned on its local time at 0. *Annals of Probability*, 29(4):1755–1779, 2001.

[10] P. Chassaing and G. Louchard. Phase transition for parking blocks, Brownian excursion and coalescence. *Random Structures Algorithms*, 21:76–119, 2002.

[11] B. Derrida. Random-energy model: an exactly solvable model of disordered systems. *Physical Review B (3)*, 24(5):2613–2626, 1981.

[12] B. Derrida. Non-self-averaging effects in sums of random variables, spin glasses, random maps and random walks. In M. Fannes, C. Maes, and A. Verbeure, editors, *On Three Levels: Micro-, Meso-, and Macro-Approaches in Physics*, NATO ASI Series, pages 125–137. Plenum Press, New York and London, 1994.

[13] B. Derrida. From random walks to spin glasses. *Phys. D*, 107(2-4):186–198, 1997. Landscape paradigms in physics and biology (Los Alamos, NM, 1996).

[14] G. Doetsch. *Theorie und Anwendung der Laplace-Transformation.* Berlin, 1937.

[15] S. Engen. *Stochastic Abundance Models with Emphasis on Biological Communities and Species Diversity.* Chapman and Hall Ltd., 1978.

[16] A. Erdélyi et al. *Higher Transcendental Functions*, volume II of *Bateman Manuscript Project.* McGraw-Hill, New York, 1953.

[17] W. J. Ewens and S. Tavaré. The Ewens sampling formula. In N. S. Johnson, S. Kotz, and N. Balakrishnan, editors, *Multivariate Discrete Distributions.* Wiley, New York, 1995.

[18] W.J. Ewens. The sampling theory of selectively neutral alleles. *Theoretical Population Biology*, 3:87 – 112, 1972.

[19] M. Grote and T. P. Speed. Approximate Ewens formulae for symmetric overdominance selection. *Annals of Applied Probability*, 12: 637 – 663, 2002.

[20] B. Hansen and J. Pitman. Prediction rules and exchangeable sequences related to species sampling. *Statistics and Probility Letters*, 46(520):251–256, 2000.

[21] F. M. Hoppe. The sampling theory of neutral alleles and an urn model in population genetics. *Journal of Mathematical Biology*, 25:123 – 159, 1987.

[22] L. C. Hsu and P. J.-S. Shiue. A unified approach to generalized Stirling numbers. *Advances in Applied Mathematics*, 20(3):366–384, 1998.

[23] L. F. James. Poisson process partition calculus with applications to exchangeable models and Bayesian nonparametrics. arXiv:math.PR/0205093, 2002.

[24] L.F. James. Bayesian calculus for gamma processes with applications to semiparametric intensity models. *Sankhyā*, 2002. to appear.

[25] S. Janson. Asymptotic distribution for the cost of linear probing hashing. *Random Structures Algorithms*, 19(3-4):438–471, 2001. Analysis of algorithms (Krynica Morska, 2000).

[26] N. L. Johnson and S. Kotz. *Continuous Univariate Distributions, volume 2.* Wiley, 1970.

[27] S. Karlin. Central limit theorems for certain infinite urn schemes. *Journal of Mathematics and Mechanics*, 17:373–401, 1967.

[28] S. Kerov. Coherent random allocations and the Ewens-Pitman formula. PDMI Preprint, Steklov Math. Institute, St. Petersburg, 1995.

[29] J. F. C. Kingman. Random discrete distributions. *Journal of the Royal Statistical Society B*, 37:1–22, 1975.

[30] J. F. C. Kingman. The representation of partition structures. *Journal of the London Mathematical Society*, 18:374–380, 1978.

[31] J. F. C. Kingman. The coalescent. *Stochastic Processes and their Applications*, 13:235–248, 1982.

[32] J. F. C. Kingman. Exchangeability and the evolution of large populations. In G. Koch and F. Spizzichino, editors, *Exchangeability in probability and statistics (Rome, 1981)*, pages 97–112. North-Holland, Amsterdam, 1982.

[33] J. F. C. Kingman. *Poisson Processes*. Clarendon Press, Oxford, 1993.

[34] U. Küchler and M. Sørensen. *Exponential families of stochastic processes*. Springer-Verlag, New York, 1997.

[35] N. N. Lebedev. *Special Functions and their Applications*. Prentice-Hall, Englewood Cliffs, N.J., 1965.

[36] P. Lévy. Sur certains processus stochastiques homogènes. *Compositio Mathematica*, 7:283–339, 1939.

[37] J. W. McCloskey. A model for the distribution of individuals by species in an environment. Ph. D. thesis, Michigan State University, 1965.

[38] G. Miermont and J. Schweinsberg. Self-similar fragmentations and stable subordinators. Technical Report 726, 2001. Prépublication du Laboratoire de Probabilités et modèles aléatoires, Universite Paris VI. Available via `http://www.proba.jussieu.fr`.

[39] J. C. P. Miller. *Tables of Weber parabolic cylinder functions*. Her Majesty's Stationery Office, London, 1955.

[40] M. Perman. Order statistics for jumps of normalized subordinators. *Stochastic Processes and their Applications*, 46:267–281, 1993.

[41] M. Perman, J. Pitman, and M. Yor. Size-biased sampling of Poisson point processes and excursions. *Probability Theory and Related Fields*, 92:21–39, 1992.

[42] J. Pitman. Poisson-Kingman partitions. Preprint available via `http://www.stat.berkeley.edu/users/pitman`, 1995.

[43] J. Pitman. Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields*, 102:145–158, 1995.

[44] J. Pitman. Random discrete distributions invariant under size-biased permutation. *Advances in Applied Probability*, 28:525–539, 1996.

[45] J. Pitman. Some developments of the Blackwell-MacQueen urn scheme. In T.S. Ferguson et al., editor, *Statistics, Probability and Game Theory; Papers in honor of David Blackwell*, volume 30 of *Lecture Notes-Monograph Series*, pages 245–267. Institute of Mathematical Statistics, Hayward, California, 1996.

[46] J. Pitman. Partition structures derived from Brownian motion and stable subordinators. *Bernoulli*, 3:79–96, 1997.

[47] J. Pitman. Brownian motion, bridge, excursion and meander characterized by sampling at independent uniform times. *Electronic Journal of Probability*, 4:Paper 11, 1–33, 1999.

[48] J. Pitman. Combinatorial Stochastic Processes. Technical Report 621, Dept. Statistics, U.C. Berkeley, 2002. Lecture notes for St. Flour course, July 2002. Available via `http://www.stat.berkeley.edu`.

[49] J. Pitman and M. Yor. Arcsine laws and interval partitions derived from a stable subordinator. *Proceedings of the London Mathematical Society (3)*, 65:326–356, 1992.

[50] J. Pitman and M. Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability*, 25:855–900, 1997.

[51] J. Pitman and M. Yor. Brownian interpretations of Mill's ratio, Hermite and parabolic cylinder functions. In preparation, 2002.

[52] H. Pollard. The representation of $e^{-x^{\lambda}}$ as a Laplace integral. *Bulletin of the American Mathematical Society*, 52:908–910, 1946.

[53] D. Ruelle. A mathematical reformulation of Derrida's REM and GREM. *Communications in Mathematical Physics*, 108:225–239, 1987.

[54] V. Seshadri. *The inverse Gaussian distribution*. The Clarendon Press, Oxford University Press, New York, 1993.

[55] M. Talagrand. *Spin glasses: a challenge to mathematicians*. Springer, 2001. Book in preparation.

[56] L. Toscano. Sulle funzioni del cilindro parabolico. *Matematiche (Catania)*, 26:104–126 (1972), 1971.

[57] S.L. Zabell. The continuum of inductive methods revisited. In J. Earman and J. D. Norton, editors, *The Cosmos of Science*, Pittsburgh-Konstanz Series in the Philosophy and History of Science, pages 351–385. University of Pittsburgh Press/Universitätsverlag Konstanz, 1997.

Diffusions on the Simplex from Brownian Motions on Hypersurfaces

Steven N. Evans

Abstract

The $(n-1)$-dimensional simplex is the collection of probability measures on a set with n points. Many applied situations result in simplex-valued data or in stochastic processes that have the simplex as their state space. In this paper we study a large class of simplex-valued diffusion processes that are constructed by first "coordinatising" the simplex with the points of a smooth hypersurface in such a way that several points on the hypersurface may correspond to a given point on the simplex, and then mapping forward the canonical Brownian motion on the hypersurface. For example, a particular instance of the Fleming-Viot process on n points arises from Brownian motion on the $(n-1)$-dimensional sphere. The Brownian motion on the hypersurface has the normalised Riemannian volume as its equilibrium distribution. It is straightforward to compute the corresponding distribution on the simplex, and this provides a large class of interesting probability measures on the simplex.

Keywords: manifold; stochastic differential equation; measure-valued process; compositional data; Riemannian volume element; Fleming-Viot process

1 Introduction

Many data sets come in the form of proportions that add to unity (that is, as points in a simplex with dimension one less than the number of proportions). For example, there is the breakdown of the composition of an ore sample into component minerals or the division of a family's expenditures into housing, food, clothing, leisure, *etc.* This type of data is often referred to as *compositional* and a standard reference for models and inference in this area is [1].

Such data can also have a temporal component. For example, there are the proportions of the population at any time having each of the possible combinations of alleles of a given set of genes (see, for example, [5]). There appears to be something of a dearth of flexible, tractable models for such stochastic processes.

Of course, stochastic processes on the simplex are an elementary instance of processes taking values in the set of probability measures on an arbitrary measurable space. However, the literature in this more general area is primarily concerned with models

such as the Fleming-Viot process that arise as continuum limits of particle systems with relatively simple dynamics (see, for example, [2]).

There is a substantial literature on diffusions on manifolds and particularly Brownian motion on manifolds (see, for example, [3, 4, 6, 8]). The approach we follow here for building diffusions on the simplex is to first take a simplicial decomposition of some compact manifold. This gives a typically many-to-one mapping of the manifold onto the simplex. We then take Brownian motion on the manifold and map it forward to obtain a continuous stochastic process on the simplex. If the manifold and the associated simplicial decomposition have suitable symmetry properties, then the resulting process on the simplex will be Markovian.

The simplest example of our construction is when the manifold is the $(n-1)$-dimensional sphere

$$\{(x^1,x^2,\ldots,x^n) : (x^1)^2+(x^2)^2+\cdots+(x^n)^2=1\}.$$

We map the sphere onto the $(n-1)$-dimensional simplex via

$$(x^1,x^2,\ldots,x^n)\mapsto((x^1)^2,(x^2)^2,\ldots,(x^n)^2).$$

If $(X_t,\mathbb{P}^x)$ is the Brownian motion on the sphere, then the distribution of the process $X=(X^1,X^2,\ldots,X^n)$ under $\mathbb{P}^{(\pm x^1,\pm x^2,\ldots,\pm x^n)}$ is the same as the distribution of $(\pm X^1, \pm X^2, \ldots, \pm X^n)$ under $\mathbb{P}^x$ for any x and any of the 2^n possible choices of sign. In particular, for any point $y=(y^1,y^2,\ldots,y^n)$ in the simplex the distribution of $((X^1)^2,(X^2)^2,\ldots,(X^n)^2)$ is the same under any of the measures $\mathbb{P}^x$ for which $((x^1)^2,(x^2)^2,\ldots,(x^n)^2)=(y^1,y^2,\ldots,y^n)$. Dynkin's criterion for a function of a Markov process to be Markovian (see Theorem 13.5 of [7]) then gives that $((X^1)^2,(X^2)^2,\ldots,(X^n)^2)$ is Markovian.

It turns out that Brownian motion on the sphere is mapped to a particular Fleming-Viot process on the set $\{1,2,\ldots,n\}$. The underlying mutation process for the Fleming-Viot process is a Markov chain that jumps at a constant rate and chooses a new state uniformly from the $(n-1)$ possibilities. The Brownian motion on the sphere has the normalised surface area measure on the sphere as its equilibrium distribution. The corresponding process on the simplex (that is, the Fleming-Viot process) has the push-forward of this measure as its equilibrium distribution and, as is well-known, this latter probability measure is the Dirichlet distribution with parameters $(\frac{1}{2},\frac{1}{2},\ldots,\frac{1}{2})$.

The plan of the paper is the following. We construct a particular class of hypersurfaces and Brownian motions on them in Section 2. We show that the Brownian motion mapped to the simplex is Markovian in Section 3, and exhibit the semimartingale decomposition of this diffusion on the simplex in Section 4. The push-forward of the normalised Riemannian volume measure is the equilibrium distribution of the diffusion on the simplex, and an explicit formula is given for this distribution in Section 5. We illustrate the general results with the special cases where the hypersurface is an ellipsoid in $\mathbb{R}^n$ or the unit sphere in $\mathbb{R}^n$ equipped with the ℓ^p norm for p an even positive integer.

2 Brownian motion on a hypersurface

Fix functions $g_i : \mathbb{R} \to \mathbb{R}_+$, $1 \leq i \leq n$, with the following properties:

i) g_i is C^∞;

ii) $g_i(0) = 0$;

iii) $g_i(-u) = g_i(u)$;

iv) $g_i'(u) > 0, u > 0$;

v) $\{u \in \mathbb{R} : g_i(u) = 1\} \neq \emptyset$.

Define $g : \mathbb{R}^n \to \mathbb{R}_+^n$ by

$$g(x^1, x^2, \ldots, x^n) := (g_1(x^1), g_2(x^2), \ldots, g_n(x^n))$$

and $G : \mathbb{R}^n \to \mathbb{R}_+$ by

$$G(x^1, x^2, \ldots, x^n) := \sum_{i=1}^n g_i(x^i).$$

The set $\mathcal{M} := \{x \in \mathbb{R}^n : G(x) = 1\}$ is a compact, connected, $(n-1)$-dimensional embedded submanifold of $\mathbb{R}^n$ and the range of g restricted to $\mathcal{M}$ is the simplex

$$\mathcal{S} := \{y \in \mathbb{R}^n : \sum_{i=1}^n y^i = 1, y^i \geq 0\}.$$

Each $y \in \mathcal{S}$ is the image of $2^{\#\{1 \leq i \leq n : y^i > 0\}}$ points of $\mathcal{M}$.

We will construct a diffusion process $Y = (Y_t, \mathbb{Q}^y)$ or $\mathcal{S}$ by letting $(Y_t)_{t \geq 0}$ under $\mathbb{Q}^y$ have the law of $(g \circ X_t)_{t \geq 0}$ under $\mathbb{P}^x$, where $X = (X_t, \mathbb{P}^x)$ is the canonical Brownian motion on $\mathcal{M}$ and x is any pre-image of y for g. The infinitesimal generator of X is a multiple of the Laplace-Beltrami operator on $\mathcal{M}$, but the most convenient way for us to describe X is as the solution of a stochastic differential equation (SDE).

Let

$$\begin{aligned} n(x) &:= \frac{\text{grad } G(x)}{\|\text{grad } G(x)\|} \\ &= \frac{(g_1'(x^1), g_2'(x^2), \ldots, g_n'(x^n))}{(\sum_{i=1}^n g_i'(x^i)^2)^{\frac{1}{2}}} \end{aligned}$$

be the unit normal to $\mathcal{M}$ at x, and write

$$P(x) := (I - n(x)n(x)^\top)$$

for the corresponding orthogonal projection onto the tangent plane to $\mathcal{M}$ at x. The mean curvature at x is given by

$$\begin{aligned} c(x) &:= -\frac{1}{2}\operatorname{div} n(x) \\ &= -\frac{1}{2}\left\{ \frac{\sum_i g_i''(x^i)}{(\sum_i g_i'(x^i)^2)^{\frac{1}{2}}} - \frac{\sum_i g_i'(x^i)^2 g_i''(x^i)}{(\sum_i g_i'(x^i)^2)^{\frac{3}{2}}} \right\} \\ &= -\frac{1}{2}\frac{\sum_{i\neq j} g_i'(x^i)^2 g_j''(x^j)}{(\sum_i g_i(x^i)^2)^{\frac{3}{2}}}. \end{aligned}$$

By [9], Brownian motion on $\mathcal{M}$ starting at $x \in \mathcal{M}$ solves the SDE

$$\begin{aligned} dX_t &= P(X_t)\, dB_t + c(X_t)\, n(X_t)\, dt \\ X_0 &= x, \end{aligned}$$

where B is a standard n-dimensional Brownian motion. Write $\mathbb{P}^x$ for the distribution of the solution of this SDE.

3 Diffusion on the simplex

Set $Y := g \circ X$. That is, $Y_t = g(X_t) \in \mathcal{S}$. We claim that Y is Markovian. As with the example on the sphere in the Introduction, this will follow from Dynkin's criterion for a function of a Markov process to be Markovian if we can show that the law of Y is the same under $\mathbb{P}^{x'}$ and $\mathbb{P}^{x''}$ for any two points $x', x'' \in \mathcal{M}$ such that $g(x') = g(x'')$ (see Theorem 13.5 of [7]).

For any $x \in \mathcal{M}$, let $X^{(x)}$ denote the solution of the SDE

$$\begin{aligned} dX_t^{(x)} &= P\left(X_t^{(x)}\right) dB_t + c\left(X_t^{(x)}\right) n\left(X_t^{(x)}\right) dt \\ X_0^{(x)} &= x. \end{aligned}$$

Fix $\varepsilon \in \{\pm 1\}^n$ and write E for the diagonal matrix $\operatorname{diag}(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$ so that for $z \in \mathbb{R}^n$, $Ez = (\varepsilon_1 z^1, \varepsilon_2 z^2, \ldots, \varepsilon_n z^n)$. Note that if $x', x'' \in \mathcal{M}$ are such that $g(x') = g(x'')$, then $x'' = Ex'$ for some such E. Observe by our assumptions on the g_i that

$$\begin{aligned} g_i'(-u) &= -g_i'(u), \\ g_i''(-u) &= g_i''(u), \end{aligned}$$

and so

$$n(Ex) = En(x),$$

$$P(Ex) = EP(x)E,$$

$$c(Ex) = c(x).$$

Thus,

$$\begin{aligned} d\left[EX_t^{(x)}\right] &= EP\left(X_t^{(x)}\right) dB_t + c\left(X_t^{(x)}\right) En\left(X_t^{(x)}\right) dt \\ &= EP\left(X_t^{(x)}\right) E\, d[EB_t] + c\left(X_t^{(x)}\right) En\left(X_t^{(x)}\right) dt \\ &= P\left(EX_t^{(x)}\right) d\tilde{B}_t + c\left(EX_t^{(x)}\right) n\left(EX_t^{(x)}\right) dt, \end{aligned}$$

where $\tilde{B} = EB$ is a standard n-dimensional Brownian motion. Moreover,

$$EX_0^{(x)} = Ex,$$

and so we conclude that $EX^{(x)}$ has the same distribution as $X^{(Ex)}$. That is, the law of EX under $\mathbb{P}^x$ is the same as that of X under $\mathbb{P}^{Ex}$, and Dynkin's criterion holds. Write $\mathbb{Q}^y$ for the distribution of Y starting at $y \in \mathcal{S}$. Because X is a Feller process and g is continuous, it follows that Y is also a Feller process.

4 Semimartingale description

By Itô's formula we have

$$\begin{aligned} dY_t^i &= g_i'(X_t^i) dX_t^i + \frac{1}{2} g_i''(X_t^i) d\langle X^i \rangle_t \\ &= g_i'(X_t^i) \sum_j P_{ij}(X_t) dB_t^j \\ &\quad + g_i'(X_t^i) c(X_t) n^i(X_t) dt + \frac{1}{2} g_i''(X_t^i) \sum_j P_{ij}(X_t)^2 dt. \end{aligned}$$

By our assumptions on g_i, for $0 \le v \le 1$ there exists a unique $u \ge 0$ such that $g_i(u) = v$. Write $u = h_i(v)$. Observe that $g_i(-h_i(v)) = v$, $g_i'(-h_i(v)) = -g_i'(h_i(v))$, and $g_i''(-h_i(v)) = g_i''(h_i(v))$. Put

$$\begin{aligned} \alpha_i(y) &:= g_i'(h_i(y^i))^2 = \frac{1}{h_i'(y^i)^2}, \\ \bar{\alpha}_i(y) &:= \frac{\alpha_i(y)}{\sum_j \alpha_j(y)}, \end{aligned}$$

and

$$\beta_i(y) := g_i''(h_i(y^i)) = -\frac{h_i''(y^i)}{h_i'(y^i)^3}.$$

Note that because $P(x)$ is a projection matrix,

$$\begin{aligned}\sum_j P_{ij}(x)^2 &= \sum_j P_{ij}(x)P_{ji}(x) \\ &= P_{ii}(x) \\ &= 1 - n^i(x)^2.\end{aligned}$$

Thus for $y = g(x)$ we have

$$\begin{aligned}&g_i'(x^i)c(x)n^i(x) + \frac{1}{2}g_i''(x^i)\sum_j P_{ij}(x)^2 \\ &= -\frac{1}{2}\alpha_i(y)\left\{\frac{\sum_j \beta_j(y)}{(\sum_j \alpha_j(y))^{\frac{1}{2}}} - \frac{\sum_j \alpha_j(y)\beta_j(y)}{(\sum_j \alpha_j(y))^{\frac{3}{2}}}\right\}\frac{1}{(\sum_j \alpha_j(y))^{\frac{1}{2}}} \\ &\quad + \frac{1}{2}\beta_i(y)\left\{1 - \frac{\alpha_i(y)}{\sum_j \alpha_j(y)}\right\} \\ &= \frac{1}{2}\left[\beta_i(y)\{1-\bar{\alpha}_i(y)\} - \bar{\alpha}_i(y)\sum_j \beta_j(y)\{1-\bar{\alpha}_j(y)\}\right].\end{aligned}$$

Note also that

$$\begin{aligned}\sum_k g_i'(x^i)P_{ik}(x)g_j'(x^j)P_{jk}(x) &= g_i'(x^i)P_{ij}(x)g_j'(x^j) \\ &= \alpha_i(y)\{\delta_{ij} - \bar{\alpha}_j(y)\}.\end{aligned}$$

Putting this all together,

$$dY_t^i = dM_t^i + \frac{1}{2}\left[\beta_i(Y_t)\{1-\bar{\alpha}_i(Y_t)\} - \bar{\alpha}_i(Y_t)\sum_j \beta_j(Y_t)\{1-\bar{\alpha}_j(Y_t)\}\right]dt$$

where $M_t = (M_t^1, \ldots, M_t^n)$ is a continuous martingale with

$$d\langle M^i, M^j\rangle_t = \alpha_i(Y_t)\{\delta_{ij} - \bar{\alpha}_j(Y_t)\}dt.$$

Example 1

Suppose that $g_i(u) = c_i u^2$ for constants $c_i > 0$, $1 \le i \le n$, so that $\mathcal{M}$ is the ellipsoid $\{(x^1, x^2, \ldots, x^n) : c_1(x^1)^2 + c_2(x^2)^2 + \cdots + c_n(x^n)^2 = 1\}$. Then

$$\begin{aligned}\alpha_i(y) &= 4c_i y^i, \\ \beta_i(y) &= 2c_i,\end{aligned}$$

and hence

$$dY_t^i = dM_t^i + \left[c_i\left\{1 - \frac{c_i Y_t^i}{\sum_j c_j Y_t^j}\right\} - \frac{c_i Y_t^i}{\sum_j c_j Y_t^j}\sum_j c_j\left\{1 - \frac{c_j Y_t^j}{\sum_k c_k Y_t^k}\right\}\right]dt,$$

where M is a continuous martingale with

$$d\langle M^i, M^j\rangle_t = 4c_i Y_t^i \left\{ \delta_{ij} - \frac{c_j Y_t^j}{\sum_k c_k Y_t^k} \right\} dt.$$

When $c_1 = c_2 = \cdots = c_n = c$ (so that $\mathcal{M}$ is the sphere with radius $\frac{1}{\sqrt{c}}$), we have

$$\begin{aligned} dY_t^i &= dM_t^i + c\left[1 - Y_t^i - Y_t^i \sum_j \{1 - Y_t^j\}\right] dt \\ &= dM_t^i + c[1 - nY_t^i]dt \\ &= dM_t^i + cn \sum_j \left\{ \frac{1}{n} - \delta_{ij} \right\} Y_t^j \, dt, \end{aligned}$$

where M is a continuous martingale with

$$d\langle M^i, M^j\rangle_t = 4cY_t^i\{\delta_{ij} - Y_t^j\}dt.$$

If we associate Y_t with the probability measure on $\{1,2,\ldots,n\}$ that assigns mass Y_t^i to i, then $(Y_t, \mathbb{Q}^y)$ is a particular case of a Fleming–Viot process (see [2]) in which the underlying mutation process jumps from each state at rate $c(n-1)$ and chooses a new state uniformly from the $(n-1)$ possibilities.

When $n = 2$, the process $Z := Y^1$ is a one-dimensional diffusion that solves the SDE

$$dZ_t = \mu(Z_t)\, dt + \sigma(Z_t)\, dB_t,$$

where

$$\begin{aligned} \mu(z) := c_1 &\left\{ 1 - \frac{c_1 z}{c_1 z + c_2(1-z)} \right\} \\ &- \frac{c_1 z}{c_1 z + c_2(1-z)} \left[c_1 \left\{ 1 - \frac{c_1 z}{c_1 z + c_2(1-z)} \right\} + c_2 \left\{ 1 - \frac{c_2(1-z)}{c_1 z + c_2(1-z)} \right\} \right], \end{aligned}$$

and

$$\sigma^2(z) := 4c_1 z \left\{ 1 - \frac{c_1 z}{c_1 z + c_2(1-z)} \right\}.$$

An interesting feature of these coefficients is that the unique zero of μ and the unique maximum of σ^2 both occur at the point $z = \sqrt{c_2}/(\sqrt{c_1} + \sqrt{c_2})$. The infinitesimal drift μ is graphed in Figures 1 and 2 for the parameter values $(c_1, c_2) = (1,1)$ and $(c_1, c_2) = (4,1)$, respectively. The infinitesimal variance σ^2 is graphed in Figures 3 and 4 for the parameter values $(c_1, c_2) = (1,1)$ and $(c_1, c_2) = (4,1)$, respectively.

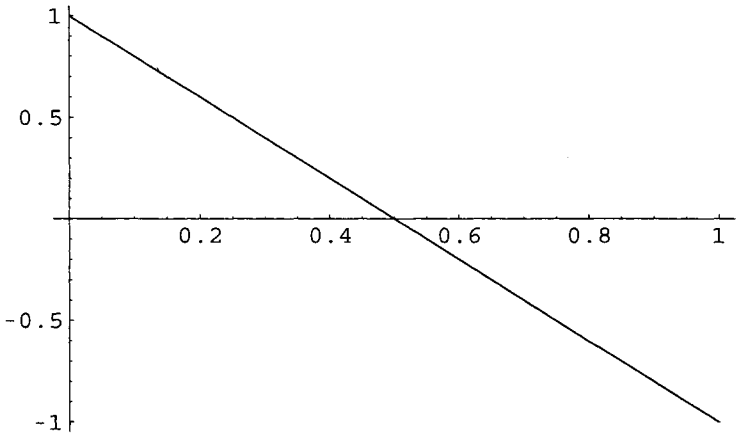


Figure 1: Drift for $c_1 = 1$ and $c_2 = 2$

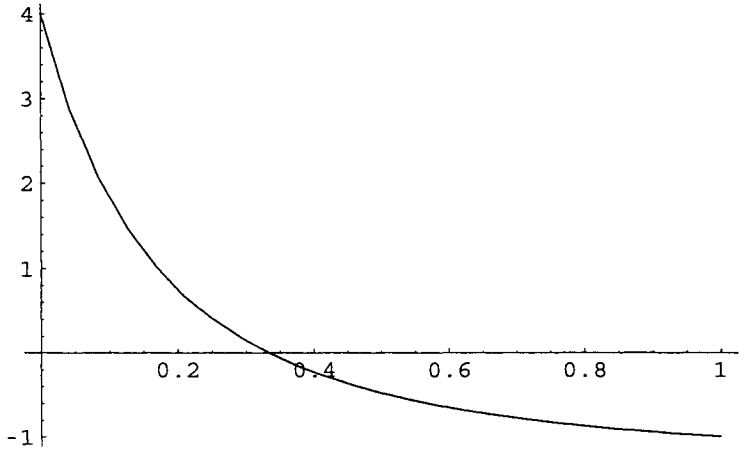


Figure 2: Drift for $c_1 = 4$ and $c_2 = 1$

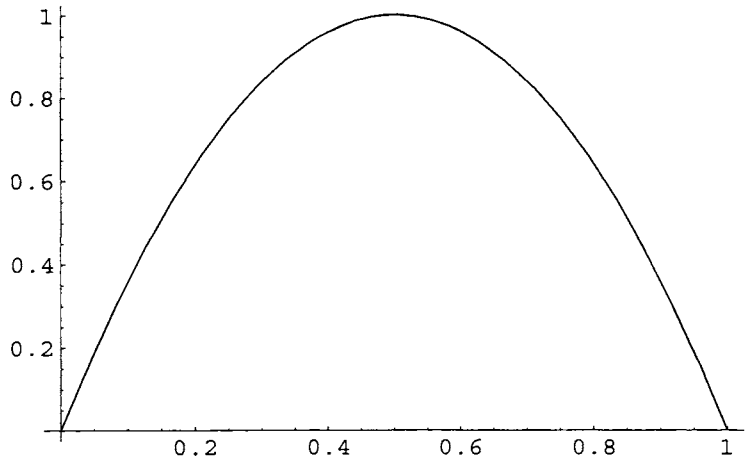


Figure 3: Variance for $c_1 = 1$ and $c_2 = 1$

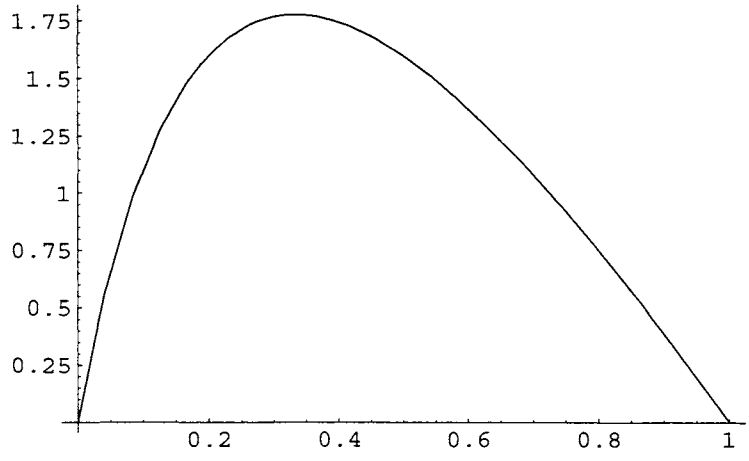


Figure 4: Variance for $c_1 = 4$ and $c_2 = 1$

Remark 1 *If we formally send $n \to \infty$ in the martingale problem for Y, then the resulting martingale problem on the infinite simplex $\{(y^1, y^2, \ldots) : \sum_i y^i = 1, y^i \geq 0\}$ makes sense when*

$$\sum_i \sup_{0 \leq v \leq 1} \left| \frac{h_i''(v)}{h_i'(v)^3} \right| < \infty$$

(in Example 1 this condition becomes $\sum_i c_i < \infty$). It would be interesting to know if this infinite-dimensional martingale problem is well-posed.

5 Equilibrium distribution

The Brownian motion X is reversible with respect to the normalised Riemannian volume measure on $\mathcal{M}$ and X_t converges in distribution to this measure as $t \to \infty$ under any $\mathbb{P}^x$. Therefore, if we let π denote the push-forward of the normalised Riemannian volume measure by g, then the diffusion Y is reversible with respect to π and Y_t converges in distribution to π as $t \to \infty$ under any $\mathbb{Q}^y$.

We can calculate the Riemannian volume measure as follows. The set

$$\{(x^1, x^2, \ldots, x^n) \in \mathcal{M} : x^n \neq 0\}$$

is the union of the two open sets

$$\left\{ \left(x^1, x^2, \ldots, x^{n-1}, \pm h_n \left(1 - \sum_{i=1}^{n-1} g_i(x^i) \right) \right) : \sum_{i=1}^{n-1} g_i(x^i) < 1 \right\}$$

and $(x^1, x^2, \ldots, x^{n-1})$ can be used as local coordinates for $\mathcal{M}$ in these two patches. The Riemannian metric in each patch is given by the matrix $I + J(x)J(x)^\top$, where $J(x)$ is the $(n-1)$-dimensional column vector

$$\left(\frac{\partial}{\partial x^i} h_n \left(1 - \sum_{j=1}^{n-1} g_j(x^j) \right) \right)_{i=1}^{n-1}.$$

The corresponding Riemannian volume measure is

$$[\det(I + J(x)J(x)^\top)]^{\frac{1}{2}} \, dx^1 dx^2 \cdots dx^{n-1} = [1 + J(x)^\top J(x)]^{\frac{1}{2}} \, dx^1 dx^2 \cdots dx^{n-1},$$

where we have used the familiar matrix fact that

$$\det(A + bb^\top) = \det(A)(1 + b^\top A^{-1} b).$$

The Jacobian matrix for the transformation

$$(x^1, x^2, \ldots, x^{n-1}) \mapsto (g_1(x^1), g^2(x^2), \ldots, g_{n-1}(x^{n-1}))$$

is the diagonal matrix $\mathrm{diag}(g_1'(x^1), g_2'(x^2), \ldots, g_{n-1}'(x^{n-1}))$. Therefore, if we coordinatise S with $\{(y^1, y^2, \ldots, y^{n-1}) : \sum_{i=1}^{n-1} y^i \leq 1,\ y^i \geq 0\}$, then π is the measure

$$C\left[1 + \sum_{i=1}^{n-1}\left\{h_n'\left(1 - \sum_{j=1}^{n-1} y^j\right) g_i'(h_i(y^i))\right\}^2\right]^{\frac{1}{2}} \prod_{i=1}^{n-1} g_i'(h_i(y^i))^{-1}\, dy^1 dy^2 \ldots dy^{n-1}$$

$$= C\left[\sum_{i=1}^{n} g_i'(h_i(y^i))^2\right]^{\frac{1}{2}} \prod_{i=1}^{n} g_i'(h_i(y^i))^{-1}\, dy^1 dy^2 \cdots dy^{n-1},$$

for a suitable normalisation constant C.

Example 2
Suppose that $g_i(u) = c_i u^2$ for constants $c_i > 0$, $1 \leq i \leq n$. Then

$$g_i'(h_i(u)) = 2c_i^{\frac{1}{2}} u^{\frac{1}{2}}$$

so that π is

$$C\left[\sum_{i=1}^{n} c_i y^i\right]^{\frac{1}{2}} \prod_{i=1}^{n} (y^i)^{-\frac{1}{2}}\, dy^1 dy^2 \cdots dy^{n-1}$$

for a suitable constant C. In particular, if $c_1 = c_2 = \cdots = c_n$, then π is the Dirichlet distribution with parameters $\left(\frac{1}{2}, \frac{1}{2}, \ldots, \frac{1}{2}\right)$.

For $n = 2$, the equilbrium density is graphed in Figures 5 and 6 for $(c_1, c_2) = (1,1)$ and $(c_1, c_2) = (4,1)$, respectively. The equilibrium density has its unique minimum at $\sqrt{c_2}/(\sqrt{c_1} + \sqrt{c_2})$. Recall from Example 1 that the infinitesimal drift coefficient vanishes and the infinitesimal variance coefficient has its maximum at this same point.

6 Another example

Suppose that $g_i(u) = u^p$, $1 \leq i \leq n$, where p is an even positive integer. Then

$$\alpha_i(y) = p^2 (y^i)^{2\left(1 - \frac{1}{p}\right)} \quad \text{and} \quad \beta_i(y) = p^2 (p-1) (y^i)^{\left(1 - \frac{2}{p}\right)}.$$

Hence, setting $r = 2\left(1 - \frac{1}{p}\right)$ and $s = \left(1 - \frac{2}{p}\right)$,

$$dY_t^i = dM_t^i + \frac{p^2(p-1)}{2}\left[(Y_t^i)^s \left\{1 - \frac{(Y_t^i)^r}{\sum_k (Y_t^k)^r}\right\} - \frac{(Y_t^i)^r}{\sum_k (Y_t^k)^r} \sum_j (Y_t^j)^s \left\{1 - \frac{(Y_t^j)^r}{\sum_k (Y_t^k)^r}\right\}\right],$$

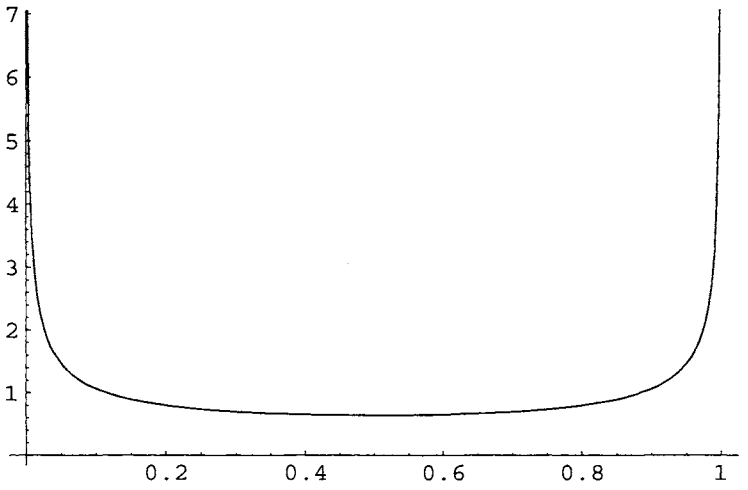


Figure 5: Equilibrium density for $c_1 = 1$ and $c_2 = 1$

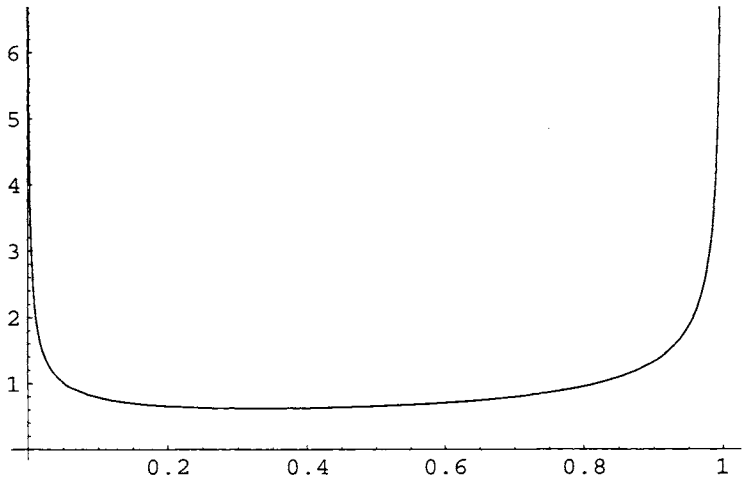


Figure 6: Equilibrium density for $c_1 = 4$ and $c_2 = 1$

where M is a continuous martingale with

$$d\langle M^i, M^j\rangle_t = p^2 (Y_t^i)^r \left\{ \delta_{ij} - \frac{(Y_t^j)^r}{\sum_k (Y_t^k)^r} \right\} dt.$$

The equilibrium measure π is

$$C \left[\sum_{i=1}^n (y^i)^r \right]^{\frac{1}{2}} \prod_{i=1}^n (y^i)^{-\frac{r}{2}} \, dy^1 dy^2 \cdots dy^{n-1}$$

for some constant C.

Acknowledgments

Research supported in part by NSF grant DMS-0071468 and a research professorship from the Miller Institute for Basic Research in Science.

Dedication

Dedicated with profound admiration to Terry, my neighbour and compatriot. Thanks for thirteen years of friendship, guidance, help and inspiration.

Steven N. Evans, Department of Statistics, University of California, Berkeley,
`evans@stat.berkeley.edu`

References

[1] J. Aitchison. *The statistical analysis of compositional data.* Chapman & Hall, London, 1986.

[2] Donald A. Dawson. Measure-valued Markov processes. In *École d'Été de Probabilités de Saint-Flour XXI—1991*, pages 1–260. Springer, Berlin, 1993.

[3] Bruce K. Driver. A primer on Riemannian geometry and stochastic analysis on path spaces. Available at `http://math.ucsd.edu/~driver`, 1995.

[4] Michel Émery. *Stochastic calculus in manifolds.* Springer-Verlag, Berlin, 1989. With an appendix by P.-A. Meyer.

[5] John H. Gillespie. *The causes of molecular evolution*, volume 2 of *Oxford series in ecology and evolution.* Oxford University Press, 1991.

[6] L. C. G. Rogers and David Williams. *Diffusions, Markov processes, and martingales. Vol. 2: Itô calculus.* John Wiley & Sons Inc., New York, 1987.

[7] Michael Sharpe. *General theory of Markov processes.* Academic Press Inc., Boston, MA, 1988.

[8] Daniel W. Stroock. *An introduction to the analysis of paths on a Riemannian manifold.* American Mathematical Society, Providence, RI, 2000.

[9] M. van den Berg and J. T. Lewis. Brownian motion on a hypersurface. *Bull. London Math. Soc.*, 17(2):144–150, 1985.

Investigating the Structure of Truncated Lévy-stable Laws

Anthony G. Pakes

Abstract

Truncations of stable laws have been proposed in the econophysics literature for modelling financial returns, often with imprecise definitions. This paper sharpens definitions of exponential truncations and attempts to expose underlying structure. Analytical comparisons are made with alternative models, leading to a tentative conclusion that the generalized hyperbolic family is more attractive for empirical work.

Keywords: Truncation; Lévy-stable laws

1 Introduction

Extensive empirical research shows that (log-)return data obtained from frequently sampled financial time series is not well fitted by a normal (Gaussian) law. Rather, the 'true' population law is more peaked around its median and it has fatter tails. Many analytically specified laws have been proposed and found to give a good fit to selected data sets. For example McDonald [22], Rydberg [31], and Voit [35, §5.3,5.4] are recent reviews representing the finance, statistics, and physics disciplines, respectively. In particular, Mandelbrot [23, E14,15] champions validity of non-normal stable laws. In fact, many return series exhibit tail behaviour which is intermediate to normal and non-normal stable behaviour. As a result, various more complicated models built from stable laws are found to mimic the stylized features of real data; see [31, 3].

The 'econophysics' school of modellers support use of so-called truncated Lévy (*i.e.*, stable) laws. See [7] for a general discussion of their use in finance, and [25] for pricing options. Let $g(x;\alpha)$ denote the density function of a stable law having index $\alpha \in (0,2)$ and symmetric about the origin, and let X be a random variable having this law. If $1 < \alpha < 2$ then $E(X) = 0$ and $var(X) = \infty$, but if $0 < \alpha \leq 1$ then neither the mean nor the variance can be defined. Econophysicists hold this to be unsatisfactory on the reasonable grounds that returns cannot be arbitrarily large in magnitude, and hence admissible models should possess finite moments of all orders. In general terms, the solution they propose is to use weighted densities

$$f(x;\alpha,w) = w(x)g(x;\alpha), \tag{1.1}$$

where $w(x) \geq 0$ is a weighting function satisfying $w(x) \to 0$ as $|x| \to \infty$, $\int f(x;\alpha,w)dx = 1$, and $\int |x|^r f(x;\alpha,w)dx < \infty$ for all $r > 0$.

This idea was introduced by Mantegna and Stanley [24] in the specific case $w(x) = c1_{[-l,l]}(x)$ where $0 < l < \infty$ is a truncation level and c is a normalization constant. Let S_n denote the sum of n independent copies of a random variable having this truncated stable law, let $f_n(x)$ denote its density function, and $v_l = E(S_1^2)$. The local version of the central limit theorem asserts that

$$\lim_{n\to\infty} \sqrt{n} f_n(0) = \frac{1}{\sqrt{2\pi v_l}}. \tag{1.2}$$

Mantegna and Stanley [24] provide simulations showing that the asymptotic regime (1.2) is approached more and more slowly as l increases. In addition, they found evidence of a quite long-lived pre-asymptotic regime during which $f_n(0)$ decays in proportion to $n^{-1/\alpha}$, the asymptotic behaviour for the parent stable law.

This modification of stable laws is analytically awkward and hence Koponen [19] recommends versions where $w(x) > 0$ for all x but decaying exponentially fast as $|x| \to \infty$. In fact, the precise nature of his definition is not clear. He asserts (for the symmetric case) that the truncated density function is

$$f(x) = c|x|^{-1-\alpha} e^{-\gamma|x|}, \tag{1.3}$$

where $\gamma > 0$ is an additional parameter and c is the normalizing constant. But clearly (1.3) does not define a density function since $\int_{[-\varepsilon,\varepsilon]} f(x)dx = \infty$ for any $\varepsilon > 0$. Koponen [19] mentions a lengthy calculation of a characteristic function (CF) whose symmetric version is

$$\sigma(t) = E\left(e^{itX}\right) = \exp\left[-A(t^2+\gamma^2)^{\alpha/2} \frac{\cos(\alpha \arctan(|t|/\gamma))}{\cos(\frac{1}{2}\alpha\pi)}\right]. \tag{1.4}$$

Paul and Baschnagel [29, p. 123] specify (1.3) holding in an asymptotic sense as $|x| \to \infty$, thus removing the singularity problem. They give a detailed derivation of (1.4) (see their Appendix D) where it is evident that the right-hand side of (1.3) is taken as the density of the Lévy measure of an infinitely divisible law. Consequently the nature of the law whose CF is (1.4) is unresolved.

Our aim here is to illumine this obscurity. We will distinguish three operations: (i) truncation as envisaged by Koponen [19], that is, exponential down-weighting a parent density function; exponential tilting, which involves multiplying a parent density by a decreasing exponential function (thus inflating the left-hand tail); and (iii) pruning an infinitely divisible (infdiv) law by truncating its Lévy measure. Pruning is implicit in Paul and Baschnagel's calculation. Briefly, it seems that truncation does not support a useful theory, whereas shrinking and tilting are almost equivalent, and they support a richer theory.

A key reason for considering truncated/pruned stable laws is to give a parametric family which exhibits a wider spectrum of tail behaviours than the stable laws, normal

and non-normal. Thus in §2 we review relevant results about infinite divisibility and convolution equivalent laws. In particular, Theorem 2.1 gives conditions ensuring the right-hand tail of the law is asymptotically proportional to the right-hand tail of its Lévy measure. This is a simple extension of known results for one-sided laws, and its proof is given in Pakes [27]. In §3 we define an exponential truncation operation on two-sided laws, observing that even though tail behaviour is in principal accessible, other structural properties such as determination of its moments present substantial analytical difficulties.

Section 4 reviews a tilting operation, familiar in other contexts, and relates it to the pruning of spectrally positive infinitely divisible laws. Particular application is made to extreme stable laws, and some limit distributions are obtained which illuminate the simulation results in Mantegna and Stanley [24]. In §5 we define the pruning of two-sided stable laws as the difference of independent tilted extreme stable random variables and Theorem 5.1 gives its characteristic function. Representation as either a tilted or truncated law is examined. We explore the representation of differences of tilted laws in terms of a truncated law, showing in particular that these pruned stable laws cannot be represented as a truncation of any two-sided stable law. In §6 we observe that pruned stable laws are generalized gamma convolutions, (GGC's) and hence their symmetric versions are normal-variance mixtures. This form of mixing is significant in financial modelling as a representation of stochastic volatility. Unfortunately, the mixing law appears to be quite complicated. Simpler representations as GGC's are found. These representations suggest comparisons with other laws which can be obtained from normal-variance mixing and tilting, and we look briefly at two special families, one being the generalized hyperbolic laws. Process and series representations of tilted and pruned stable laws are examined in §7. Here we give a self-contained and elementary account of random series representations of a broad class of infinitely divisible laws, and demonstrate that although there are many representations of tilted and pruned stable laws, finding one with simple explicit generating elements is problematic. Some final comments are given in §8, where we recommend the generalized hyperbolic family as being far better suited for empirical work than truncated or pruned stable laws.

2 Infinitely divisible laws

Our context will be the infinitely divisible (infdiv) laws, and there are several equivalent ways of defining this notion. We will agree that the random variable X with distribution function $F(x)$ has an infdiv law if its characteristic function (CF) $\phi(t) = E[e^{itX}]$ has the form $\phi(t) = \exp(-\psi(t))$ where the *characteristic exponent* is

$$\psi(t) = -Ait + \tfrac{1}{2}Vt^2 + \int_{|x|<1} \left[1 - e^{itx} + itx\right] \nu(dx) + \int_{|x|\geq 1} \left[1 - e^{itx}\right] \nu(dx), \qquad (2.1)$$

A is a real constant, $V \geq 0$, and ν is a measure on $(-\infty,\infty)$ satisfying $\nu\{0\} = 0$ and $\int(x^2 \wedge 1)\nu(dx) < \infty$, and called the Lévy measure. (We use the notation $\nu\mathcal{E}$ to denote

the measure assigned by ν to the set $\mathcal{E}$.) Thus $\mathcal{L}(X)$, the law of X, is comprised of three independent components, the constant A, a $\mathcal{N}(0,V)$ normal component, and a superposition of compound Poisson laws (Sato (1999) for example). Infdiv laws comprise the totality of laws which satisfy the partition property: For any integer $n \geq 1$, X can be written as a sum $\sum_{j=1}^{n} \varepsilon_{jn}$ of independent and identically distributed random variables (and clearly their law has the CF $(\phi(t))^{1/n}$)). The centering term itx in the first integral of (2.1) is often expressed in different ways, but this only affects the value of A. Infdiv laws are always unbounded in at least one direction and hence the Mantegna-Stanley [24] truncation always results in a law which is not infdiv.

We identify the important special case of spectrally positive infdiv laws (SPID laws), defined by the constraint $\nu(-\infty,0) = 0$. In this case the Laplace-Stieltjes transform (LST) $E(e^{-\theta X}) := L(\theta) = \exp(-\kappa(\theta))$ is finite ($\theta \geq 0$) where the *cumulant function* has the representation

$$\kappa(\theta) = A\theta - \tfrac{1}{2}V\theta^2 + \int_0^{1-} \left[1 - e^{-\theta x} - \theta x\right] \nu(dx) + \int_1^{\infty} \left[1 - e^{-\theta x}\right] \nu(dx). \tag{2.2}$$

To minimise algebraic details, we will always assume that $V = 0$. It is clear that the general infdiv law can be decomposed as $X = A + X_1 - X_2$, where X_1 and X_2 are independent SPID random variables with zero constant terms. In this sense SPID laws are fundamental.

There are two types of SPID law. Type 2 is defined by the condition $\int_0^1 x\nu(dx) = \infty$, in which case $\text{supp}(\mathcal{L}(X)) = \mathbb{R}$; X can assume any positive or negative value. But since $L(\theta) < \infty$ if $\theta > 0$, the left-hand tail $P(X < -x)$ ($x > 0$) decreases to zero faster than any exponentially decreasing function. Indeed, Ohkubo [26, p. 78] shows that $P(X < -x) = O(\exp(-x\log x))$ for large x. Thus $\mathcal{L}(X)$ is 'almost' one-sided.

Type 1 is defined by the condition $\int_0^1 x\nu(dx) < \infty$ in which case the cumulant function has the slightly simpler representation

$$\kappa(\theta) = B\theta + \int_0^{\infty} \left[1 - e^{-\theta x}\right] \nu(dx), \tag{2.3}$$

where $B = A - \int_0^{1-} x\nu(dx)$, and $\mathcal{L}(X)$ is one-sided with support $[B,\infty)$. This includes the fundamental compound Poisson case where $\rho = \nu[0,\infty) < \infty$. In this case $\rho^{-1}\nu[0,x]$ is a distribution function and we can write

$$X = B + \sum_{j=1}^{N} \eta_j,$$

where N has a Poisson(ρ) law and the η_j are independent with distribution function $\rho^{-1}\nu[0,x]$ and independent of N. We shall see below that the asymptotic behaviour of $P(X - A > x)$ is determined by the rate at which $\nu[x,\infty)$ tends to zero as $x \to \infty$; in other words, it is determined by the compound Poisson component of the infdiv law. It typically is the case that an infdiv or SPID law is such that it is not possible to explicitly

exhibit its distribution function, whereas its Lévy measure often has a simple form. (This is certainly true for stable laws which we later consider.) Thus, it is important to somehow relate the upper tail $P(X > x)$ to properties of $\nu(dx)$. Sato [33] discusses these matters for cases where there exists a constant $\gamma > 0$ such that the transform $\hat{\nu}(\theta) = \int_1^\infty e^{-tx}\nu(dx)$ converges in the open interval $(-\gamma, \infty)$ and diverges otherwise. Here we are concerned with cases where $\gamma \geq 0$ and $\hat{\nu}(-\gamma) < \infty$. The following concepts embrace this situation.

We begin the following definition, slightly extending Definition 1 in Cline [10]. Denote the survivor function of a distribution function $G(x)$ by $\overline{G}(x) := 1 - G(x)$.

Definition 2.1. A distribution function $G(\cdot)$ has an exponential tail with rate $\gamma \geq 0$, written $G(\cdot) \in \mathcal{L}_\gamma$, if

$$\lim_{x\to\infty} \frac{\overline{G}(x-y)}{\overline{G}(x)} = e^{\gamma y} \qquad (-\infty < y < \infty).$$

For each $y' > 0$ the limit holds uniformly for $y \leq y'$ if $\gamma > 0$, and uniformly in $[-y', y']$ if $\gamma = 0$. Speaking of an exponential tail with rate $\gamma = 0$ is somewhat contradictory, and we observe that, for our purposes, $\mathcal{L}_0$ is a very substantial class of long-tailed distribution functions in that $\lim_{x\to\infty} e^{\varepsilon x}\overline{G}(x) = \infty$ for each $\varepsilon > 0$.

We will see that stable and pruned stable laws belong to the following general class of laws. Denote the convolution of distribution functions G and H by $G * H$, and convolution powers by, for example, G^{*2}.

Definition 2.2. If $G \in \mathcal{L}_\gamma$ for some $\gamma \geq 0$, say that it is convolution equivalent, written $G \in \mathcal{S}_\gamma$, if

$$\lim_{x\to\infty} \frac{\overline{G^{*2}}(x)}{\overline{G}(x)} = 2M, \tag{2.4}$$

where $M < \infty$.

Bingham *et al.* [5] make some remarks about convolution equivalence on the line, and the unpublished report [37] develops some properties of two-sided convolution equivalence. Apart from these references, general theory for convolution equivalent distribution functions assumes that $G(0-) = 0$, that is, that the corresponding random variables are non-negative. In this case Cline [11, p. 355] shows that $M = M_G(\gamma) := \int e^{\gamma x} dG(x)$, the moment generating function of G. (Unrestricted integrals are taken over the real line.) Pakes [27] extend this to the two-sided case. For positive laws, $G(0-) = 0$, the boundary case $\gamma = 0$ usually is defined by (2.4) alone with the additional condition $M = 1$, giving the so-called subexponential class $\mathcal{S}$, which is a proper subset of $\mathcal{L}_0$. The subexponential class was introduced by Chistyakov [8] for estimating the long-term mean size of certain population processes, and it contains virtually

any long-tailed law occuring in financial modelling and other applications. We define $\mathcal{S}$ to comprise the laws satisfying (2.4) with $M = 1$. An important example is $\overline{G} \in \mathcal{R}_{-a}$, the class of regularly varying (at infinity) functions with index $-a$. Another is $\overline{G}(x) = const.\exp(-cx^a)$ $(x > 0)$ where $c > 0$ and $0 < a < 1$. In these cases $G \in \mathcal{S}$.

The definition (2.4) is equivalent to $\lim_{x\to\infty} \overline{G}^{*n}(x)/\overline{G}(x) = nM^{n-1}$ for some (and hence all) integers $n \geq 1$. Thus if $\gamma = 0$, the definition has the probabilistic meaning

$$\lim_{x\to\infty} \frac{P(Y_1 + \cdots + Y_n > x)}{P(\max(Y_1, \ldots, Y_n) > x)} = 1,$$

where the Y_j are independent with distribution function $G(\cdot)$. If $\gamma > 0$, then $\overline{G}(x) = e^{-\gamma x}\tau(x)$, where $\int_1^\infty \tau(x)dx < \infty$. Bingham *et al.* [5] thus use the term 'close to exponential' for members of $\cup_{\gamma>0}\mathcal{S}_\gamma$. Observe that exponential and gamma laws with scale parameter γ belong to $\mathcal{L}_\gamma$ but not to $\mathcal{S}_\gamma$. Cline [10] gives several criteria for membership of $\mathcal{S}_\gamma$ $(\gamma \geq 0)$. The following lemma, embracing the laws we consider here, is a direct consequence of his Corollary 2.

Lemma 2.3. Suppose that

$$\overline{G}(x) = x^{-\delta}L(x)e^{-\gamma x - cx^\omega},$$

where $\gamma, c \geq 0$, $\omega < 1$, $L(\cdot)$ is normalized slowly varying, and if $c = 0$ then either $\delta > 1$ or $\delta = 1$ and $\int_1^\infty (L(x)/x)dx < \infty$. Then $G \in \mathcal{S}_\gamma$.

Proof. Observe that if $c = 0$ then $M_G(\gamma) < \infty$ iff the conditions on δ hold. Write $\overline{G}(x) = \exp[-\xi(x)]$ and observe that

$$\xi'(x) = \gamma + c\omega x^{\omega-1} + \delta/x - \varepsilon(x)/x$$

where $\varepsilon(x) \to 0$ $(x \to \infty)$ is the index function of $L(\cdot)$; see [5, pp. 12-15]. The function $\xi_1(x) = \gamma x + cx^\omega + \delta \log x$ is concave and $x|\xi'(x) - \xi_1'(x)| \to 0$, thus fulfilling Cline's conditions. □

The following theorem relates the asymptotic behaviour of $\overline{F}(x)$ for an infdiv law and the distribution function

$$J(x) = \lambda^{-1}\nu(x,\infty)1_{[1,\infty)}(x),$$

where $\lambda = \nu(1,\infty)$, of its positive jump components exceeding unity. Its proof with other details and references are given in Pakes [27].

Theorem 2.1. Suppose that $\gamma \geq 0$ and F is an infdiv distribution function. The following assertions are equivalent:

(i) $J \in \mathcal{S}_\gamma$;
(ii) $J \in \mathcal{L}_\gamma$ and

$$\lim_{x\to\infty} \frac{\overline{F}(x)}{\nu(x,\infty)} = M_F(\gamma); \tag{2.5}$$

(iii) $F \in \mathcal{S}_\gamma$.

As an application of his results for two-sided convolution equivalence, Willekens [37] proved the equivalence of (i) and (ii) and that each implies (iii). In addition to the Tauberian converse, Pakes [27] shows Theorem 2.1 is essentially a consequence of the one-sided case. We emphasize that for our purposes the most significant part of Theorem 2.1 is the assertion that (i) implies (ii) and (iii). Observe also that reflection about the origin shows that if $\tilde{J}(x) = \nu[-x,-1)/\nu(-\infty,-1) \in \mathcal{S}_\gamma$ then

$$\lim_{x\to\infty} \frac{P(X<-x)}{\nu(-\infty,-x)} = M_F(-\gamma).$$

Consider the important case of the general stable law with index $\alpha \in (0,2)$, denoted by stable(α,b,p). This is defined by the absolutely continuous Lévy measure $\nu(dx) = n(x)dx$ where

$$n(x) = \begin{cases} \frac{\alpha b p}{\Gamma(2-\alpha)} x^{-1-\alpha} & \text{if } x>0, \\ \frac{\alpha b q}{\Gamma(2-\alpha)} |x|^{-1-\alpha} & \text{if } x<0, \end{cases}$$

where $b>0$ and $0 \le p = 1-q \le 1$. Then the CF is given by

$$\psi(t) = -Ait + \begin{cases} c|t|^\alpha \left(1 - i\beta \operatorname{sgn}(t) \tan(\tfrac{1}{2}\pi\alpha)\right) & \text{if } \alpha \neq 1, \\ c|t| \left(1 + i\beta \operatorname{sgn}(t) \tfrac{2}{\pi} \log(|t|)\right) & \text{if } \alpha = 1, \end{cases}$$

where

$$c = \begin{cases} \frac{b}{\alpha-1} \cos(\tfrac{1}{2}\pi\alpha) & \text{if } \alpha \neq 1, \\ \tfrac{1}{2} b\pi & \text{if } \alpha = 1, \end{cases}$$

and $\beta = p - q$. Note that many textbook renditions err in the sign attached to β; see [15]. The most error-proof derivation of this result is synthesizing it from the cumulant function of the spectrally positive version, $p = 1$. In this case

$$\kappa(\theta) = A\theta - \begin{cases} \frac{b}{\alpha-1}\theta^\alpha & \text{if } \alpha \neq 1, \\ b\theta \log\theta & \text{if } \alpha = 1, \end{cases}$$

obtained by integrating the relation $\kappa''(\theta) = \int_0^\infty e^{-\theta x} x^2 n(x)dx = -\alpha b \theta^{-2+\alpha}$. Some manipulation with complex algebra leads to the above CF.

Stable laws are subexponential ($\gamma = 0$), and hence Theorem 2.1 gives the asymptotic estimate

$$P(X > x) \sim \nu(x, \infty) = \frac{bp}{\Gamma(2-\alpha)} x^{-\alpha} \qquad (x \to \infty). \tag{2.6}$$

For later reference observe that the parameter b functions as a scale constant, and that it affects only the constant multiplier in (2.6). In particular, all one-dimensional laws of the embedding process are tail equivalent in the sense that

$$\lim_{x\to\infty} \frac{P(\Lambda(\tau_1) > x)}{P(\Lambda(\tau_2) > x} = \frac{\tau_1}{\tau_2}.$$

3 Truncating the density

Suppose a random variable X has a density $f(x) > 0$ for all real x, and with $\gamma > 0$ let X_γ denote the random variable having the symmetrically truncated density

$$f_\gamma(x) = Ke^{-\gamma|x|} f(x),$$

where K is a normalizing constant. In principle, this specifies an explicit density function with easily determined tail behaviour. Thus if $f(\cdot)$ is a stable density, it follows from (2.6) that $P(X_\gamma > x) \sim const.x^{-1-\alpha}e^{-\gamma x}$. However, it seems difficult to gain further structural information, such as moments or the CF, or to determine if $\mathcal{L}(X_\gamma)$ is infdiv, or to give a probabilistic characterization of this construction. In addition, it is hard to see how truncation could be put into a process framework. Specifically, if $X = \Lambda(1)$ where $(\Lambda(s))$ is a Lévy process then is there a process $(\Lambda_\gamma(s))$ such that $\Lambda_\gamma(1)$ has density $f_\gamma(x)$ and properties which are relevant to the modelling context?

Recalling that $\phi(t)$ is the CF of $f(\cdot)$ and observing that the Fourier transform of the kernel $e^{-\gamma|x|}$ equals $\gamma/(\gamma^2 + t^2)$, we can write the CF of the truncated density as the convolution (or Poisson integral)

$$\phi_\gamma(t) = \gamma \int \frac{\phi(u)}{\gamma^2 + (t-u)^2} du.$$

This however seems to offer little insight into the nature of $\mathcal{L}(X_\gamma)$, even for quite specific cases such as symmetric stable laws.

One exception is the Cauchy density $f(x) = c/\pi(c^2 + x^2)$. In this case, reference to a table of integrals [14] shows that the truncated law has the LST

$$E\left(e^{-\theta X_\gamma}\right) = \frac{I(\gamma+\theta) + I(\gamma-\theta)}{2I(\gamma)} \qquad (|\theta| < \gamma)$$

where

$$I(\theta) = \int_0^\infty e^{-\theta x} \frac{c}{c^2 + x^2} dx = \mathrm{ci}(c\theta)\sin(c\theta) - \mathrm{si}(c\theta)\cos(c\theta)$$

and

$$\mathrm{ci}(x) = -\int_x^\infty \frac{\cos v}{v} dv \qquad \& \qquad \mathrm{si}(x) - \int_x^\infty \frac{\sin v}{v} dv$$

are cosine and sine integrals, respectively. In addition, $E(X_\gamma) = 0$ (by symmetry) and

$$var(X_\gamma) = \frac{c}{\gamma l(\gamma)} - c^2 \sim \frac{2c}{\pi\gamma} \qquad (\gamma \downarrow 0).$$

We conclude that truncation is not a fruitful concept.

4 Tilting and pruning

In this section, we recall an asymmetric exponential weighting operation which has been much studied in other contexts. So we let X be a random variable with arbitrary distribution function $F(\cdot)$ satisfying $L(\theta) = \int e^{-\theta x} dF(x) < \infty$ for all $\theta \geq 0$. Fix a constant $\gamma > 0$ and define the law of a random variable X_γ, the (exponential) γ-tilt of $\mathcal{L}(X)$, which has the distribution function

$$F(x;\gamma) = \frac{\int_{-\infty}^x e^{-\gamma x} dF(x)}{L(\gamma)} \qquad (x \in \mathbb{R}). \tag{4.1}$$

The LST of X_γ is

$$L(\theta;\gamma) = \frac{L(\theta+\gamma)}{L(\gamma)}.$$

The family of laws obtained by varying γ through the largest open interval such that $L(\gamma) < \infty$ is called the natural exponential family (NEF) generated by $\mathcal{L}(X)$. See Seshadri [34] for these matters, but note that we adopt an opposite sign convention for the exponent parameter. Tilting is used for obtaining asymptotic expansions for sums of random variables, and in the theory of random walk. See Feller [13] for the latter application, where he uses the term 'associated distribution'.

If X has a SPID law, as defined by (2.2), we can define an operation of *pruning* whereby ν is replaced by the truncation

$$\nu_\gamma(dx) = \tau e^{-\gamma x} \nu(dx), \tag{4.2}$$

where $\tau > 0$ is a constant. Some manipulation shows that the cumulant function κ is transformed into

$$\kappa(\theta;\gamma,\tau) = \tau[\kappa(\theta+\gamma) - \kappa(\gamma) + B_\gamma \theta],$$

where

$$B_\gamma = \int_0^{1-} \left(1 - e^{-\gamma x}\right) x\nu(dx).$$

A random variable $\hat{X}_{\gamma,\tau}$ with this pruned law has the LST is

$$L(\theta;\gamma,\tau) := E\left(e^{-\theta \hat{X}_{\gamma,\tau}}\right) = \left[\frac{L(\theta+\gamma)}{L(\gamma)}\right]^\tau e^{-\tau B_\gamma \theta}. \tag{4.3}$$

This shows that the exponential tilt of $\mathcal{L}(X)$ is equivalent to weighting ν according to (4.2) with $\tau = 1$, together with a shift B_γ to the left of the pruned law. For the reverse direction, let $(\Lambda(\tau) : \tau \geq 0)$ denote the spectrally positive Lévy process with cumulant function $\kappa(\theta)$. Then the transformation (4.2) maps $\mathcal{L}(X)$ to $\mathcal{L}(\Lambda_\gamma(\tau) + \tau B_\gamma)$, the exponential tilt of the process at time τ with a translation τB_γ to the right. In particular, if $\tau = 1$ then, in obvious notation,

$$\hat{X}_\gamma \stackrel{L}{=} X_\gamma + B_\gamma.$$

In the Type 1 case, we can apply the tilting operation to (2.3) to obtain the same form with B unchanged (*i.e.*, $B_\gamma = 0$) and Lévy measure (4.2) with $\tau = 1$. In any event, we see that the non-symmetric tilting operation thins the right-hand tail of $\mathcal{L}(X)$ in the manner recommended by [19, 24]. In contrast, the left-hand tail, if it is non-trivial, is inflated by exponential tilting but it still decreases faster than exponential. In addition, if $\nu(\cdot)$ has a density $n(\cdot)$, that is $\nu(0,x] = \int_0^x n(y)dy$, then $F(\cdot)$ has a density function $f(x)$, and $\mathcal{L}(X_\gamma)$ has a density function and an absolutely continuous Lévy measure given respectively by

$$f(x;\gamma) = e^{-\gamma x} f(x)/L(\gamma) \qquad \& \qquad \nu_\gamma(dx) = e^{-\gamma x} n(x)dx.$$

Finally, any relation connecting the tail behaviours of $F(x)$ and $\nu(dx)$ translates to a parallel relation between $F(x;\gamma)$ and $\nu_\gamma(dx)$. In general, it is analytically more convenient to work with tilting rather than pruning.

We shall now consider the effect of these transformations on the spectrally positive stable law stable$(\alpha, b, 1)$. First, ignoring the change of location, note that the effect of the parameter τ in (4.2) is simply to multiply the parameter b. Consequently, for our present considerations, we lose no generality in setting $\tau = 1$, and we do this until further notice. Tilting shrinks the Lévy measure to $\nu_\gamma(dx) = [\alpha b/\Gamma(2-\alpha)]e^{-\gamma x}x^{-1-\alpha}dx$, yielding the cumulant function

$$\kappa(\theta;\gamma) = \kappa(\theta+\gamma) - \kappa(\gamma) = A\theta - \begin{cases} \frac{b}{\alpha-1}\left((\theta+\gamma)^\alpha - \gamma^\alpha\right) & \text{if } \alpha \neq 1, \\ b[(\theta+\gamma)\log(\theta+\gamma) - \gamma\log\gamma] & \text{if } \alpha = 1. \end{cases} \tag{4.4}$$

We denote this tilted law by t-stable$(\alpha, b; \gamma)$, where the notation is understood to imply that the asymmetry parameter $\beta = 1$. The case $\alpha < 1$ defines the so-called Hougaard laws [17], used to model lifetime distributions in a heterogeneous population. Seshadri [34] mentions some earlier formulations. The special case $\alpha = \frac{1}{2}$ gives the inverse-Gaussian law whose density function in the case $A = 0$ is

$$f(x;\gamma) = \frac{b}{\sqrt{\pi x^3}} \exp\left[2b\sqrt{\gamma} - \left(\frac{b^2}{x} + \gamma x\right)\right], \qquad (x > 0).$$

The gamma laws occur as the limit of the t-stable$(\alpha, b/\alpha; \gamma)$ as $\alpha \to 0$. In no other case is it possible to express $f(x;\gamma)$ in terms of elementary functions. This is a consequence of the corresponding intractability of stable densities. However Hoffmann-Jorgensen

[16] gives expressions for stable densities in terms of an incomplete hypergeometric function, valid in all cases except $\alpha = 1$ and $\beta \neq 0$. Williams [38] gives an elegant demonstration that the stable density for the case $\alpha = 1/3$ and $\beta = 1$ has a simple form in terms of a Bessel function. See Zolotarev [39, pp. 155–158] for some representations in terms of Whittaker functions.

The mean and variance of X_γ are given respectively by

$$\mu_\gamma = A - \begin{cases} \frac{\alpha b}{\alpha - 1}\gamma^{\alpha-1}, \\ b(1+\log\gamma), \end{cases} \qquad \& \quad \sigma_\gamma^2 = \begin{cases} \alpha b \gamma^{\alpha-2} & \text{if } \alpha \neq 1, \\ \frac{b}{\gamma} & \text{if } \alpha = 1. \end{cases}$$

These quantities are finite, and indeed all moments are finite. Observing that

$$x^{1+\alpha}\int_x^\infty y^{-1-\alpha}e^{-\gamma y}dy = e^{-\gamma x}\int_0^\infty \left(\frac{x}{x+y}\right)^{1+\alpha} e^{-\gamma y}dy \sim \gamma^{-1}e^{-\gamma x} \qquad (x \to \infty),$$

we see that Theorem 2.1, or the tilting construction itself, implies that

$$P(X_\gamma > x) \sim \frac{\alpha b}{\gamma L(\gamma)\Gamma(2-\alpha)} e^{-\gamma x}x^{-1-\alpha} \qquad (x \to \infty). \tag{4.5}$$

Note that in the case of heavy tilting, $\gamma \gg 1$, $\mu_\gamma \approx 0$ if $\alpha < 1$ and $\mu_\gamma \approx -\infty$ if $\alpha \geq 1$, and $\sigma_\gamma \approx 0$ in both cases. It is easy to show that $(X_\gamma - \mu_\gamma)/\sigma_\gamma$ is approximately standard normal when γ is large, a regime which is unlikely to be relevant for financial applications.

As mentioned in §1, Mantegna and Stanley [24], on the basis of simulations, identify two limit regimes as n increases for the sum S_n of n identical copies of random variables having their version of the truncated Lévy law. In addition, they assert a value of n, denoted by $n_\times$, which is claimed to characterize the transition from stable limit behaviour to ultimate normal limit behaviour. The basis for this is an assertion that the density function of S_n, evaluated at the origin, has a stable form when n is small and a normal form when n is large. (The precise nature of these forms result from local limit theorems.) The critical value $n_\times$ is obtained by equating the two density values.

The following considerations identify three limit regimes for the tilted spectrally positive stable law. The cumulant function of S_n is $n\kappa(\theta;\gamma)$ and since, from (4.4), the factor n merely inflates the paramter b we can, and shall, set $n = 1$ and let $b \to \infty$. We will see that the limit behaviour of $\mathcal{L}(X_\gamma)$ is determined by a critical parameter $\xi = \alpha b\gamma^\alpha$. There is some tension in the literature on modelling financial returns between whether they exhibit algebraically decreasing (heavy) tails or whether there also is a truncation factor $e^{-\gamma x}$ present (called semi-heavy tails by some). If this factor is present, then it may be that $\gamma \ll 1$: see §8 for further remarks. In such a case we can envisage that even with b large, there are three possible regimes, $\xi \approx 0$, $\xi = O(1)$, and $\xi \gg 1$, the last being attained in the limit $b \to \infty$. The following theorem deals with the first two possibilities. The proof is omitted since it involves only a straightforward manipulation of $\kappa(\theta b^{-1/\alpha};\gamma)$.

Theorem 4.1. Let $b \to \infty$ and $\xi \to \alpha\zeta$ where $0 \leq \zeta < \infty$. If $\alpha \neq 1$ then

$$V_b(\alpha) := b^{-1/\alpha}(X_\gamma - A) \xrightarrow{L} \text{t-stable}(\alpha, 1; \zeta^{1/\alpha}).$$

If $\alpha = 1$ then

$$V_b(1) := b^{-1}(X_\gamma - A - \log b) \xrightarrow{L} \text{t-stable}(1, 1; \zeta).$$

The proof shows in the case $\alpha = 1$ that the limit actually is an identity in law if $\zeta \equiv \xi/\alpha$. (We define $\zeta \log \zeta := 0$ if $\zeta = 0$.) Theorem 4.1 asserts that if b is large but γ is so small that $\xi \ll 1$ then $\mathcal{L}(V_b)$ is close to stable. As b grows further so that ξ is moderate, then $\mathcal{L}(V_b)$ retains the tilted stable form.

As b becomes larger still the critical parameter ξ becomes large. A straightforward application of the binomial theorem to the cumulant function of the normed variable $Z_b := (X_\gamma - \mu_\gamma)/\sigma_\gamma$ yields the expansion

$$\log E\left(e^{-\theta Z_b}\right) = \tfrac{1}{2}\theta^2 + \sum_{j=3}^{\infty} a_j(-\theta)^j \xi^{-\frac{1}{2}j+1}, \tag{4.6}$$

where

$$a_j = \frac{\alpha\Gamma(j-\alpha)}{j!\Gamma(2-\alpha)}.$$

This expansion is valid for $0 < \alpha < 2$. The following result characterizing the third (limiting) regime follows immediately.

Theorem 4.2. If $\xi \to \infty$ as $b \to \infty$ then $Z_b \xrightarrow{L} \mathcal{N}(0,1)$, the standard normal law.

Observe that the norming in Theorem 4.1 when $\zeta > 0$ is equivalent to that used for Theorem 4.2 since $\sigma_\gamma \sim \sqrt{\alpha\zeta}(b/\zeta)^{1/\alpha}$ and, if $\alpha \neq 1$, $(\mu_\gamma - A)/\sigma_\gamma \to (\alpha - 1)^{-1}\sqrt{\alpha\zeta}$. It is not at all clear how one might quantify the transition from one regime to the next.

The expansion (4.6) makes it clear that the normal limit is approached only after $\sqrt{\xi}$ becomes large. Indeed, this expansion can be inverted using the Fourier methodology described by Feller [13, Chapter XVI]. If $g_b(x)$ denotes the density function of Z_b and $\varphi(x)$ is the standard normal density function, then

$$g_b(x) - \varphi(x) = \varphi(x) \sum_{j=3}^{r} \xi^{-\frac{1}{2}j+1} P_j(x) = o(\xi^{-\frac{1}{2}r+1}),$$

where $P_j(x)$ is a polynomial of degree j which is independent of ξ. The case $r \leq 5$ gives the approximation

$$g_b(x) - \varphi(x) = \tfrac{1}{2}\varphi(x) \sum_{j=3}^{r} a_j \xi^{-\frac{1}{2}j+1} H_j(x) + O(\xi^{-\frac{1}{2}(r-1)}),$$

where

$$H_j(x) = (-1)^j e^{\frac{1}{2}x^2} \frac{d^j}{dx^j} e^{-\frac{1}{2}x^2}$$

is a (version of a) Hermite polynomial. Thus $H_3(x) = x^3 - 3x$, $H_4(x) = x^4 - 6x^2 + 3$ and $H_5(x) = x^5 - 10x^3 + 15x$.

5 Pruning two-sided stable laws

The construction implicit in the calculation of Paul and Baschnagel [29] can be expressed in general terms as follows. Let $X(j)$ $(j = 1,2)$ be independent random variables having a SPID law with Lévy measure ν_j, constant term A_j, and LST $L_j(\theta)$. Next, let $\hat{X}_\gamma(j)$ denote a random variable with the pruned law obtained from (4.2), that is, with Lévy measure $\nu_{j,\gamma}(dx) = \tau_j e^{-\gamma x}\nu_j(dx)$. Thus the scaling constant, but not the shrinkage parameter γ, may depend on j. Alteration of details below allows relaxation of this restriction. Then $X = \hat{X}_\gamma(1) - \hat{X}_\gamma(2)$ has a two-sided infdiv law, and from (4.3) its mgf is

$$M_F(\theta) = \left(\frac{L_1(\gamma-\theta)}{L_1(\gamma)}\right)^{\tau_1}\left(\frac{L_2(\gamma+\theta)}{L_2(\gamma)}\right)^{\tau_2} e^{\theta(\tau_1 B_{1,\gamma} - \tau_2 B_{2,\gamma})},$$

which is finite in an interval containing $[-\gamma,\gamma]$. Equivalently, we can define $X = X_\gamma(1) - X_\gamma(2)$, in which case the exponential factor is absent.

Assume X is centered so that $\tau_1(A_1 + B_{1,\gamma}) - \tau_2(A_2 + B_{2,\gamma}) = 0$. If

$$J_1(x) = [\nu_{1,\gamma}(x,\infty)/\nu_{1,\gamma}(1,\infty)]1_{[1,\infty)}(x) \in \mathcal{S}_\gamma$$

then Theorem 2.1 implies that

$$\lim_{x\to\infty}\frac{P(X>x)}{\nu_{1,\gamma}(x,\infty)} = (L_1(\gamma))^{-\tau_1}\left(\frac{L_2(2\gamma)}{L_2(\gamma)}\right)^{\tau_2}, \tag{5.1}$$

with a similar relation for $P(X < -x)$.

In the sequel, we will work with the tilted rather than the pruned version. So we let $X(j) - A_j \sim \text{stable}(\alpha, b_j, 1)$ $(j = 1,2)$ where $b_j > 0$, A_j is real, $\gamma > 0$. The CF of $X = X_\gamma(1) - X_\gamma(2)$ can easily be computed from (4.4) by substituting $\theta = -it$ and converting to the polar representation $\gamma - it = \sqrt{\gamma^2+t^2}e^{-i\omega\,\mathrm{sgn}(t)}$ where $\omega = \arctan(|t|/\gamma)$. Algebraic manipulation yields the following result, agreeing with Paul and Baschnagel [29], and with Koponen [19], apart from scaling of some of the parameters.

Theorem 5.1. The CF of the pruned law $\mathcal{L}(X)$, where $X = X_\gamma(1) - X_\gamma(2)$, is $\exp(-\psi(t))$ where

$$\psi(t) - Ait = \begin{cases} -\frac{b}{\alpha-1}\left\{(\gamma^2+t^2)^{\alpha/2}[\cos(\alpha\omega) - i\beta\,\mathrm{sgn}(t)\sin(\alpha\omega)] - \gamma^\alpha\right\} & \text{if } \alpha \neq 1 \\ -b\left\{(\gamma^2+t^2)^{\frac{1}{2}}[\tfrac{1}{2}\log(\gamma^2+t^2)\cos\omega - \omega\sin\omega - i\beta\,\mathrm{sgn}(t)(\tfrac{1}{2}\log(\gamma^2+t^2)\sin\omega + \omega\cos\omega)] - \gamma\log\gamma\right\} & \text{if } \alpha = 1, \end{cases}$$

where

$$A = A_1 - A_2, \qquad b = \tau_1 b_1 + \tau_2 b_2, \quad \& \quad \beta = \frac{\tau_1 b_1 - \tau_2 b_2}{\tau_1 b_1 + \tau_2 b_2}.$$

We shall, without any real loss in generality, take $\tau_1 = \tau_2 = 1$, and let p-stable$(\alpha, b, \beta; \gamma)$ denote the resulting law.

Moments of $\mathcal{L}(X)$ and its asymptotic behaviour for large b can be inferred from the results in the previous section. For later reference we record tail estimates, assuming the location parameter $A = 0$:

$$\bar{F}(x) \sim \frac{B_+}{\gamma} x^{-1-\alpha} e^{-\gamma x} \quad \& \quad F(-x) \sim \frac{B_-}{\gamma} x^{-1-\alpha} e^{-\gamma x} \quad (x \to \infty) \tag{5.2}$$

where

$$B_+ = \frac{\alpha b p}{\Gamma(2-\alpha)} \frac{L_2(2\gamma)}{L_1(\gamma) L_2(\gamma)} \quad \& \quad B_- = \frac{\alpha b q}{\Gamma(2-\alpha)} \frac{L_1(2\gamma)}{L_1(\gamma) L_2(\gamma)}.$$

The following considerations will make it clear that $\mathcal{L}(X)$ cannot be realized as a two-sided exponential truncation (1.1) of a stable law.

We can gain some appreciation of the structure of $\mathcal{L}(X)$ as follows. Suppose for now that $X(j)$ $(j = 1,2)$ are independent with density functions $f_j(x)$ positive at least in $(0,\infty)$ and that $L_j(\theta) = \int e^{-\theta x} f_j(x) dx < \infty$ if $0 \le \theta \le 2\gamma$. Then the density of $X = X_\gamma(1) - X_\gamma(2)$ is

$$f(x) = \frac{e^{-\gamma x}}{L_1(\gamma) L_2(\gamma)} \int f_1(x+y) f_2(y) e^{-2\gamma y} dy. \tag{5.3}$$

Observe that

$$\int e^{-\theta x} \left(\int f_1(x+y) f_2(y) e^{-2\gamma y} dy \right) dx = L_1(\theta) \int f_2(y) e^{-y(2\gamma-\theta)} dy = L_1(\theta) L_2(2\gamma - \theta),$$

which is finite if $0 \le \theta \le 2\gamma$. Hence,

$$g(x) := \frac{\int f_1(x+y) f_2(y) e^{-2\gamma y} dy}{L_2(2\gamma)} \qquad (-\infty < x < \infty) \tag{5.4}$$

is a density function positive on the real line and with (bilateral) LST

$$L_G(\theta) = L_1(\theta) L_2(2\gamma - \theta) / L_2(2\gamma).$$

This is the LST of a random variable $U := X(1) - X_{2\gamma}(2)$, which clearly is infdiv if the $X(j)$ are SPID. In particular, if $X(j)$ has a stable$(\alpha, b_j, 1)$ law then U has the Lévy density

$$n_G(x) = \begin{cases} \frac{\alpha b_1}{\Gamma(2-\alpha)} x^{-1-\alpha} & \text{if } x > 0, \\ \frac{\alpha b_2}{\Gamma(2-\alpha)} |x|^{-1-\alpha} e^{-2|x|} & \text{if } x < 0, \end{cases}$$

showing that $\mathcal{L}(U)$ is asymmetric with tail probabilities $P(U > x) = O(x^{-\alpha})$ and $P(U < -x) = O(x^{-1-\alpha} e^{-2\gamma x})$.

Returning to the general case, we now can interpret (5.3) as specifying the exponential tilt $\mathcal{L}(X) = \mathcal{L}(U_\gamma)$, noting that $\mathcal{L}(U)$ depends on γ. This interpretation can be

extended to show that $\mathcal{L}(X)$ can indeed be represented as an exponential truncation of a two-sided law. Denoting the integral in (5.3) by $I(x)$, observe that

$$K_1 := \int_0^\infty I(x)dx = \int \bar{F}_1(y)f_2(y)e^{-2\gamma y}dy = L_2(2\gamma)P(X(1) > X_{2\gamma}(2)).$$

Similarly

$$K_2 := \int_{-\infty}^0 e^{-2\gamma x}I(x)dx = L_1(2\gamma)P(X(2) > X_{2\gamma}(1)).$$

Thus $h_+(x;\gamma) := (I(x)/K_1)1_{(0,\infty)}(x)$ is the conditional density of $X(1) - X_{2\gamma}(2)$, given this difference is positive. Similarly $h_-(x;\gamma) := (e^{-2\gamma x}I(x)/K_2)1_{(-\infty,0)}(x)$ is the conditional density of $X_{2\gamma}(1) - X(2)$, given this difference is negative. Now define the family of two-sided densities $h_r(x;\gamma) = rh_+(x;\gamma) + (1-r)h_-(x;\gamma)$, where $0 < r < 1$. For each such r let $m_1(r) = (K_1/rL_1(\gamma)L_2(\gamma))$ and $m_2(r) = (K_2/(1-r)L_1(\gamma)L_2(\gamma))$. Then we have the truncation (*c.f.* (1.1))

$$f(x) = \begin{cases} m_1(r)e^{-\gamma x}h_r(x;\gamma) & \text{if } x > 0, \\ m_2(r)e^{\gamma x}h_r(x;\gamma) & \text{if } x < 0, \end{cases}$$

thus representing $f(x)$ somewhat in the manner (apparently) envisaged by Koponen [19]. We emphasize that the law we are truncating here depends on the parameter γ, and it is obvious that this construction gives the only possible truncation representation. Observe however that in the stable case (5.2) implies that $f(x) \sim B_+x^{-1-\alpha}e^{-\gamma x}$ $(x \to \infty)$ and hence that $h_r(x;\gamma) \sim const.|x|^{-1-\alpha}$ as $|x| \to \infty$, with a different constant for each tail. Hence the truncated density has tails which decay at the same algebraic rate as a stable$(\alpha,\cdot,\beta)$ with $|\beta| < 1$. The nature of these laws is an open question, for example, it is not clear whether or not they are infdiv.

6 Normal-variance mixtures

Following Bondesson [6], let $\mathcal{T}_e$ denote the class of extended generalized gamma convolutions, that is, the closure of laws obtained as finite linear combinations of independent gamma distributed random variables. The subclass $\mathcal{T}$ of generalized gamma convolutions (GGC's) is generated from the linear combinations having positive coefficients. Members of $\mathcal{T}_e$ are infdiv and absolutely continuous. Moreover the symmetric members of $\mathcal{T}_e$ are normal-variance mixtures. More specifically, if $\mathcal{L}(X) \in \mathcal{T}_e$ and it is symmetric, then $X \stackrel{L}{=} Z\sqrt{Y}$ where $\mathcal{L}(Z) = \mathcal{N}(0,2)$, $\mathcal{L}(Y) \in \mathcal{T}$ and Y and Z are independent. The parametrization for $\mathcal{L}(Z)$ is chosen so that

$$M_F(\theta) = M_G(\theta^2), \tag{6.1}$$

where $G(\cdot)$ is the distribution function of Y, and we assume the mgf's are finite. Normal-variance mixing is important in financial modelling as a way of modelling stochastic

volatility. More precisely, if $(Y(\tau) : \tau \geq 0)$ is the Lévy process with $Y(1) \stackrel{L}{=} Y$ and $(B(\tau) : \tau \geq 0)$ is a Brownian motion process with $B(1) \stackrel{L}{=} Z$ then the *subordinated* process $\Lambda(\tau) := B(Y(\tau))$ is an embedding Lévy process, $\Lambda(1) \stackrel{L}{=} X$. Many marginal laws used for financial modelling can be represented in this way. See [18] for a catalogue and references. The following considerations show that pruning is accomodated by this framework.

The Laplace transform relation

$$x^{-1-\alpha}e^{-\gamma x} = \frac{1}{(\Gamma(1+\alpha))}\int_0^\infty [(v-\gamma)^+]^\alpha e^{-xv}dv$$

implies that t-stable$(\alpha, b; \gamma) \in \mathcal{T}$ and p-stable$(\alpha, b, \beta; \gamma) \in \mathcal{T}_e$; see Bondesson [6, pp. 30,107]. In particular, p-stable$(\alpha, b, 0; \gamma)$ is a normal-variance mixture, a fact which is not evident from inspection of its CF in Theorem 5.1 or its cumulant function

$$\kappa_s(\theta; \gamma) = -\frac{b}{2(\alpha-1)}[(\gamma+\theta)^\alpha + (\gamma-\theta)^\alpha - 2\gamma^\alpha], \tag{6.2}$$

finite if $|\theta| < \gamma$. The following results will show that the mixing law arises from tilting another positive law, and that this more basic law has a complicated form.

Suppose that $\mathcal{L}(X)$ =p-stable$(\alpha, b, 0; \gamma)$ and denote its Lévy density by $n(x)$. Hence, $n(x) = (\alpha b/2\Gamma(2-\alpha))|x|^{-1-\alpha}e^{-\gamma|x|}$ (all real x). As above, $X \stackrel{L}{=} Z\sqrt{Y}$, and $m(x)$ will denote the Lévy density of the mixing law $\mathcal{L}(Y)$. The general relation (6.1) can be expressed for infdiv laws as

$$\log M_F(\theta) = \int_0^\infty \left(e^{\theta^2 y} - 1\right) m(y)dy. \tag{6.3}$$

We use the following general result relating the Lévy densities of a normal-variance mixture.

Lemma 6.1. Suppose $F(x)$ is an absolutely continuous and symmetric infdiv law and that it is a normal-variance mixture as specified by (6.1). Then the corresponding Lévy densities are related by

$$n(x) = \frac{1}{2\sqrt{\pi}}\int_0^\infty e^{-x^2/4y} m(y)\frac{dy}{\sqrt{y}},$$

or

$$4\sqrt{\pi}n(\sqrt{s}) = \int_0^\infty e^{-sv} m(1/4v)v^{-3/2}dv. \tag{6.4}$$

Proof. The right-hand side of (6.3) has the representation $\int_0^\infty E\left[e^{\theta Z\sqrt{y}}-1\right]m(y)dy$. Equating this expression to $\kappa(-\theta)$, as obtained from (2.3) with $A=V=0$, and differentiating twice with respect to θ yields the identity

$$\int e^{\theta x}x^2 n(x)dx = E\left[Z^2\int_0^\infty e^{\theta Z\sqrt{y}}ym(y)dy\right] = E\left[Z^2\int_0^\infty\left(\int e^{\theta x}\delta(x-Z\sqrt{y})dx\right)ym(y)dy\right],$$

where $\delta(\cdot)$ is the Dirac delta-function. Interchanging the integrals on the right-hand side and inverting the bilateral Laplace transforms yields

$$x^2 n(x) = E\left[Z^2\int_0^\infty \delta(x-Z\sqrt{y})ym(y)dy\right].$$

Evaluating the expectation and cancelling the x^2 factor common to both sides leads to (6.4). □

Lemma 6.2. If $\hat{\eta}(\theta)$ is the Laplace transform of the function $\eta(x)$, then $\hat{\eta}(\sqrt{s})=\hat{h}(s)$ where

$$h(v)=\frac{1}{2\sqrt{\pi v^3}}\int_0^\infty ue^{-u^2/4v}\eta(u)du.$$

Proof. Simply observe that $\hat{h}(s)=\int_0^\infty e^{-u\sqrt{s}}\eta(u)du$ and that the exponential factor in the integrand is the Laplace transform of the stable$(\frac{1}{2},\frac{1}{2}u,1)$ law. □

Theorem 6.1. The Lévy density of the mixing law $\mathcal{L}(Y)$ for the p-stable$(\alpha,b,0;\gamma)$ is

$$m(x)=\frac{2\alpha b}{\Gamma(\alpha)\Gamma(1-\alpha)x}\int_0^\infty u^{\alpha-1}e^{-x(u+\gamma)^2}du=\frac{2^{1-\alpha}\alpha b}{\Gamma(1-\alpha)}\cdot x^{-\frac{1}{2}\alpha-1}e^{-\gamma^2 x}\Psi(\tfrac{1}{2}\alpha,\tfrac{1}{2};\gamma^2 x), \tag{6.5}$$

where Ψ is the Kummer (confluent hypergeometric) function of the second kind.

Proof. The left-hand side of (6.4) has the form $\hat{\eta}(\sqrt{s})$ where

$$\eta(u)=\frac{4\sqrt{\pi}\alpha b}{\Gamma(2-\alpha)}\cdot\frac{[(u-\gamma)^+]^\alpha}{\Gamma(1+\alpha)},$$

and hence Lemma 6.2 leads to the evaluation

$$h(v)=\frac{2bv^{-3/2}}{\Gamma(\alpha)\Gamma(2-\alpha)}I(1/v)\qquad\&\qquad I(x)=\int_0^\infty(u+\gamma)u^\alpha e^{-x(u+\gamma)^2}du.$$

But $\hat{h}(s)$ must equal the right-hand side of (6.4), that is, $h(v)=v^{-3/2}m(1/4v)$. Integrating $I(x)$ by parts leads to the integral representation in (6.5). The final form comes

from a substitution in the identity $\int_0^\infty u^{\alpha-1}e^{-u^2-2uz}du = \Gamma(\alpha)H_{-\alpha}(z)$, where the Hermite function $H_{-\alpha}(z) = 2^{-\alpha}\Psi(\frac{1}{2}\alpha, \frac{1}{2}; z^2)$ [20, pp. 285,290]. □

Expanding the exponent in (6.5) yields $m(x) = e^{-\gamma^2 x}\ell(x)$ where

$$\ell(x) = \frac{2\alpha K}{x}\int_0^\infty u^{\alpha-1}e^{-x(u^2+2\gamma u)}du \qquad \& \qquad K = \frac{b}{\Gamma(\alpha)\Gamma(2-\alpha)}. \tag{6.6}$$

Lemma 6.3. The mixing law is a tilted law, $Y \stackrel{L}{=} W_{\gamma^2}$ where $\mathcal{L}(W)$ is a positive infdiv law with Lévy density $\ell(x)$. Moreover, $\int_0^1 \ell(x)dx = \infty$, and

$$\ell(x) \sim \alpha K\Gamma(\tfrac{1}{2}\alpha)x^{-1-\alpha/2} \ (x \to 0) \qquad \& \qquad \ell(x) \sim \frac{\alpha b}{2^{\alpha-1}\gamma^\alpha\Gamma(2-\alpha)}x^{-1-\alpha} \ (x \to \infty).$$

Proof. We show first that $\ell(x)$ is a Lévy density. Clearly $\ell(x) \downarrow 0$ as $x \uparrow \infty$, and $\ell(0+) = \infty$. The substitution $v = u\sqrt{x}$ yields $\ell(x) = 2\alpha K x^{-1-\alpha/2}\int_0^\infty v^{\alpha-1}e^{-v^2-2\gamma v\sqrt{x}}dv$, and the integral converges to $\frac{1}{2}\Gamma(\frac{1}{2}\alpha)$ as $x \to 0$. Consequently $\int_0^1 x\ell(x)dx < \infty$. Next, observing that $v^{\alpha-1}e^{-v^2} \sim v^{\alpha-1}$ as $v \to 0$, a Tauberian theorem implies that the last integral is asymptotically equal to $\Gamma(\alpha)(2\gamma\sqrt{x})^{-\alpha}$ as $x \to \infty$, and the second asymptotic relation follows. In particular $\int_1^\infty \ell(x)dx < \infty$. □

It is clear from Lemma 6.3 that $\mathcal{L}(W)$ is not stable, although $P(W > x)$ is asymptotically proportional to the right-hand tail of a stable(α) law (provided it is not spectrally negative). We have not been able to relate $\mathcal{L}(W)$ to simpler known laws. The integral expression (6.6) yields the Laplace transform expression

$$\int_0^\infty x\ell(x)e^{-\theta x}dx = 2\alpha K\int_0^\infty \frac{u^{\alpha-1}}{u^2+2\gamma u+\theta}du = \frac{2\alpha K\pi}{\sin(\alpha\pi)}\cdot\frac{(z_-(\theta))^{\alpha-1} - (z_+(\theta))^{\alpha-1}}{z_+(\theta) - z_-(\theta)} \tag{6.7}$$

where $z_\pm = \gamma \pm \sqrt{\gamma^2 - \theta}$, $|\theta| < \gamma^2$, and the second equality follows from Gradshteyn and Ryzhik [14, 3.223,#1], and it holds for $\alpha \neq 1$. The expression for $\alpha = 1$ is given by

$$2K[\log(z_+(\theta)/z_-(\theta))]/[z_+(\theta) - z_-(\theta)].$$

A further integration yields the explicit expression for the cumulant function of $\mathcal{L}(W)$,

$$\int_0^\infty (1-e^{-\theta x})m(x)dx = \frac{2\pi b}{\Gamma(\alpha)\gamma(2-\alpha)}\frac{(2\gamma)^\alpha - (z_+(\theta))^\alpha - (z_-(\theta))^\alpha}{\sin(\alpha\pi)}, \quad (\alpha \neq 1),$$

valid for $\theta \leq \gamma^2$. This representation is too narrowly defined to give the cumulant function of $Y = W_{\gamma^2}$. An explicit expression for the cumulant function of Y can be given in terms of trigonometric functions, but it yields little insight.

We now explore the GGC properties of $\mathcal{L}(W)$ and $\mathcal{L}(Y)$, beginning by rendering Bondesson's [6, p. 29] definition as follows. We say that the positive law $\mathcal{L}(W)$ is a

GGC if

$$E(e^{-\theta W}) = \exp\left[-\Delta\theta - \int_0^\infty \log\left(1+\frac{\theta}{x}\right) d\mathcal{V}(x)\right], \tag{6.8}$$

where Δ is a constant and $\mathcal{V}(x) \geq 0$ is non-decreasing on $(0,\infty)$ and satisfies the conditions $\int_{0+}^1 |\log x| d\mathcal{V}(x) < \infty$ and $\int_1^\infty x^{-1} d\mathcal{V}(x) < \infty$. We have the following stochastic integral representation,

$$W \stackrel{L}{=} \int_0^\infty \tau^{-1} d_\tau \mathcal{G}(\mathcal{V}(\tau)) \tag{6.9}$$

where $(\mathcal{G}(\tau) : \tau \geq 0)$ is a (standard) gamma process with cumulant function $\log(1+\theta)$, and $\mathcal{V}(\tau)$ functions as a deterministic time transformation. This representation arises from the easily demonstrated result for the stochastic integral $I = \int_0^T k(\tau) d\Lambda(\mathcal{V}(\tau))$, where $k(\tau)$ and $T \leq \infty$ are deterministic, and the L'evy process Λ has characteristic exponent $\psi(t)$: The CF of I is $\exp[-\int_0^T \psi(tk(\tau)) d\mathcal{V}(\tau)]$.

The following result shows that $\mathcal{L}(W)$ is a GGC.

Lemma 6.4. The cumulant function $\kappa_W(\theta)$ of $\mathcal{L}(W)$ has the canonical form (6.8) with $\Delta = 0$ and

$$\mathcal{V}(x) = K\left[\sqrt{\gamma^2 + x} - \gamma\right]^\alpha. \tag{6.10}$$

Proof. Observing that the left-hand side of (6.7) is $\kappa'_W(\theta)$, integration of the second term yields

$$\kappa_W(\theta) = 2\alpha K \int_0^\infty u^{\alpha-1} \log\left(1 + \frac{\theta}{u^2 + 2\gamma u}\right) du.$$

The change of variable $y = u^2 + 2\gamma u$ reduces this to the GGC canonical form (6.8) as asserted. The density $v(x)$ of the measure (6.10) satisfies $v(x) \sim K(x/2\alpha)^{\alpha-1}$ as $x \to 0+$, showing that $\int_0^1 |\log x| v(x) dx < \infty$, and $v(x) \sim \alpha K x^{\frac{1}{2}\alpha - 1}$ as $x \to \infty$, showing $\int_1^\infty [v(x)/x] dx < \infty$. Thus all conditions for GGC membership are satisfied. □

A little manipulation with (6.8) shows that the a-tilt of a GGC is again a GGC with canonical measure $\mathcal{V}(x-a)$. Applying this to (6.10) shows that $Y = W_{\gamma^2}$ is a GGC with canonical measure $\mathcal{V}_\gamma(x) = K([\sqrt{x} - \gamma]^+)^\alpha$. The stochastic integral representations (6.9) of these laws could form the basis of data simulation.

The above decompositions leading to Lemma 6.3 shows the existence of a positive law $\mathcal{L}(W)$ which can be tilted, then used to mix a normal variance, thus yielding the symmetric pruned stable law $X = Z\sqrt{W_{\gamma^2}}$. If desired, asymmetry can be introduced by a further tilting operation as follows. Let $-\gamma < \zeta < \gamma$, and define the law $\mathcal{L}(X_\zeta)$ whose density is proportional to $e^{-\zeta x} f(x)$. From (6.2), we see that its cumulant function is

$$-\frac{b}{2(\alpha-1)}\left[(\gamma+\zeta+\theta)^\alpha + (\gamma-\zeta-\theta)^\alpha - 2\gamma^\alpha\right],$$

showing that this law is realized as $X_\zeta \stackrel{L}{=} V_{\gamma+\zeta}(1) - V_{\gamma-\zeta}(2)$ where the $V(j)$ are independent stable$(\alpha, b, 1)$ variates.

This chain of construction can be applied to any initial positive law. For example, a structurally simpler way of truncating the tails of stable laws could begin with W having a positive stable$(a, b, 1)$ law where $0 < a < 1$. Then W_{γ^2} has LST

$$L(\theta;\zeta) = \exp\left(-\frac{b}{1-a}[(\gamma^2+\theta)^a - \gamma^{2a}]\right).$$

As before, let Z be independent of W_{γ^2} with a normal $\mathcal{N}(0,2)$ law and $X = Z\sqrt{W_{\gamma^2}}$. The CF of X is $\phi(t) = \exp(-\psi(t))$ where

$$\psi(t) = \frac{b}{1-a}\left[(\gamma^2+t^2)^{\alpha/2} - \gamma^\alpha\right], \tag{6.11}$$

where $\alpha = 2a$. The normal inverse Gaussian family corresponds to $\alpha = 1$ [1], and the normal-variance gamma laws arise as the limiting case $\alpha \to 0$ after replacing b with $2b/\alpha$. Note the similarity of (6.11) and the case $\alpha \neq 1$ and $\beta = 0$ of Theorem 5.1; there is no $\cos(\alpha\omega)$ term or dependence on the sign of $\alpha - 1$ in (6.11). The following result lists properties relevant to our theme of this tilted-stable mixture law. Let $K_\lambda(\cdot)$ denote the modified Bessel function of the third kind.

Theorem 6.2. The law defined by (6.11) is infdiv with a symmetric Lévy density

$$n(x) = \frac{ab}{\Gamma(2-a)\sqrt{\pi}}\left(\frac{2\gamma}{|x|}\right)^{a+\frac{1}{2}} K_{a+\frac{1}{2}}(\gamma|x|). \tag{6.12}$$

As $x \to \infty$,

$$n(x) \sim \frac{ab}{\Gamma(2-a)}(2\gamma)^a x^{-1-a} e^{-\gamma x} \tag{6.13}$$

and

$$\bar{F}(x) \sim \frac{ab}{\gamma\Gamma(2-a)}(2\gamma)^a \exp\left(\frac{b\gamma^\alpha}{1-a}\right) x^{-1-a} e^{-\gamma x}. \tag{6.14}$$

The density function has the series representation

$$f(x) = -\frac{M}{|x|}\sum_{j=1}^{\infty} \frac{\Gamma(aj+1)\sin(a\pi j)}{j!}\left[-\rho\left(\frac{|x|}{2\gamma}\right)^a\right]^j K_{aj+\frac{1}{2}}(\gamma|x|), \tag{6.15}$$

where $\rho = b/(1-a)$ and $M = (2\pi)^{-3/2}\gamma^{-\frac{1}{2}}\exp(\rho\gamma^\alpha)$.

Proof. The Lévy density of W_{γ^2} is

$$m(x) = \frac{ab}{\Gamma(2-a)} x^{-1-a} e^{-\gamma^2 x} 1_{(0,\infty)}(x).$$

Lemma 6.1 yields a standard integral [14, p. 340,#9], giving (6.12). The asymptotic form (6.13) immediately follows [14, p. 963, #6], and then (6.14) from Theorem 2.1. For (6.15), let $g_{\gamma^2}(x)$ be the density of W_{γ^2} and observe that the specification of $f(x)$ as a normal-variance mixture entails

$$f(x) = \frac{1}{2\sqrt{\pi}} \int_0^\infty e^{-x^2/4y} g_{\gamma^2}(y) y^{-\frac{1}{2}} dy = \frac{\exp(\rho\alpha^a)}{2\sqrt{\pi}} \int_0^\infty e^{-\nu x^2/4 - \gamma^2 \nu} g_0(1/\nu) \nu^{-3/2} d\nu$$

where $g_0(x)$ is the stable$(a, b, 1)$ density. Inserting the power series representation of $g_0(1/\nu)$ and integrating term-by-term leads to (6.15). □

A key difference between the symmetric pruned stable law and this variance mixture is seen in the differing algebraic factors in the expansions (5.2) and (6.14), $x^{-1-\alpha}$ and $x^{-1-\alpha/2}$, respectively. The pruned stable law allows a little more scope for fitting tails than the tilted stable mixture. Another difference lies in their variances,

$$v_{PS} = \frac{2\alpha b}{2-\alpha} \gamma^{\alpha-2} \qquad \& \qquad v_{TSM} = \frac{2ab}{1-a} \gamma^{\alpha-2},$$

for the pruned stable law and the tilted-stable mixture, respectively. The first is obtained by differentiation of the characteristic exponent in Theorem 5.1 and the second from $v_{NM} = E(Z^2 W_{\gamma^2})$.

There is no reason to expect a closed expression for the sum (6.15), just as there is no closed expression for general stable densities. By contrast, the density for the pruned stable is completely intractable. The case $\alpha = 1$ for the tilted-stable mixture admits the explicit result

$$f(x) = \frac{2b\gamma \exp(2b\gamma)}{\pi} \cdot \frac{K_1(\gamma\sqrt{4b^2 x^2}}{\sqrt{4b^2 + x^2}},$$

which we recognize as the symmetric normal inverse Gaussian law because the base law is the positive stable$(\frac{1}{2})$ law whose tilt is an inverse Gaussian. See [1, 2] for financial applications. An asymmetric extension of (6.11) is produced by a further ζ-tilting operation, that is, multiplying the Lévy density (6.13) by $e^{-\zeta x}$.

An algebraically even simpler starting point gives the very flexible *generalized hyperbolic* (GH) family. Define a measure μ on $(0,\infty)$ by $\mu(dx) = x^{\lambda-1} e^{-\delta^2/x} dx$, where $\delta \geq 0$ and $\lambda \in \mathbb{R}$. This measure is finite iff $\lambda < 0$ and $\delta > 0$, in which case normalization gives the density of the reciprocal gamma family. A normal-variance mixture using this family gives the Student t-laws. If $\delta > 0$ then the γ^2-tilt $e^{-\gamma^2 x} \mu(dx)$ after normalization gives the generalized inverse Gaussian family [34]. Using this family to form a normal-variance mixture gives the symmetric GH family, and ζ-tilting (with $|\zeta| < \gamma$) gives the full GH family, which we denote by GH$(\lambda, \delta; \gamma, \zeta)$. Note that we use a parametrization

differing slightly to the usual accounts in order to more easily compare with the p-stable and t-stable mixture laws. Specifically, the mgf of the GH$(\lambda, \delta; \gamma, \zeta)$ law is

$$M_{GH}(\theta) = \left(\frac{\gamma^2 - \zeta^2}{\gamma^2 - (\theta+\zeta)^2}\right)^{\frac{1}{2}\lambda} \cdot \frac{K_\lambda(2\delta\sqrt{\gamma^2 - (\theta+\zeta)^2})}{K_\lambda(2\delta\sqrt{\gamma^2 - \zeta^2})}$$

and there is an explicit expression for the corresponding density, also in terms of a Bessel function. The tail behaviour of the GH law can be expressed via its density as $f(x) \sim Cx^{\lambda-1}e^{-(\gamma-\zeta)x}$ $(x \to \infty)$. Note that the exponent in the algebraic factor can take any real value. See [12] and references therein for accounts of the GH family.

A related construction of the GH family (*e.g.*, [3, p. 173]) is via $X \stackrel{L}{=} \zeta Y + Z\sqrt{Y}$, where Y has the generalized inverse Gaussian law. The random mean term accomplishes the second tilting operation, but with the modified parametrization GH$(\lambda, \delta; \sqrt{\gamma^2 + \zeta^2}, \zeta)$. Analagous outcomes occur if Y has the mixing law of Theorem 6.1, or the t-stable$(\alpha, b; \gamma^2)$ law.

Rydberg [31], and Barndorff-Nielsen and Shephard [3, 4] are recent surveys of the application to financial data of the GH and related models and the highly developed methodology developed for them.

7 Process and series representations

It hardly needs saying that t-stable and p-stable random variables can be embedded in a Lévy process. The random measure representation of this embedding process does not significantly simplify, except insofar as discussed by Eberlein [12, p.326] whose remarks apply whenever the process has a finite mean. Barndorff-Nielsen and Shephard [4] construct stationary models of Ornstein-Uhlenbeck (OU) type built on the fact that a law $\mathcal{L}(X)$ is self-decomposable iff $X = \int_0^\infty e^{-\tau} d\mathcal{B}(\tau)$ where the integrator is a Lévy process, called the background driving Lévy process (BDLP). The corresponding stationary OU process is $X(\tau) = e^{-\tau}X(0) + \int_0^t e^{-(\tau-u)} d\mathcal{B}(u)$, where $X(0) \stackrel{L}{=} X$. If n and ℓ denote the Lévy densities of X and $\mathcal{B}(1)$, respectively, then $\ell(x) = -(d/dx)(xn(x))$ [4, p. 302]. This construction forms the basis of their coherent modelling methodology mentioned above.

Barndorff-Nielsen and Shephard [3, 4] choose the GH family for $\mathcal{L}(X)$. In principle their general approach is applicable to t-stable and p-stable laws. For example, if we fix $A > 0$ and let $\mathcal{L}(X)$ have the Lévy density

$$n_\gamma(x) = Ax^{-1-\alpha}e^{-\gamma x}1_{(0,\infty)}(x)$$

then

$$\ell(x) = \alpha n_\gamma(x) + \gamma A x^{-\alpha} e^{-\gamma x} 1_{(0,\infty)}(x),$$

which clearly is a Lévy density. It follows that the t-stable law is self-decomposable and its BDLP is the sum of two independent Lévy processes. The first has Lévy density

$\alpha n_\gamma(x)$ corresponding to the time dilated embedding process $(\Lambda(\alpha\tau))$. If $\alpha < 1$ then the second component is a compound Poisson process having a gamma jump law which in obvious notation we denote by $\mathrm{Gam}(1-\alpha,\gamma)$. If $\alpha = 1$ then the second component is a gamma process, and if $1 < \alpha < 2$ then it is generated by a t-stable$(\alpha - 1, b; \gamma)$ law, where $b = (A/\alpha)\Gamma(2-\alpha)$. Similar representations hold for p-stable laws, even in the asymmetric case.

The tilted-stable mixture law of Theorem 6.2 being a normal-variance mixture with a GGC mixing law is a member of $\mathcal{T}_e$, and hence it too is self-decomposable. It follows that

$$\ell(x) = 2an(x) + \frac{2ab\gamma^2}{\Gamma(2-a)\sqrt{\pi}} \left(\frac{2\gamma}{|x|}\right)^{a-\frac{1}{2}} K_{a-\frac{1}{2}}(\gamma|x|).$$

The proof involves differentiating (6.12) and using the fact $xK'_\nu(x) + \nu K_\nu(x) = xK_{\nu-1}(x)$. So again the BDLP resolves into independent components with the first a time dilated version of the embedding process. The second component can be shown to be a compound Poisson process of normal-variance mixture type, $C_\tau = \sum_{j=1}^{N_\tau} Z_j\sqrt{V_j}$ where (N_τ) is a Poisson process with rate $2ab\gamma^2/(1-a)$, the Z_j are independent copies of $Z \sim \mathcal{N}(0,2)$, and the V_j are independent $\mathrm{Gam}(a,\gamma^2)$ variates. This decomposition generalizes Proposition 6.2 in [4] $(a = \frac{1}{2})$ and it represents the second component in a simpler and more explicit form than they achieve.

Motivated in part by the search for prior laws for Bayesian nonparametric inference, there is a body of work on random series representations of stable, and more generally, of infdiv variates. See [21, 30, 32, 36] for a fairly complete listing of the literature. Practicable ways of simulating stably distributed data appears to be a subsidiary motivation, but it is generally agreed now that the series converge too slowly to be useful for this purpose. As we now show, an elementary treatment results from imposing a regular variation condition on a Lévy measure. This condition holds for all modelling applications we know of.

Let (N_τ) be a unit rate Poisson process with event times $T_1 < T_2 < \cdots$. Given a Lévy measure μ, define $M(x) = \mu(x,\infty)$ and suppose there is a constant $\beta > 0$ and a function L slowly varying at infinity such that

$$M(x) = x^{-\beta}L(x^{-\beta}) \qquad (x > 0),$$

that is, M is regularly varying at the origin. The constraint $\int_0^1 x^2\mu(dx) < \infty$ implies that $\beta \leq 2$. The function M has an asymptotic inverse

$$\rho(v) = \left(vL^{\#}(v)\right)^{-1/\beta} > 0, \qquad (x > 0) \tag{7.1}$$

where $L^{\#}$ is the slowly varying conjugate of L [5]. Finally, let $\{Y_n : n \geq 1\}$ denote independent copies of Y which has CF σ and first moment ξ_1, when it is defined.

Theorem 7.1. (i) Suppose $I_\rho := \int_1^\infty \rho(v)dv < \infty$. Then the series

$$X = \sum_{n=1}^{\infty} Y_n \rho(T_n) \tag{7.2}$$

converges absolutely almost surely iff

$$\beta < 1 \qquad \& \qquad E|Y|^\beta L(|Y|^\beta) < \infty$$

or

$$\beta = 1 \qquad \& \qquad E|Y|\ell(|Y|) < \infty,$$

where $\ell(x) = \int_x^\infty y^{-1}L(y)dy < \infty$. If either condition holds then the CF of X is

$$\phi(t) = \exp\left[-\int_0^\infty (1-\sigma(tx))\mu(dx)\right]. \tag{7.3}$$

(ii) Suppose $1 \le \beta < 2$, $I_\rho = \infty$, ζ_1 is finite, and L is normalized slowly varying. Then the series

$$\bar{X} = \sum_{n=1}^{\infty} (Y_n \rho(T_n) - \zeta_1 \rho(n)) \tag{7.4}$$

converges unconditionally almost surely if $E|Y|^p < \infty$ for some $p > \beta$. If $\zeta_1 = 0$ then $\bar{X}$ has the CF (7.3). If $\zeta_1 \neq 0$ then

$$\dot{X} = \sum_{n=1}^{\infty} \left(Y_n \rho(T_n) - \zeta_1 \int_n^{n+1} \rho(v)dv \right) - B, \tag{7.5}$$

where $B = \zeta_1 \int_1^{M(1)} \rho(v)dv$, converges unconditionally almost surely under the above moment condition, and its CF is

$$\tilde{\phi}(t) = \exp\left[-\int_1^\infty (1-\sigma(tx))\mu(dx) - \int_0^1 (1-\sigma(tx)+i\zeta_1 tx)\mu(dx)\right].$$

Proof. (i) The law of large numbers and (7.1) imply that $\rho(T_n) \sim \rho(n)$ and the absolute convergence assertions are an immediate consequence of general convergence criteria for random Dirichlet series: See Corollary 2.2(b,c) in Pakes [28], observing that $\beta \le 1$ is necessary for $I_\rho < \infty$.

The form (7.5) of the CF is derived essentially as in [21]. If $X_n = \sum_{j=1}^{N_n} Y_j \rho(T_j)$ then

$$\begin{aligned} E\left(e^{itX_n}\right) &= E\left[E\left(e^{itX_n}\right)|N_n\right] = E\left[\left(n^{-1}\int_0^n \sigma(t\rho(v)dv\right)^{N_n}\right], \\ &= \exp\left[-\int_0^n (1-\sigma(t\rho(v)))dv\right] = \exp\left[-\int_{\rho(n)}^\infty (1-\sigma(tx))\mu(dx)\right] \end{aligned} \tag{7.6}$$

and (7.3) follows since $X_n \overset{a.s.}{\to} X$.
(ii) Our assumptions imply that $I_\rho = \infty$ whence $\sum_{n\geq 1} Y_n \rho(T_n)$ is almost surely divergent (Pakes [28, Corollary 2(a,c)]). Express the summands in (7.4) as $Y_n[\rho(T_n) - \rho(n)] + [Y_n - \zeta_1]\rho(n) \equiv U_{1n} + U_{2n}$ and let $\bar{\kappa} = \sup\{\kappa \geq 1 : E|Y|^\kappa < \infty\}$. Now $\sum_{n\geq 1} U_{2n}$ is a random Dirichlet series if it is regarded as a function of β^{-1}, and Theorem 3.2 (b) of Pakes [28] asserts that its abscissa of unconditional convergence is $\max(\frac{1}{2}, \bar{\kappa}^{-1})$. Choose $p < 2$ and note that since $\beta^{-1} > \frac{1}{2}$ we have $\beta^{-1} > p^{-1} \geq \bar{m}^{-1}$, and hence this series is almost surely unconditionally convergent under our moment hypothesis.

Observe that the normalization assumption on L implies that $L^\#$ is normalized slowly varying, and hence that $|L^\#(T_n)/L^\#(n) - 1| = o(|T_n - n|/n)$. This estimate and the mean value theorem imply that

$$|\rho(T_n) - \rho(n)| \sim \frac{\rho(n)}{\beta n}|T_n - n|.$$

We infer from the Marcinkiewicz-Zygmund strong law [9, p.122] that a.s. $|T_n - n| = o(n^{1/p})$ and hence that $|\rho(T_n) - \rho(n)| = o(n^{-1-(\beta^{-1}-p^{-1})}L^\#(n))$. It follows that $\sum_{n\geq 1} U_{1n}$ is almost surely absolutely convergent. If $\zeta_1 = 0$ then (7.6) still holds and hence the integral has a finite limit as $n \to \infty$.

If $\zeta_1 \neq 0$ write the nth partial sum of (7.5) is

$$\mathcal{P}(n) = \sum_{j=1}^{n} (Y_j \rho(T_j) - \zeta_1 \rho(j)) + \sum_{j=1}^{n} \left(\rho(j) - \int_j^{j+1} \rho(v)dv\right) - B.$$

The terms in the second sum are non-negative and bounded above by $\rho(j) - \rho(j+1)$, and hence that sum converges as $n \to \infty$. This establishes the unconditional convergence of the series (7.5). Observe now that

$$X_n'' := X_n - \zeta_1 \int_{N_n}^{n} \rho(v)dv = \mathcal{P}(N_n) + \zeta_1 H(n) - B,$$

where

$$H(n) = \int_n^{1+N_n} \rho(v)dv = O(\rho(n)|N_n - n|) = \rho(n)o(\sqrt{n}\log n) \to 0,$$

and the final estimate is a consequence of a strong law of Kolmogorov. (In Feller's rendering [13, p. 239], for example, take his independent summands X_k to have the same law and $b_k = \sqrt{k}\log k$. Of course, the above estimate follows from the more abstruse law of the iterated logarithm.) It follows that X_n'' has a limit law coinciding with $\mathcal{L}(\tilde{X})$.

But since $\int_{N(1)}^{n} \rho(v)dv = \int_{\rho(n)}^{1} x\mu(dx)$, it follows from (7.6) that

$$E\left(e^{itX_n''}\right) = \exp\left[-\int_1^\infty (1 - \sigma(tx))\mu(dx) - \int_{\rho(n)}^1 (1 - \sigma(tx) + i\zeta_1 tx)\mu(dx)\right],$$

and this converges to $\tilde{\phi}(t)$ because the series (7.5) converges. □

Two boundary cases, which are not covered by Theorem 7.1, are stated without proof in the next result. Its proof is similar to that above, but using Corollary 2.3 and Theorem 3.2 in Pakes [28].

Theorem 7.2. If M is slowly varying at the origin then the series (7.2) converges absolutely almost surely iff $EM(|Y|^{-1}) < \infty$. If $\beta = 2$ and $\zeta_1 = 0$ then (7.2) converges unconditionally almost surely iff $\tilde{\ell}(x) = \int_x^\infty M(y^{-2})dy/y < \infty$ $(x > 0)$ and $E|Y^2|\tilde{\ell}(|Y|) < \infty$. If either convergence criterion is satisfied then (7.3) holds.

Observe that if $M(0+) < \infty$ then $\rho(v) \equiv 0$ if $v > M(0+)$ and the series (7.2) has finitely many non-zero terms, that is, X has a compound Poisson law. In the case $\beta = 2$ note that $\tilde{\ell}$ is slowly varying. Finally, we mention that representations for the embedding process, with its time parameter restricted to $[0,1]$, are obtained by replacing Y with $Y1_{(0,U]}(\tau)$ in the above series, where U has the uniform law on $(0,1]$.

It is clear that the laws of X and $\tilde{X}$ are infdiv since the Lévy measure $\tau\mu$ induces the function $\rho(v/\tau)$. In particular, if μ has density m and Y has density f, then X and $\tilde{X}$ have the Lévy density

$$n(x) = \int_0^\infty f(x/y)m(y)dy/y = \int_0^\infty f(y\mathrm{sgn}(x))m(|x|/y)dy/y. \tag{7.7}$$

Thus desired functional forms of n in principle can be tailored from convenient choices of m and f. For example, it is known that $m(x) = Ax^{-1-\alpha}$ yields spectrally positive (respectively, two-sided) stable(α)) laws for any one-sided (respectively, two-sided) $\mathcal{L}(Y)$, provided the convergence criteria are satisfied. The common choice is the point mass at unity (respectively, $P(Y = \pm 1) = \frac{1}{2}$). If $m(x) = Ax^{-1-\alpha}e^{-\gamma x}1_{(0,\infty)}(x)$ then these choices for $\mathcal{L}(Y)$ give $n(x) = m(x)$ in the first case (t-stable), and $n(x) = \frac{1}{2}m(|x|)$ in the second case (p-stable). Integration and changing variables yields

$$M(x) = ax^{-1-\alpha}E_{1+\alpha}(\gamma x)$$

where $E_{1+\alpha}$ is an exponential integral.

If f is the $\mathcal{N}(0,2)$ density, then (7.7) has the normal mixture form in Lemma 6.1 after replacing $m(x)$ there with $2xm(x^2)$. To realize the t-stable or symmetric p-stable laws, it follows from (6.5) that we must have

$$m(x) = A2^{1-\alpha}x^{-1-\alpha}e^{-\gamma^2x^2}\Psi(\tfrac{1}{2}\alpha, \tfrac{1}{2}; \gamma^2x^2).$$

Similarly, p-stable laws can be achieved by taking

$$f(x) = \frac{\gamma^\delta}{2\Gamma(\delta)} \cdot |x|^{\delta-1}e^{-\gamma|x|} \qquad \& \qquad m(x) = \frac{2A\Gamma(\delta)\gamma^\alpha}{\Gamma(\alpha+\delta)} \cdot x^{-1-\alpha}(1-x)^{\alpha+\delta-1}1_{(0,1)}(x).$$

Explicit determination of ρ in any of these cases is problematic.

8 Final remarks

We have illuminated confused definitions of exponentially truncated stable laws and elicited some properties of the pruned version. Symmetric pruned stable laws have been contrasted with normal variance mixtures using a tilted positive stable law or an inverse Gaussian law. Two characteristics stand out. The first is the absence of explicit expressions for the densities of the pruned stable or the tilted-stable mixture. The second is the restricted ranges which are permissable for the exponent values in the algebraic factors occuring in the tail estimates; $(1,3)$ for (5.2) and $(1,2)$ for (6.14).

Proponents of truncated/pruned stable laws argue their case in terms of the good fits obtained for selected data sets. It seems to us that a consistent application of this criterion should cause abandonment of these laws in favour of generalized hyperbolic laws. The GH family of laws is more useful because

- There are explicit expressions in terms of Bessel functions for their density functions, their Lévy densities, and for their moment generating functions for all parameter values;
- The family includes several commonly used sub-families;
- Members have the financially desirable property of being represented as normal variance mixtures;
- The family is more flexible for data fitting purposes, particularly by having greater scope in its tail behaviour.

The last point is nicely illustrated by Hurst and Platen [18] in their examination of five major world market index series. They fit eight types of symmetric law to these series, including the stable, Student-t, and the GH family. The Student-t family is declared the 'winner' in the sense of achieving a uniformly better fit according to a likelihood ratio criterion. The GH family fits equally well, but at the expense of an extra parameter. The Student t-law density function $f(x) \sim const.x^{-1-d}$ as $x \to \infty$, where $d > 0$ is the degrees-of-freedom parameter. In all cases its estimated value $\hat{d}$ lies outside the interval $(0,2)$, the permissable range of the stable index α. The estimates $\hat{\gamma}$ of the exponential decay factor for symmetric GH laws all appear to be very close to zero, though it should be noted that only estimates of $2\delta\gamma$ are actually reported. However, the fact that $\hat{\gamma} > 0$ implies that $-\hat{\lambda}$ should be smaller than $\hat{d}$, as indeed it is. In fact $-\hat{\lambda} > 0$, and it lies outside the interval $(0,2)$ for the Australian index series, but not for the other series. For all series, the best fitting laws have smaller tails than any non-normal stable law can attain. One anticipates that fitting a symmetric p-stable law will show little improvement over the stable, and not so good a fit as the symmetric GH. It would be worthwhile fitting the p-stable, if only to eliminate it from further consideration.

Anthony G. Pakes, Department of Mathematics and Statistics, The University of Western Australia, 35 Stirling Highway, 6009 Crawley, Western Australia,
pakes@maths.uwa.edu.au

References

[1] O. Barndorff-Nielsen. Normal inverse Gaussian distributions and stochastic volatility. *Scandinavian Journal of Statistics*, 24:1–13, 1997.

[2] O. Barndorff-Nielsen. Apparent scaling. *Finance and Stochastics*, 5:103–113, 2001.

[3] O. Barndorff-Nielsen and N. Shephard. Non-Gaussian Ornstein-Uhlenbeck-based models and some of their uses in financial economics. *Journal of the Royal Statistical Society*, 63(Part 2):167–241, 2001.

[4] O. Barndorff-Nielsen and N. Shephard. Modelling by Lévy processes for financial econometrics. In O. Barndorff-Nielsen, T. Mikosch, and S. Resnick, editors, *Lévy Process – Theory and Applications*, pages 283–318, Birkhäuser, Boston, 2001.

[5] N. H. Bingham, C. M. Goldie, and J. L. Teugels. *Regular Variation*. Cambridge University Press, Cambridge, 1987.

[6] L. Bondesson. *Generalized Gamma Convolutions and Related Classes of Distributions*. Springer-Verlag, New York, 1992.

[7] J.-P. Bouchard and M. Potters. *Theory of Financial Risks*. Cambridge University Press, Cambridge, 2000.

[8] V. P. Chistyakov. A theorem on sums of independent positive random variables and its applications to branching random processes. *Theory of Probability and its Applications*, 9:640–648, 1964.

[9] Y. S. Chow and H. Teicher. *Probability Theory*. Springer-Verlag, New York, 1978.

[10] D. B. H. Cline. Convolution tails, product tails and domains of attraction. *Probability Theory and Related Fields*, 72:529–557, 1986.

[11] D. B. H. Cline. Convolutions of distributions with exponential and subexponential tails. *Journal of the Australian Mathematical Society (Series A)*, 43:347–365, 1987.

[12] E. Eberlein. Application of generalized hyperbolic Lévy motions to finance. In O. Barndorff-Nielsen, T. Mikosch, and S. Resnick, editors, *Lévy Process – Theory and Applications*, pages 319–336, Birkhäuser, Boston, 2001.

[13] W. Feller. *An Introduction to Probability Theory and its Applications, Vol. II.* Wiley, New York, 1971.

[14] I. S. Gradshteyn and I. M. Ryzhik. *Table of Integrals, Series, and Products.* Academic Press, New York, 1980.

[15] P. Hall. A comedy of errors: the canonical form for a stable characteristic function. *Bulletin of the London Mathematical Society*, 13:23–28, 1981.

[16] J. Hoffmann-Jorgensen. Stable densities. *Theory of Probability and its Applications*, 38:350–355, 1994.

[17] P. Hougaard. Survival models for heterogeneous populations derived from stable distributions. *Biometrika*, 73:387–396, 1986.

[18] S. R. Hurst and E. Platen. The marginal distributions of returns and volatility. *IMS Lecture Notes – Monograph Series Volume 31*, Hayward, CA, 1997.

[19] I. Koponen. Analytic approach to the problem of convergence of truncated Lévy flights towards the Gaussian stochastic process. *Physical Review E*, 52:1197–1198, 1995.

[20] N. N. Lebedev. *Special Functions and their Applications.* Dover Publications, New York, 1972.

[21] R. LePage. Multidimensional infinitely divisible variables and processes. Part I: Stable case. In S. Cambanis and A. Weron, editors, *Probability Theory on Vector Spaces IV. Proceedings 1987*, pages 153–163, Springer, Berlin, 1989.

[22] J. B. McDonald. Probability distributions for financial models. In G. S. Maddala and C. R. Rao, editors, *Handbook of Statistics Vol. 14*, pages 427–461, Elsevier Science, Amsterdam, 1996.

[23] B. Mandelbrot. *Fractals and Scaling in Finance.* Springer, New York, 1997.

[24] R. N. Mantegna and H. E. Stanley. Stochastic process with ultraslow convergence to a Gaussian: The truncated Lévy flight. *Physical Review Letters*, 73:2946–2949, 1994.

[25] A. Matacz. Financial modelling and option theory with the truncated Lévy process. *International Journal of Theoretical and Applied Finance*, 3:143–160, 2000.

[26] H. Ohkubo. On the asymptotic tail behaviour of infinitely divisible distributions. *Yokohama Mathematical Journal*, 27:77–89, 1979.

[27] A. G. Pakes. Convolution equivalence and infinite divisibility. Submitted, 2002.

[28] A. G. Pakes. Convergence and divergence of random series. *Australian & New Zealand Journal of Statistics*, Submitted, 2002.

[29] W. Paul and J. Baschnagel. *Stochastic Processes from Physics to Finance*. Springer, Berlin, 1999.

[30] J. Rosiński, J. Series representations of Lévy processes from the perspective of point processes. In O. Barndorff-Nielsen, T. Mikosch, and S. Resnick, editors, *Lévy Process - Theory and Applications*, pages 401–415, Birkhäuser, Boston, 2001.

[31] T. H. Rydberg. Realistic statistical modelling of financial data. *International Statistical Review*, 68:233–258, 2000.

[32] G. Samorodnitsky and M. S. Taqqu. *Stable non-Gaussian Random Processes*. Chapman & Hall, New York, 1994.

[33] K. Sato. *Lévy Processes and Infinitely Divisible Distributions*. Cambridge University Press, Cambridge, 1999.

[34] V. Seshadri. *The Inverse Gaussian Distribution: A Case Study in Exponential Families*. Clarendon Press, Oxford, 1993.

[35] J. Voit. *The Statistical Mechanics of Financial Markets*. Springer, Berlin, 2001.

[36] S. Walker and P. Damien. Representations of Lévy processes without Gaussian components. *Biometrika*, 87:477–483, 2000.

[37] E. Willekens. Subexponentiality on the real line. Unpublished Research Report, 1987.

[38] E. J. Williams. Some representations of stable random variables as products. *Biometrika* 64:167–169, 1977.

[39] V. M. Zolotarev. *One-Dimensional Stable Distributions*. American Mathematical Society, Providence, RI, 1986.

Designs on Association Schemes

R. A. Bailey

Abstract

An *association scheme* partitions a finite set Ω into symmetric subsets, one of which is the diagonal subset. This paper develops the idea of a design map between two association schemes. In many designed experiments, the structure on the experimental units is an orthogonal block structure. These appear to be the structures where both the components-of-variance and patterns-of-covariance approaches (almost) agree. By replacing orthogonal block structures by association schemes, only the patterns-of-covariance model generalizes.

Keywords: association schemes; balanced design; experimental design; general balance; Latin square; orthogonal block structures

1 Introduction

Terry Speed and I worked together in the 1980s on problems in the analysis of variance. My motivation was to understand how an analysis of variance could be defined by the randomization used in setting up the experiment [3]; his was more fundamental, seeking to answer the question 'What is an analysis of variance?' [29]. We were both heavily influenced by John Nelder's two papers [25, 26], in which he defines simple orthogonal block structures, makes an unsubstantiated claim about randomization, defines general balance, and shows how to analyse data from generally balanced experiments with many strata.

In joint work with Cheryl Praeger and Chris Rowley [7], we were able to generalize Nelder's simple orthogonal block structures to a class which I now call poset block structures, and prove that Nelder's claim about randomization holds in poset block structures. The other three authors extended this work in [27], while I showed in [4] that poset block structures are the same as the 'complete balanced response structures' which Kempthorne and his team at Ames, Iowa had studied extensively [21, 22, 32, 36].

More surprisingly, in [30, 31] Speed and I found that if you ignore the question of randomization then you can define an even wider class of structures in which all of Nelder's theory carries through, with rather easy proofs. Today I use the term 'orthogonal block structure' for structures in this class [4]. An important input from Speed was to recognise that these orthogonal block structures are association schemes: this insight has influenced my own subsequent work enormously. A second key input from Speed was to introduce concepts from partial orders, most importantly the Möbius function,

which enables us to give explicit formulae which do not involve matrix inverses. In conversation in 1990, Oscar Kempthorne told me how important he thought the introduction of the Möbius function was to the subject. He said that the Möbius function really did the job; he wished that he and his colleagues had known about it.

Orthogonal block structures are reviewed in Section 2. They give a context for the remainder of the paper. In a very large proportion of designed experiments, the structure on the experimental units is an orthogonal block structure, but other association schemes do occur.

In [20], Houtman and Speed examined general balance in detail. In order to include as many covariance structures as possible, they did not restrict their attention to structures defined by combinatorial concepts such as 'in the same block'. Instead, they defined a linear model to 'have orthogonal block structure' if all the eigenspaces of the covariance matrix are known. Everything about general balance and estimation was worked through in this framework. It is certainly true that general balance can be fruitfully defined whenever the eigenspaces of the covariance matrix are known. However, I prefer to retain the term 'orthogonal block structure' for the combinatorial structures defined in Section 2.

Section 5.2 of [20] discusses partially balanced incomplete-block designs. These have an association scheme defined on the set of treatments: indeed, this is the context in which association schemes were defined [9, 10]. It is fairly natural to extend the idea of partial balance to other orthogonal block structures: see [8, 18, 19] for nested block designs and [16] for nested row-column designs. However, Section 5.2 went far beyond that, because it proposed that both the set of treatments and the set of experimental units could have an arbitrary association scheme defined on them.

This idea, of two association schemes and a design map from one to the other, was given less than two pages in [20]. It is developed in the main part of this paper.

There are two rather natural ways of defining a covariance matrix on a structured set of random variables. If the structure is defined by partitions on the set, then independent random variables can be associated with each class (part) of each partition: those associated with the same partition have the same variance. This gives the components-of-variance model, which is widely used: see [28]. On the other hand, if the structure is defined by a partition on the ordered pairs from the set, one can demand that the covariance is the same for all pairs in the same part. This gives the patterns-of-covariance model, which is natural if the model is justified by randomization: see [3]. Orthogonal block structures appear to be precisely those structures where not only are both approaches possible and tractable but also the two approaches (almost) agree, as shown in Section 2. However, in generalizing from orthogonal block structures to association schemes, only the second approach is possible.

2 Orthogonal block structures

Let F be a partition of a finite set Ω. Define the subspace V_F of the real vector space $\mathbb{R}^\Omega$ to consist of all those vectors which are constant on every class of F. Then $\dim V_F$ is equal to n_F, the number of classes of F.

Two $\Omega \times \Omega$ real matrices are defined by F. The first is the *relation matrix* R_F, whose (α,β)-entry is equal to 1 if $F(\alpha) = F(\beta)$ and to 0 otherwise. Here we are writing $F(\alpha)$ for the class of F which contains α. The second is the *projection matrix* P_F. There is a natural inner product $\langle\ ,\ \rangle$ on $\mathbb{R}^\Omega$ given by

$$\langle v, w\rangle = \sum_{\omega\in\Omega} v_\omega w_\omega;$$

this defines orthogonality, and P_F is just the matrix of orthogonal projection onto V_F. The (α,β)-entry of P_F is equal to $1/|F(\alpha)|$ if $F(\alpha) = F(\beta)$; otherwise it is zero.

The partition F is defined to be *uniform* if all of its classes have the same size, which must be $|\Omega|/n_F$. If F is uniform then $|\Omega|P_F = n_F R_F$.

There are two trivial uniform partitions of Ω. The *universal* partition U has a single class. Thus V_U is the 1-dimensional subspace consisting of the constant vectors. At the other extreme, the classes of the *equality* partition E are all singletons, so $V_E = \mathbb{R}^\Omega$.

Suppose that F and G are two partitions of Ω. We say that F is *finer* than G, and write $F \preccurlyeq G$, if every F-class is contained in a G-class. In this case, $V_G \leqslant V_F$. In particular, $E \preccurlyeq F \preccurlyeq U$ for every partition F of Ω.

More generally, the *infimum* $F \wedge G$ of F and G is defined to be the coarsest partition which is finer than both F and G. Its classes are the non-empty intersections of F-classes with G-classes. Dually, the *supremum* $F \vee G$ of F and G is the finest partition which is coarser than both F and G. Its classes are the connected components of the graph whose vertices are the elements of Ω and whose edges are the pairs $\{\alpha,\beta\}$ for which $F(\alpha) = F(\beta)$ or $G(\alpha) = G(\beta)$. It follows that $V_{F\vee G} = V_F \cap V_G$; however, there is no simple expression for $V_{F\wedge G}$.

Partitions F and G are defined to be *orthogonal* to each other if P_F commutes with P_G; that is, if V_F is geometrically orthogonal to V_G in the sense that $V_F \cap V_{F\vee G}^\perp$ is orthogonal to $V_G \cap V_{F\vee G}^\perp$: see [33]. If $F \preccurlyeq G$ then $F \vee G = G$ so $V_G \cap V_{F\vee G}^\perp$ is the zero subspace, which is orthogonal to all subspaces, so F is orthogonal to G. In particular, F is orthogonal to U, E and itself.

Orthogonality is equivalent to a combinatorial condition that statisticians will recognise as 'proportional meeting'. Figure 1 shows five examples where the set Ω is a rectangle. In each case F is the partition into rows, G is the partition into columns, and the numbers show the size of the row-column intersections. In (a), (c) and (d), each of F, G and $F \wedge G$ is uniform; in (e), F and G are uniform but neither $F \wedge G$ nor $F \vee G$ is; in (a)–(d), $F \vee G = U$; in (c) and (d), $F \wedge G = E$; in (a), (b), (d)and (e), F is orthogonal to G.

If F, G, $F \wedge G$ and $F \vee G$ are all uniform then there is a simple criterion for orthogonality: F is orthogonal to G if and only if, for all pairs α and β, $F(\alpha) \cap G(\beta)$ is

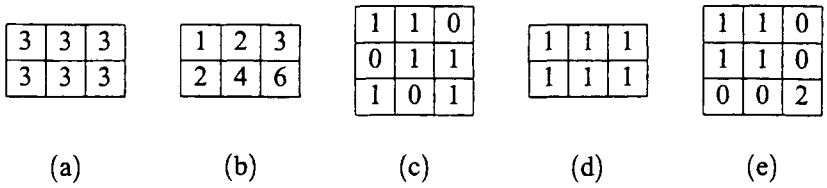


Figure 1: Examples to demonstrate orthogonality

non-empty if and only if $G(\alpha) \cap F(\beta)$ is non-empty.

Sections 42 and 76 of [17] show that if F is orthogonal to G then

$$P_F P_G = P_{F \vee G}. \tag{1}$$

Definition

An *orthogonal block structure* on Ω is a set $\mathcal{F}$ of uniform partitions of Ω such that

(i) $\mathcal{F}$ contains E and U;

(ii) if $F \in \mathcal{F}$ and $G \in \mathcal{F}$ then $F \wedge G \in \mathcal{F}$ and $F \vee G \in \mathcal{F}$;

(iii) if $F \in \mathcal{F}$ and $G \in \mathcal{F}$ then F is orthogonal to G.

Suppose that $\mathcal{F}$ is an orthogonal block structure. Then $\mathcal{F}$ defines a partition of $\Omega \times \Omega$ into *associate classes* C_F labelled by elements of $\mathcal{F}$, as follows. Let α and β be in Ω. Since $\mathcal{F}$ is closed under $\wedge$, there is a unique finest F in $\mathcal{F}$ such that $F(\alpha) = F(\beta)$. Now the class $C(\alpha, \beta)$ containing (α, β) is C_F, and we call α and β *F-associates*. In other words, $(\alpha, \beta) \in C_F$ if and only if (i) $F(\alpha) = F(\beta)$ and (ii) if $G \in \mathcal{F}$ and $G(\alpha) = G(\beta)$ then $F \preccurlyeq G$. The $\Omega \times \Omega$ *adjacency matrix* A_F is defined to have (α, β)-entry equal to 1 if α and β are F-associates; otherwise it is zero.

Example 1

Suppose that Ω consists of b blocks, each of which is an $n \times m$ rectangular array. Let B be the partition into the blocks, F the partition into the bn rows and G the partition into the bm columns. Then $\{E, F, G, B, U\}$ is an orthogonal block structure. Moreover (α, β) is in

C_E if $\alpha = \beta$
C_F if $\alpha \neq \beta$ but α and β are in the same row
C_G if $\alpha \neq \beta$ but α and β are in the same column
C_B if α and β are in the same block but different rows and columns
C_U if α and β are in different blocks.

Also, given an orthogonal block structure $\mathcal{F}$, define

$$W_F = V_F \cap \bigcap_{G \in \mathcal{F},\, F \prec G} V_G^{\perp} \tag{2}$$

for F in $\mathcal{F}$. Since $\mathcal{F}$ is closed under $\vee$ and satisfies the orthogonality condition, it is fairly easy to show that $W_F \perp W_G$ if $F \neq G$, and that

$$V_F = \bigoplus_{G \in \mathcal{F},\, F \preccurlyeq G} W_G \tag{3}$$

for F in $\mathcal{F}$.

$$\zeta(F,G) = \begin{cases} 1 & \text{if} F \preccurlyeq G \\ 0 & \text{otherwise.} \end{cases}$$

The elements of $\mathcal{F}$ can be written in such an order that, as a matrix, ζ is upper triangular with all diagonal elements equal to 1. Therefore, ζ has an inverse matrix μ, and it is this which is called the Möbius function.

The definition of A_F shows that

$$R_F = \sum_{G \in \mathcal{F}} \zeta(G,F) A_G$$

for all F in $\mathcal{F}$. Hence

$$A_F = \sum_{G \in \mathcal{F}} \mu(G,F) R_G$$

for all F in $\mathcal{F}$, and $\operatorname{span}\{A_F : F \in \mathcal{F}\} = \operatorname{span}\{R_F : F \in \mathcal{F}\}$. Since all the partitions are uniform, $|\Omega| P_F = n_F R_F$ for all F in $\mathcal{F}$, and $\operatorname{span}\{P_F : F \in \mathcal{F}\} = \operatorname{span}\{R_F : F \in \mathcal{F}\}$. Finally, let S_F be the matrix of orthogonal projection onto W_F. Equation (3) shows that

$$P_F = \sum_{G \in \mathcal{F}} \zeta(F,G) S_G$$

for all F in $\mathcal{F}$, and hence

$$S_F = \sum_{G \in \mathcal{F}} \mu(F,G) P_G.$$

Therefore

$$\begin{aligned} \operatorname{span}\{A_F : F \in \mathcal{F}\} &= \operatorname{span}\{R_F : F \in \mathcal{F}\} = \operatorname{span}\{P_F : F \in \mathcal{F}\} \\ &= \operatorname{span}\{S_F : F \in \mathcal{F}\}. \end{aligned} \tag{4}$$

Now suppose that Ω is the set of experimental units in an experiment. We observe a data vector that is a realization of a random vector Y. What should we assume about the covariance matrix $\mathrm{Cov}(Y)$?

One common assumption is that there are independent random variables associated with every class of every partition in $\mathcal{F}$: all those associated with F have variance σ_F^2. This gives

$$\mathrm{Cov}(Y) = \sum_{F \in \mathcal{F}} \sigma_F^2 R_F, \tag{5}$$

which is called the *components-of-variance* model. A second assumption is that all pairs of F-associates have the same covariance γ_F, for all F in $\mathcal{F}$. This gives the *patterns-of-covariance* model

$$\mathrm{Cov}(Y) = \sum_{F \in \mathcal{F}} \gamma_F A_F. \tag{6}$$

Because of Equation (4), both of Equations (5) and (6) can be reparametrized as

$$\mathrm{Cov}(Y) = \sum_{F \in \mathcal{F}} \xi_F S_F. \tag{7}$$

This shows that the spaces W_F are eigenspaces of $\mathrm{Cov}(Y)$ in both cases, with eigenvalues ξ_F. Nelder called these eigenspaces *strata*, so the quantities ξ_F are called the *stratum variances*. His proposed analysis of the data begins by projecting the data onto each stratum, where it has effectively a scalar covariance matrix, so that ordinary least squares can be applied: see also [1].

However, models (5) and (6) are not identical. A covariance matrix is non-negative definite, so Equation (7) is constrained by

$$\xi_F \geqslant 0 \quad \text{for all } F \text{ in } \mathcal{F}. \tag{8}$$

Variances must also be non-negative, so (5) is constrained by

$$\sigma_F^2 \geqslant 0 \quad \text{for all } F \text{ in } \mathcal{F}. \tag{9}$$

Now,

$$\sum_F \sigma_F^2 R_F = \sum_F \sigma_F^2 \frac{|\Omega|}{n_F} P_F = \sum_F \frac{|\Omega|}{n_F} \sigma_F^2 \sum_G \zeta(F,G) S_G$$

so

$$\xi_G = \sum_F \zeta(F,G) \frac{|\Omega|}{n_F} \sigma_F^2,$$

and therefore condition (9) is stronger than condition (8).

In [20], Houtman and Speed effectively started with Equation (7) for known projectors S_F. By replacing orthogonal block structures by association schemes, we can also retain Equation (6) for known adjacency matrices A_F. That is, the patterns-of-covariance model generalizes but the components-of-variance model does not.

3 Association schemes

A subset of $\Omega \times \Omega$ can be identified with its $\Omega \times \Omega$ *adjacency matrix* A, whose (α, β)-entry is equal to 1 if (α, β) is in the subset and to 0 otherwise. The subset is said to be *symmetric* if its adjacency matrix is a symmetric matrix. The *diagonal* subset is $\{(\omega, \omega) : \omega \in \Omega\}$: its adjacency matrix is the identity matrix I. The adjacency matrix of $\Omega \times \Omega$ is the all-1 matrix J.

Definition
An *association scheme* on Ω is a partition of Ω into symmetric subsets, called *associate classes*, one of which is the diagonal subset, such that the product of any two of its adjacency matrices is a real linear combination of the adjacency matrices of associate classes.

The *trivial* association scheme has just one non-diagonal associate class. If B is a non-trivial uniform partition of Ω then B defines a *group-divisible* association scheme on Ω: its non-diagonal classes are

$$\{(\alpha, \beta) \in \Omega \times \Omega : B(\alpha) = B(\beta) \text{ but } \alpha \neq \beta\} \text{ and}$$

$$\{(\alpha, \beta) \in \Omega \times \Omega : B(\alpha) \neq B(\beta)\}.$$

If $\mathcal{P}$ is an association scheme, the set $\mathcal{A}(\mathcal{P})$ of all real linear combinations of its adjacency matrices forms an algebra, called the *Bose–Mesner algebra*. A key theorem for association schemes (see [14, Chapter 17]) is that $\mathcal{A}(\mathcal{P})$ is commutative and hence has a basis $\{S_e : e \in \mathcal{E}\}$ consisting of the matrices of orthogonal projection onto its mutual eigenspaces W_e, for e in some suitable index set $\mathcal{E}$. If the adjacency matrices are A_i for i in I then $|I| = |\mathcal{E}| = \dim \mathcal{A}(\mathcal{P})$, but there is not usually any canonical bijection between I and $\mathcal{E}$. The subspace V_U is always a common eigenspace, with projector $|\Omega|^{-1} J$.

Equations (4) and (1) show that the non-zero adjacency matrices A_F of an orthogonal block structure $\mathcal{F}$ form an association scheme, and the common eigenspaces are the non-zero strata W_F defined by Equation (2). It is convenient to extend the term 'stratum' to all association schemes. If none of the A_F is zero then $I = \mathcal{F} = \mathcal{E}$ and none of the W_F is zero: here there is a natural bijection between the associate classes and the strata.

4 Designs

I take the view, explained in [2], that a design is a function h from one structured set Ω, consisting of the experimental units, to another structured set Θ, consisting of the treatments. The treatment assigned to experimental unit ω is just $h(\omega)$. In this paper, the structures on Ω and Θ are both association schemes.

Information about the design map can be recorded in the $\Omega \times \Theta$ design matrix X, whose (ω, θ)-entry is equal to 1 if $h(\omega) = \theta$ and to 0 otherwise. If A is the adjacency matrix of a subset Δ of $\Omega \times \Omega$, then the (θ, ϕ)-entry in $X'AX$ is equal to

$$|\{(\alpha, \beta) \in \Delta : h(\alpha) = \theta \text{ and } h(\beta) = \phi\}|.$$

Here X' denotes the transpose of X. In particular, $X'X = X'IX$ is diagonal with (θ, θ)-entry equal to the replication of treatment θ, which is $|h^{-1}(\theta)|$, while the (θ, ϕ)-entry of $X'JX$ is equal to $|h^{-1}(\theta)|\,|h^{-1}(\phi)|$.

Definition
Let $\mathcal{P}$ be an association scheme on Ω with adjacency matrices A_i, for i in I, and let $\mathcal{Q}$ be an association scheme on Θ with adjacency matrices B_j, for j in J. Let $h: \Omega \to \Theta$ be a design with design matrix X. Then h is *partially balanced* for $\mathcal{P}$ with respect to $\mathcal{Q}$ if there are integers λ_{ij} for (i, j) in $I \times J$ such that

$$X'A_iX = \sum_j \lambda_{ij} B_j$$

for all i in I; that is, if θ and ϕ are j-th associates in Θ then there are λ_{ij} pairs of i-th associates α and β in Ω such that $h(\alpha) = \theta$ and $h(\beta) = \phi$.

When $\mathcal{P}$ is group divisible, this definition agrees with the usual definition of a partially balanced block design. In general, the definition is identical to the definition of $(\mathcal{P}, \mathcal{Q})$-balance in Section 5.2 of [20]. However, the usual definition of a balanced block design is more restrictive: a block design is balanced if it is partially balanced, in the above sense, with respect to the trivial association scheme on Θ. It therefore seems less confusing to reserve the unqualified term 'balance' for the case in which $\mathcal{Q}$ is trivial: that is, h is balanced for $\mathcal{P}$ if it is partially balanced for $\mathcal{P}$ with respect to the trivial association scheme on Θ. Such balanced designs are investigated in [6].

If $\mathcal{P}$ is the association scheme defined by an orthogonal block structure then Equation (4) shows that an equivalent definition of partial balance is that there are integers λ^*_{ij} such that $X'R_iX = \sum_j \lambda^*_{ij} B_j$ for all i. Thus Figure 2 shows a design which is partially balanced for the association scheme of the orthogonal block structure in Example 1 (with $b = n = 2$ and $m = 3$) with respect to the group-divisible scheme defined by the partition $A, B \,||\, C, D \,||\, E, F$.

It is usual to use the label 0 to index the diagonal associate class. In a partially balanced design every treatment has replication λ_{00}, so the design is equi-replicate. It is conventional to write r for λ_{00}.

A	C	E
D	F	B

A	D	F
C	E	B

D	E	A
B	C	F

B	D	E
C	F	A

Figure 2: A partially balanced design on the orthogonal block structure in Ex. 1

Given a random vector Y on Ω, a natural assumption is that

$$\mathrm{Cov}(Y) = \sum_j \gamma_j A_j; \tag{10}$$

that is, that $\mathrm{cov}(Y_\alpha, Y_\beta)$ depends only on the associate class containing (α, β). Equation (10) can be reparametrized as

$$\mathrm{Cov}(Y) = \sum_e \xi_e S_e, \tag{11}$$

where S_e are the stratum projectors in $\mathcal{P}$ and ξ_e are the stratum variances.

The other assumption for a linear model for a designed experiment is that

$$\mathbb{E}(Y) = X\tau$$

for some unknown vector τ in $\mathbb{R}^\Theta$. Projection onto the stratum W_e gives

$$\mathbb{E}(S_e Y) = S_e X\tau \text{ and}$$

$$\mathrm{Cov}(S_e Y) = S_e \mathrm{Cov}(Y) S_e' = \xi_e S_e,$$

which is scalar on W_e.

Put $L_e = X' S_e X$, which is called the *information matrix* for stratum W_e. If $x \in \mathrm{Im}\, L_e$ then there is a vector z in $\mathbb{R}^\Theta$ such that $L_e z = x$. Ordinary least-squares theory shows that the best linear unbiased estimator of $\langle x, \tau \rangle$ from $S_e Y$ is $z' X' S_e Y$, whose variance is $z' X' S_e' (\xi_e S_e) S_e X z = \xi_e z' L_e z$. In particular, if x is an eigenvector of L_e with eigenvalue $r\varepsilon$ then this variance is equal to $\xi_e x' x / r\varepsilon$.

In the textbook situation, where $\mathrm{Cov}(Y) = \sigma^2 I$, the variance is $x' x \sigma^2 / r$. The ratio $\sigma^2 \varepsilon / \xi_e$ is called the *efficiency* for x in stratum W_e, while ε, which depends on the design and not on the values of the stratum variances, is called the *efficiency factor* for x in stratum W_e.

Now, S_e is a linear combination of the adjacency matrices A_i, so L_e is a linear combination of the matrices $X' A_i X$. If the design is partially balanced for $\mathcal{P}$ with respect to Q then each of the matrices $X' A_i X$ is in $\mathcal{A}(Q)$, so $L_e \in \mathcal{A}(Q)$. Therefore the strata of Q are (contained in) eigenspaces of L_e. Write ε_{ef} for the efficiency factor for vectors from stratum f (in Q) in stratum W_e (of $\mathcal{P}$). If the strata in Q have projection matrices T_f for f in $\mathcal{F}$ then

$$L_e = r \sum_{f \in \mathcal{F}} \varepsilon_{ef} T_f. \tag{12}$$

The matrices L_e are non-negative definite and sum to rI, so, for each fixed f in $\mathcal{F}$, the efficiency factors ε_{ef} are non-negative and sum to 1. If there is any e such that $\varepsilon_{ef} = 1$ then any contrast $\langle x, \tau \rangle$ with x in $\mathrm{Im}\, T_f$ is estimated only in stratum W_e. Otherwise,

information has to be combined from two or more strata, as described in [20]. If every efficiency factor is equal to 0 or 1 then no combining is needed and the design is said to be *orthogonal*.

Both $\mathcal{P}$ and Q have the one-dimensional stratum labelled U. Moreover,

$$L_U = |\Omega|^{-1} X' J_\Omega X = |\Omega|^{-1} r^2 J_\Theta = r|\Theta|^{-1} J_\Theta = rT_U.$$

Therefore, $\varepsilon_{UU} = 1$, $\varepsilon_{Uf} = 0$ if $f \neq U$ and $\varepsilon_{eU} = 0$ if $e \neq U$.

If the design is balanced, $V_U^\perp$ is the only other stratum in Q. It is convenient to give it no label, and write ε_e for the eigenvalue of L_e on $V_U^\perp$.

Just as for incomplete-block designs, for a more general association scheme $\mathcal{P}$ the $\mathcal{E} \times \mathcal{F}$ table of efficiency factors gives important information about the design. Proposed designs for an experiment are compared on the basis of these tables. In Section 6 onwards, some partially balanced designs and their efficiency factors are given for those association schemes which are not orthogonal block structures but which are plausible for the set of experimental units in a designed experiment, as noted in [3]. First, Section 5 gives some theory which aids subsequent calculations.

5 Composite designs

If $h_1 : \Omega \to \Theta$ and $h_2 : \Theta \to \Psi$ are functions then we can form the composite function $h_2 \circ h_1 : \Omega \to \Psi$, as shown in Figure 3. If h_1 and h_2 are both designs, then so is $h_2 \circ h_1$, and it is natural to call $h_2 \circ h_1$ a *composite* design, although this conflicts with the terminology in [11]. If h_i is equi-replicate with replication r_i for $i = 1, 2$ then $h_2 \circ h_1$ is equi-replicate with replication $r_1 r_2$.

$$\begin{array}{ccccc} \Omega & \xrightarrow{h_1} & \Theta & \xrightarrow{h_2} & \Psi \\ \mathcal{P} & & Q & & \mathcal{R} \end{array}$$

Figure 3: A composite design

Theorem 1
Let $\mathcal{P}$, Q and $\mathcal{R}$ be association schemes on Ω, Θ and Ψ respectively. Let $h_1 : \Omega \to \Theta$ and $h_2 : \Theta \to \Psi$ be designs. If h_1 is partially balanced for $\mathcal{P}$ with respect to Q and h_2 is partially balanced for Q with respect to $\mathcal{R}$ then $h_2 \circ h_1$ is partially balanced for $\mathcal{P}$ with respect to $\mathcal{R}$.

Proof
Let the adjacency matrices for $\mathcal{P}$ be A_i for i in I, for Q be B_j for j in $\mathcal{J}$, and for $\mathcal{R}$ be C_k for k in $\mathcal{K}$. For $i = 1, 2$ let X_i be the design matrix for h_i. There are integers λ_{ij}, for

(i,j) in $I \times J$, and ν_{jk}, for (j,k) in $J \times K$, such that

$$X_1' A_i' X_1 = \sum_j \lambda_{ij} B_j$$

for i in I and

$$X_2' B_j' X_2 = \sum_k \nu_{jk} C_k$$

for j in J. Now, the design matrix for $h_2 \circ h_1$ is $X_1 X_2$, and

$$(X_1X_2)'A_i(X_1X_2) = X_2'X_1'A_iX_1X_2 = X_2' \sum_j \lambda_{ij} B_j X_2 = \sum_k \sum_j \lambda_{ij} \nu_{jk} C_k,$$

for all i in I, and so $h_2 \circ h_1$ is partially balanced for $\mathcal{P}$ with respect to $\mathcal{R}$. ∎

The following theorem gives a partial converse.

Theorem 2

Let $\mathcal{P}$, $\mathcal{Q}$ and $\mathcal{R}$ be association schemes on Ω, Θ and Ψ respectively. Let A_i, for i in I, be the adjacency matrices for $\mathcal{P}$. Let $h_1 : \Omega \to \Theta$ and $h_2 : \Theta \to \Psi$ be designs. If $\{X_1'A_iX_1 : i \in I\}$ spans $\mathcal{A}(\mathcal{Q})$ and $h_2 \circ h_1$ is partially balanced for $\mathcal{P}$ with respect to $\mathcal{R}$ then h_1 is partially balanced for $\mathcal{P}$ with respect to $\mathcal{Q}$ and h_2 is partially balanced for $\mathcal{Q}$ with respect to $\mathcal{R}$.

Proof

If $\{X_1'A_iX_1 : i \in I\}$ spans $\mathcal{A}(\mathcal{Q})$ then $X_1'A_iX_1 \in \mathcal{A}(\mathcal{Q})$ for all i in I and so h_1 is partially balanced for $\mathcal{P}$ with respect to $\mathcal{Q}$. Moreover, if B_j is an adjacency matrix for $\mathcal{Q}$ then $B_j = X_1'MX_1$ for some M in $\mathcal{A}(\mathcal{P})$. If $h_2 \circ h_1$ is partially balanced for $\mathcal{P}$ with respect to $\mathcal{R}$ then $(X_1X_2)'M(X_1X_2) \in \mathcal{A}(\mathcal{R})$; that is, $X_2'B_jX_2 \in \mathcal{A}(\mathcal{R})$. Hence h_2 is partially balanced for $\mathcal{Q}$ with respect to $\mathcal{R}$. ∎

If $\{X_1'A_iX_1 : i \in I\}$ spans $\mathcal{A}(\mathcal{Q})$ then the information matrices for h_1 span $\mathcal{A}(\mathcal{Q})$, so their mutual eigenspaces are precisely the strata in $\mathcal{Q}$. Otherwise there is at least one pair of strata in $\mathcal{Q}$ with the same efficiency factors in every stratum of $\mathcal{P}$. In some sense, a design h_1 in which $\{X_1'A_iX_1 : i \in I\}$ spans $\mathcal{A}(\mathcal{Q})$ has *full rank* with respect to $\mathcal{P}$ and $\mathcal{Q}$.

Theorem 3

Let $\mathcal{P}$, $\mathcal{Q}$ and $\mathcal{R}$ be association schemes on Ω, Θ and Ψ respectively, with stratum projectors S_e for e in $\mathcal{E}$, T_f for f in $\mathcal{F}$, and U_g for g in $\mathcal{G}$ respectively. Let $h_1 : \Omega \to \Theta$ be a partially balanced design for $\mathcal{P}$ with respect to $\mathcal{Q}$ whose efficiency factors are ε_{ef} for (e,f) in $\mathcal{E} \times \mathcal{F}$, and let $h_2 : \Theta \to \Psi$ be a partially balanced design for $\mathcal{Q}$ with respect to $\mathcal{R}$ whose efficiency factors are ε_{fg} for (f,g) in $\mathcal{F} \times \mathcal{G}$. Then the efficiency factors ε_{eg} of $h_2 \circ h_1$ are given by

$$\varepsilon_{eg} = \sum_{f \in \mathcal{F}} \varepsilon_{ef} \varepsilon_{fg}$$

for (e,g) in $\mathcal{E} \times \mathcal{G}$.

Proof

Let r_1 and r_2 be the replications of h_1 and h_2 respectively. Then Equation (12) gives

$$X_1'S_eX_1 = r_1 \sum_{f\in\mathcal{F}} \varepsilon_{ef}T_f \quad \text{and} \quad X_2'T_fX_2 = r_2 \sum_{g\in\mathcal{G}} \varepsilon_{fg}U_g.$$

Hence

$$(X_1X_2)'S_e(X_1X_2) = r_1r_2 \sum_{g\in\mathcal{G}} \left(\sum_{f\in\mathcal{F}} \varepsilon_{ef}\varepsilon_{fg} \right) U_g. \quad \blacksquare$$

A version of Theorem 1 is used in [13] for the multitiered experiments described in [12] to show that if the component designs h_1 and h_2 are generally balanced then so is their composite. For example, $h_2 \circ h_1$ can be a two-phase experiment. In the first phase, treatments Ψ are applied to field plots Θ according to design h_2. In the second phase, the treatments are the produce from Θ, which are allocated to evaluation-occasions Ω according to design h_1. [13] uses Theorem 3 to construct analysis-of-variance tables for the composite designs.

By contrast, we shall use Theorems 2 and 3 in the case that $\mathcal{P}$ is group-divisible. Then h_1 and $h_2 \circ h_1$ are both block designs. Knowledge about block designs will be exploited to deduce properties of h_2.

Thus we now switch notation so that h_2 is the design function h of Section 4, with the associated notation for adjacency matrices and stratum projectors. Meanwhile, h_1 becomes a design function g from Γ to Ω, where Γ has the group-divisible association scheme defined by the orthogonal block structure $\{U,B,E\}$ for some non-trivial uniform partition B of Γ. See Figure 4, which applies to the next two sections.

$$\begin{array}{ccccc} \Gamma & \xrightarrow{g} & \Omega & \xrightarrow{h} & \Theta \\ \{U,B,E\} & & \mathcal{P} & & \mathcal{Q} \end{array}$$

Figure 4: Another composite design

6 Triangular association schemes

If $\mathcal{P}$ is a triangular scheme $\mathrm{T}(n)$ then Ω consists of all unordered pairs from an n-set: two elements of Ω are i-th associates if their intersection has size $2-i$, for $i = 0$, 1, 2. This can happen in an experiment where the treatments are tasks to be carried out by teams of two people playing the same role. It can also happen in half-diallel experiments, where the experimental units consist of all crosses between n parental lines, excluding self-crosses, in situations where the gender of the parent is irrelevant.

A design h on $\mathcal{P}$ can conveniently be shown as a symmetric square with the diagonal missing, as in Figures 5–6. The symbol in row a and column b is $h(\{a,b\})$, the

treatment on the element $\{a,b\}$ of Ω. This square layout also suggests a suitable block design for g: it has n blocks of size $n-1$, and block a contains every pair $\{a,b\}$ with $b \neq a$. Now the composite design $h \circ g$ also has n blocks of size $n-1$; the treatments in block a are the symbols occurring in row a of the square.

The diallel context gives a way of naming the strata for $\mathrm{T}(n)$. They are:

W_0	the one-dimensional space V_U;
W_p	the $(n-1)$-dimensional space for contrasts between parents;
W_q	$(W_0 + W_p)^{\perp}$.

The efficiency factors for g are

$$\begin{array}{lll} \varepsilon_{U0} = 1 & \varepsilon_{Up} = 0 & \varepsilon_{Uq} = 0 \\ \varepsilon_{B0} = 0 & \varepsilon_{Bp} = \dfrac{n-2}{2(n-1)} & \varepsilon_{Bq} = 0 \\ \varepsilon_{E0} = 0 & \varepsilon_{Ep} = \dfrac{n}{2(n-1)} & \varepsilon_{Eq} = 1. \end{array}$$

No two columns are identical, so g has full rank. Therefore, design h is partially balanced for $\mathrm{T}(n)$ with respect to an association scheme Q on Θ if and only if the block design $h \circ g$ is partially balanced with respect to Q. Theorem 3 shows that, for stratum f in Q,

$$\varepsilon_{Bf} = \frac{n-2}{2(n-1)}\varepsilon_{pf} \tag{13}$$

$$\varepsilon_{Ef} = \frac{n}{2(n-1)}\varepsilon_{pf} + \varepsilon_{qf}. \tag{14}$$

In a block design we usually want the efficiency factors ε_{Bf} to be as small as possible. In a design on $\mathrm{T}(n)$, it is plausible that $\xi_p >> \xi_q$, so we also want the efficiency factors ε_{pf} to be as small as possible. Thus a strategy for finding a good design h is to find a good design g' and see if it can be arranged in a symmetric square so that $h \circ g = g'$: not all block designs g' can be so arranged.

Example 2

Figure 5 gives two balanced designs for seven treatments on the association scheme $\mathrm{T}(7)$. The design h is constructed by omitting the main diagonal of a symmetric idempotent Latin square. Its composite design $h \circ g$ is a binary balanced block design with $\varepsilon_B = 1/36$ and $\varepsilon_E = 35/36$. Hence h is balanced with $\varepsilon_p = 1/15$ and $\varepsilon_q = 14/15$, by Equations (13) and (14). Although the design h' is also balanced, its composite design $h' \circ g$ is not binary. Now the composite design has $\varepsilon_B = 2/9$ and $\varepsilon_E = 7/9$ so h' has $\varepsilon_p = 8/15$ and $\varepsilon_q = 7/15$. Thus h is better than h'.

Example 3

Figure 6 shows a design h for 12 treatments $A, \ldots, L$ in T(9). The composite design is a binary incomplete-block design which is partially balanced with respect to the group-divisible association scheme defined by the partition $A,B,C \,||\, D,E,F \,||\, G,H,I \,||\, J,K,L$.

Although this is an orthogonal block structure, we shall label the classes and strata without reference to U and E, to avoid confusion with the labels B and E for the block design g. Label the within-group class (pairs such as $\{A,B\}$) by 1 and the between-group class (pairs such as $\{D,H\}$) by 2. Label the strata so that

$$\begin{aligned} W_0 &= V_U \\ W_g &= \text{the space for contrasts between groups} \\ W_w &= (W_0 + W_g)^{\perp}. \end{aligned}$$

In the composite design, $\lambda_{B1} = 3$, $\lambda_{B2} = 4$ and $\varepsilon_{Bg} = 0$. Theorem 2.2 of [15] shows that the composite design is optimal among binary incomplete-block designs in the sense of maximizing the harmonic mean of the efficiency factors in stratum W_E, counted according to multiplicity. Equations (13)–(14) suggest that h will therefore be a good design for T(9).

The design h is constructed by taking T(9) to consist of unordered pairs of points in the affine plane over GF(3). The letters $A, \ldots, L$ are the twelve lines of the plane, in their four parallel classes. Let π be a permutation of the parallel classes of cycle type 2^2. Any two points a and b in the plane lie on a line ℓ containing a third point c. The line ℓ lies in a parallel class $\mathcal{L}$. Define $h(\{a,b\})$ to be the line through c in parallel class $\pi(\mathcal{L})$. Then row a of the square contains all lines which do not pass through a.

7 Latin-square schemes

Let Ω consist of the n^2 cells of a square array on which there are $s-2$ mutually orthogonal Latin squares of order n, for some s with $2 \leqslant s \leqslant n-1$. Let F_1 be the partition

	B	C	D	E	F	G
B		D	E	F	G	A
C	D		F	G	A	B
D	E	F		A	B	C
E	F	G	A		C	D
F	G	A	B	C		E
G	A	B	C	D	E	

Design h

	A	G	A	E	E	G
A		B	A	B	F	F
G	B		C	B	C	G
A	A	C		D	C	D
E	B	B	D		E	D
E	F	C	C	E		F
G	F	G	D	D	F	

Design h'

Figure 5: Two balanced designs for 7 treatments in T(7)

	A	*B*	*F*	*E*	*I*	*H*	*K*	*L*
A		*C*	*G*	*J*	*E*	*L*	*F*	*H*
B	*C*		*J*	*G*	*K*	*F*	*I*	*E*
F	*G*	*J*		*D*	*L*	*B*	*A*	*I*
E	*J*	*G*	*D*		*A*	*K*	*H*	*B*
I	*E*	*K*	*L*	*A*		*G*	*D*	*C*
H	*L*	*F*	*B*	*K*	*G*		*C*	*D*
K	*F*	*I*	*A*	*H*	*D*	*C*		*J*
L	*H*	*E*	*I*	*B*	*C*	*D*	*J*	

Figure 6: Group-divisible design for 12 treatments in T(9)

of Ω into rows, F_2 the partition of Ω into columns, and, for $i = 3, \ldots, s$, let F_i be the partition of Ω into subsets defined by the letters of square i. Then $\{U, E, F_1, \ldots, F_s\}$ is an orthogonal block structure on Ω. Put

$$\begin{aligned} A_0 &= I \\ A_b &= A_{F_1} + \cdots + A_{F_s} \\ A_c &= J - A_0 - A_b. \end{aligned}$$

Then A_0, A_b and A_c are the adjacency matrices of an association scheme on Ω which is said to have *Latin-square* type $\mathrm{L}(s,n)$. Its strata are

$$\begin{aligned} W_0 &= V_U \\ W_b &= W_{F_1} + \cdots + W_{F_s} \\ W_c &= (W_0 + W_b)^{\perp}. \end{aligned}$$

We are mostly concerned with the case that $s = 2$.

If the plots in a field trial have an $n \times m$ rectangular array, it is usually appropriate to regard them as having the rectangular association scheme $\mathrm{R}(n,m)$, which is the orthogonal block structure whose two non-trivial partitions correspond to the rows and columns. Even if $n = m$ the rectangular scheme may still be appropriate, because the plots may not be square or the columns may be in the direction of ploughing. However, if $m = n$ and the plots are square and cultivation is by hand then $\mathrm{L}(2,n)$ may be appropriate.

A design h on $\mathrm{L}(s,n)$ can obviously be shown in a square array: see Figures 7 and 8. If $s = 2$ the labels $h(\omega)$, for ω in the square array, give all the information. If $s \geqslant 3$ then the letters of the Latin squares must also be shown. The natural choice for the block design g is a square lattice design [34]. It has sn blocks of size n, whose 'treatments' are the elements of Ω in the classes of $F_1, \ldots, F_s$. The composite design $h \circ g$ also has

sn blocks of size n; the treatments in a block are those occurring in a row, or a column, or a letter of a Latin square, in the square array.

The efficiency factors for the lattice design g are

$$\begin{array}{lll} \varepsilon_{U0} = 1 & \varepsilon_{Ub} = 0 & \varepsilon_{Uc} = 0 \\ \varepsilon_{B0} = 0 & \varepsilon_{Bb} = \dfrac{1}{s} & \varepsilon_{Bc} = 0 \\ \varepsilon_{E0} = 0 & \varepsilon_{Eb} = \dfrac{s-1}{s} & \varepsilon_{Ec} = 1. \end{array}$$

Hence g has full rank, so h is partially balanced for $\mathrm{L}(s,n)$ with respect to Q if and only if the block design $h \circ g$ is partially balanced with respect to Q. Moreover,

$$\varepsilon_{Bf} = \frac{1}{s}\varepsilon_{bf} \tag{15}$$

$$\varepsilon_{Ef} = \frac{1}{s}[(s-1)\varepsilon_{bf} + s\varepsilon_{cf}] = 1 - \frac{1}{s}\varepsilon_{bf} \tag{16}$$

for strata f of Q.

For the association scheme $\mathrm{L}(s,n)$ it is plausible that $\xi_b >> \xi_c$, so we want ε_{bf} to be as small as possible for all f. Once again, it appears that h will be a good design if $h \circ g$ is good.

Example 4

Figure 7 shows a design h for treatments $A, \ldots, G$ on $\mathrm{L}(2,4)$. The composite design $h \circ g$ is partially balanced with respect to the group-divisible scheme defined by the partition

$$A,B \parallel C,D \parallel E,F \parallel G,H$$

of Θ. Labelling the strata of the latter scheme as in Example 3, we find that the efficiency factors for $h \circ g$ are

$$\begin{array}{ll} \varepsilon_{Bg} = 0 & \varepsilon_{Bw} = \dfrac{1}{4} \\ \varepsilon_{Eg} = 1 & \varepsilon_{Ew} = \dfrac{3}{4}. \end{array}$$

Equations (15) and (16) show that those for h are

$$\begin{array}{ll} \varepsilon_{bg} = 0 & \varepsilon_{bw} = \dfrac{1}{2} \\ \varepsilon_{cg} = 1 & \varepsilon_{cw} = \dfrac{1}{2}. \end{array}$$

The cyclic block design for eight treatments with initial block $\{0,1,2,4\}$ is more efficient than $h \circ g$, but it cannot be arranged as the rows and columns of a 4×4 square.

A	*C*	*E*	*G*
C	*B*	*G*	*F*
F	*H*	*A*	*D*
H	*E*	*D*	*B*

Figure 7: Design for 8 treatments in L(2,4)

Example 5
Houtman and Speed [20] discuss the design h in Figure 8, originally given by Kshirsagar [23]. They regard the association scheme on the 6×6 square Ω as R(6,6), and show that h is partially balanced for R(6,6) with respect to the association scheme L(2,3) on Θ shown in Figure 9. However, if we regard the association scheme on Ω as L(2,6) then the design h is balanced.

B	*D*	*H*	*G*	*F*	*C*
C	*E*	*G*	*B*	*D*	*I*
E	*F*	*C*	*A*	*G*	*H*
D	*I*	*A*	*H*	*C*	*E*
F	*G*	*I*	*E*	*A*	*B*
A	*H*	*B*	*D*	*I*	*F*

Figure 8: Design h on Ω in Ex. 5; Ω may carry R(6,6) or L(2,6)

A	*B*	*C*
D	*E*	*F*
G	*H*	*I*

Figure 9: Treatment set Θ for Ex. 5; its association scheme may be L(2,3) or trivial

In nested row-column designs the experimental units carry the orthogonal block structure $\underline{b}/\mathrm{R}(n,m)$, which consists of b copies of R(n,m). If $n = m$ then R(n,m) can be replaced by L$(2,n)$. Some nested row-column designs with $n = m$ bear a double interpretation similar to the one in Example 5. A family of such examples consists of the lattice square designs of Yates [35].

8 Pair schemes

In a full diallel experiment without self-crosses, the experimental units are all ordered crosses between n parental lines; that is, the gender of the parent is deemed important. Similarly, an experiment on tasks may need ordered pairs of people if the two people play different roles.

Now the appropriate association scheme is Pair(n), which was introduced by Nair [24] in the context of rectangular lattice designs, called the *square* association scheme in [5] and Pair(n) in [6]. The set Ω consists of all ordered pairs of distinct elements from an n-set, where $n \geqslant 4$. For ω in Ω, if $\omega = (x,y)$ then put $\bar{\omega} = (y,x)$. The associate

classes are defined so that α and β are

0th associates	if $\alpha = \beta$
1st associates	if $\bar{\alpha} = \beta$
2nd associates	if α and β are in the same row or column but $\alpha \neq \beta$
3rd associates	if $\bar{\alpha}$ and β are in the same row or column but $\bar{\alpha} \neq \beta$
4th associates	otherwise.

Call a vector in $\mathbb{R}^\Omega$ *symmetric* if $v_\omega = v_{\bar{\omega}}$ for all ω in Ω, and *antisymmetric* if $v_\omega = -v_{\bar{\omega}}$ for all ω in Ω. Then the strata are as follows.

W_0 = V_U

W_1 = the space of symmetric vectors spanned by row and column contrasts (dimension $n-1$)

W_2 = the space of antisymmetric vectors spanned by row and column contrasts (dimension $n-1$)

W_s = the space of symmetric vectors orthogonal to row and column contrasts (dimension $n(n-3)/2$)

W_a = the space of antisymmetric vectors orthogonal to row and column contrasts (dimension $(n-1)(n-2)/2$)

The stratum projectors are

$$S_0 = \frac{1}{n(n-1)}J$$

$$S_1 = \frac{1}{2(n-2)}[2(I+A_1)+A_2+A_3-\frac{4}{n}J]$$

$$S_2 = \frac{1}{2n}[2(I-A_1)+A_2-A_3]$$

$$S_s = \frac{1}{2(n-2)}[(n-4)(I+A_1)-A_2-A_3+\frac{2}{n-1}J]$$

$$S_a = \frac{1}{2n}[(n-2)(I-A_1)-A_2+A_3].$$

Put $R = I + A_1$. Then R is the relation matrix of the uniform partition B of Ω into mirror-image pairs $\{\omega, \bar{\omega}\}$. Let $\mathcal{R}$ be the group-divisible association scheme on Ω defined by B.

It is reasonable to assume that ξ_1 and ξ_2 are much bigger than ξ_s and ξ_a, so that only W_s and W_a are used for estimation. (There is also a randomization argument for using only these two strata: see Section 12 of [3].) Thus we want efficiency factors in W_1 and W_2 to be as small as possible.

One way to achieve this is to use a unipotent Latin square of order n and omit its main diagonal: recall that a Latin square is unipotent if it has the same letter throughout

	A	B	C	D	E	F	G
A		C	D	E	F	G	B
B	C		E	F	G	A	D
C	D	E		G	A	B	F
D	E	F	G		B	C	A
E	F	G	A	B		D	C
F	G	A	B	C	D		E
G	B	D	F	A	C	E	

Figure 10: Orthogonal balanced design for 7 treatments in Pair(8), obtained from a symmetric unipotent Latin square

	A	B	C	D	E	F
A		C	D	E	F	G
B	C		E	F	G	A
C	D	E		G	A	B
D	E	F	G		B	C
E	F	G	A	B		D
F	G	A	B	C	D	

Figure 11: Balanced design for 7 treatments in Pair(7), obtained from a symmetric idempotent Latin square

its main diagonal. Examples are shown in Figures 10, 12 and 14. Then there are $n-1$ treatments, each replicated n times.

For such a design,

$$X'(2I+A_2)X = X'(2A_1+A_3)X = 2nJ,$$

and $X'JX = n^2J$, so $L_1 = L_2 = 0$. Moreover, $X'IX = nI$, so

$$\begin{aligned} L_s &= \frac{1}{2}X'RX - nT_0 \\ L_a &= nI - \frac{1}{2}X'RX. \end{aligned}$$

(Here T_0 denotes the projector onto stratum V_U in $\mathcal{Q}$.) Therefore, such a design h on Ω is partially balanced for Pair(n) with respect to $\mathcal{Q}$ if and only if it is partially balanced for $\mathcal{R}$ with respect to $\mathcal{Q}$. Moreover, $\varepsilon_{sf} = \varepsilon_{Bf}$ and $\varepsilon_{af} = \varepsilon_{Ef}$ for all strata f in $\mathcal{Q}$.

There are three obvious ways to construct h as a block design for n treatments in $n(n-1)/2$ blocks of size 2. The first is to apply each treatment to both experimental units in each of $n/2$ blocks. Then $X'RX = 2nI$, so $L_a = 0$ and $L_s = n(I - T_0)$. The design is orthogonal and balanced, with all estimation taking place in stratum W_s. A unipotent Latin square gives such a block design if and only if it is symmetric. Such a square exists if and only if n is even. An example with $n = 8$ is in Figure 10.

The second is to have each pair of treatments occurring together in a single block, and each treatment occurring on both experimental units in one block. Then $X'RX = (n+1)I + J$, so the design is balanced with $\varepsilon_B = \varepsilon_s = (n+1)/2n$ and $\varepsilon_E = \varepsilon_a = (n-1)/2n$. Construction of a unipotent Latin square with this property is possible when n is even and 3 does not divide $n-1$. An example with $n = 8$ is in Figure 12.

The third is to divide the $n-1$ treatments into $(n-1)/2$ groups of two and ensure that each pair $\{\omega, \bar{\omega}\}$ is allocated one of these groups. Then the design is orthogonal and group divisible, with contrasts between groups estimated in stratum W_s and contrasts

	A	B	C	D	E	F	G
B		D	E	F	G	A	C
D	E		G	A	B	C	F
F	G	A		C	D	E	B
A	B	C	D		F	G	E
C	D	E	F	G		B	A
E	F	G	A	B	C		D
G	C	F	B	E	A	D	

Figure 12: Balanced design for 7 treatments in Pair(8), obtained from a unipotent Latin square

	A	B	C	D	E	F
B		D	E	F	G	A
D	E		G	A	B	C
F	G	A		C	D	E
A	B	C	D		F	G
C	D	E	F	G		B
E	F	G	A	B	C	

Figure 13: Balanced design for 7 treatments in Pair(7), obtained from an idempotent Latin square

	A	B	C	D	E	F
F		A	B	C	D	E
E	F		A	B	C	D
D	E	F		A	B	C
C	D	E	F		A	B
B	C	D	E	F		A
A	B	C	D	E	F	

Figure 14: Orthogonal group-divisible design for 6 treatments in Pair(7), obtained from a unipotent Latin square

	A	B	C	D	E
F		A	B	C	D
E	F		A	B	C
D	E	F		A	B
C	D	E	F		A
B	C	D	E	F	

Figure 15: Group-divisible design for 6 treatments in Pair(6)

within groups estimated in stratum W_a. A unipotent Latin square with this property exists if and only if n is odd. One construction for odd n is to label the rows and columns of Ω by the integers modulo n, and put $y-x \pmod n$ in cell (x,y). An example with $n=7$ is in Figure 14.

A similar family of three types of design is available for n treatments with replication $n-1$. This time we start with an idempotent Latin square of order n; that is, one in which every letter occurs once on the main diagonal. Omitting the diagonal leaves each letter in all rows except one and in all columns except one; the exceptional row does not meet the exceptional column. Therefore

$$X'(2I+A_2)X = X'(2A_1+A_3)X = 2I+2(n-2)J,$$

$X'IX = (n-1)I$ and $X'JX = (n-1)^2J$, so

$$L_1 = \frac{2}{n-2}(I-T_0),$$

which has efficiency factor $2/(n-1)(n-2)$ on all treatment contrasts. Meanwhile,

$L_2 = 0,$

$$L_s = \frac{1}{2}X'RX - \frac{2}{n-2}I - \frac{n(n-3)}{n-2}T_0$$

and

$$L_a = (n-1)I - \frac{1}{2}X'RX.$$

Once again, the design is partially balanced for Pair(n) with respect to Q if and only if it is partially balanced for $\mathcal{R}$ with respect to Q. This time, $\varepsilon_{sf} = \varepsilon_{Bf} - 2/(n-1)(n-2)$ and $\varepsilon_{af} = \varepsilon_{Ef}$ for all strata f in Q.

If the Latin square is symmetric then $X'RX = 2(n-1)I$ so the design is balanced with $\varepsilon_1 = 2/(n-1)(n-2)$, $\varepsilon_2 = 0$, $\varepsilon_s = n(n-3)/(n-1)(n-2)$ and $\varepsilon_a = 0$. A symmetric idempotent Latin square exists if and only if n is odd. One construction for odd n is to label the rows and columns of Ω by the integers modulo n, and put $x+y \pmod n$ in cell (x,y). An example with $n = 7$ is in Figure 11. A unipotent Latin square can be obtained from this by moving the letter on cell (x,x) to the cell in row x of an additional column and column x of a new row, and putting a new letter on the main diagonal. The design in Figure 10 is obtained in this way from the design in Figure 11.

In the second type of design, each pair of treatments occurs together in a single block. Thus $X'RX = (n-2)I + J$ so the design is balanced with $\varepsilon_1 = 2/(n-1)(n-2)$, $\varepsilon_2 = 0$, $\varepsilon_s = n(n-4)/2(n-1)(n-2)$ and $\varepsilon_a = n/2(n-1)$. If n is odd and not divisible by 3 then such an idempotent Latin square can be constructed by labelling the rows and columns of Ω by the integers modulo n and putting $2x+y \pmod n$ in cell (x,y). An example is in Figure 13. A unipotent Latin square of order $n+1$ can be constructed from this just as in the previous case. Thus the design in Figure 12 is obtained from the design in Figure 13.

For the third type of design, we do not use an idempotent Latin square. If treatments A and B always occur on mirror-image pairs then the row which omits A passes through the same diagonal cell as the column which omits B. Hence

$$X'(2I + A_2)X = 2I + 2(n-2)J$$

but

$$X'(2A_1 + A_3)X = \frac{2}{n-1}X'RX - 2I + 2(n-2)J.$$

Therefore

$$\begin{aligned}
L_1 &= \frac{1}{n-2}\left[\frac{1}{n-1}X'RX - 2T_0\right] &&= \frac{2}{n-2}T_g \\
L_2 &= \frac{1}{n}\left[2I - \frac{1}{n-1}X'RX\right] &&= \frac{2}{n}T_w \\
L_s &= \frac{n(n-3)}{2(n-2)}\left[\frac{1}{n-1}X'RX - 2T_0\right] &&= \frac{n(n-3)}{n-2}T_g \\
L_a &= \frac{(n-2)(n+1)}{2n}\left[2I - \frac{1}{n-1}X'RX\right] &&= \frac{(n-2)(n+1)}{n}T_w.
\end{aligned}$$

The design is group divisible. Such a design can be constructed from the third type of design based on unipotent squares, by simply omitting the final row and column. An example is in Figure 15.

R. A. Bailey, School of Mathematical Sciences, Queen Mary, University of London, Mile End Road, London E1 4NS, UK, r.a.bailey@qmw.ac.uk

References

[1] R. A. Bailey. A unified approach to design of experiments, *Journal of the Royal Statistical Society, Series A*, 144: 214–223, 1981.

[2] R. A. Bailey. Designs: mappings between structured sets, In: *Surveys in Combinatorics, 1989* (ed. J. Siemons), *London Mathematical Society Lecture Note Series*, 141, Cambridge University Press, Cambridge, 1989.

[3] R. A. Bailey. Strata for randomized experiments, *Journal of the Royal Statistical Society, Series B*, 53: 27–78, 1991.

[4] R. A. Bailey. Orthogonal partitions in designed experiments, *Designs, Codes and Cryptography*, 8:45–77, 1996.

[5] R. A. Bailey. Suprema and infima of association schemes, *Discrete Mathematics*, 248:1–16, 2002.

[6] R. A. Bailey. Balanced colourings of strongly regular graphs, submitted to *Discrete Mathematics*.

[7] R. A. Bailey, Cheryl E. Praeger, C. A. Rowley and T. P. Speed. Generalized wreath products of permutation groups, *Proceedings of the London Mathematical Society*, 47:69–82, 1983.

[8] S. Banerjee and S. Kageyama. Methods of constructing nested partially balanced incomplete block designs, *Utilitas Mathematica*, 43:3–6, 1993.

[9] R. C. Bose and K. R. Nair. Partially balanced incomplete block designs, *Sankhyā*, 4:337–372, 1939.

[10] R. C. Bose and T. Shimamoto. Classification and analysis of partially balanced incomplete block designs with two associate classes, *Journal of the American Statistical Association*, 47:151–184, 1952.

[11] G. E. P. Box and K. J. Wilson. On the experimental attainment of optimum conditions, *Journal of the Royal Statistical Society, Series B*, 13:1–45, 1951.

[12] C. J. Brien and R. A. Bailey. Multitiered experiments: I. Design and randomization, in preparation.

[13] C. J. Brien and R. A. Bailey. Multitiered experiments: II. Structure and analysis, in preparation.

[14] P. J. Cameron and J. H. van Lint. *Designs, Graphs, Codes and their Links*, *London Mathematical Society Student Texts*, 22, Cambridge University Press, Cambridge, 1991.

[15] C.-S. Cheng and R. A. Bailey. Optimality of some two-associate-class partially balanced incomplete-block designs, *Annals of Statistics*, 19:1667–1671, 1991.

[16] S. C. Gupta and M. Singh. Partially balanced incomplete block designs with nested rows and columns, *Utilitas Mathematica*, 40:291–302, 1991.

[17] P. R. Halmos. *Finite-Dimensional Vector Spaces*, second edition, Van Nostrand, Princeton, 1958.

[18] R. Homel and J. Robinson. Nested partially balanced incomplete block designs, In: *Proceedings of the First Australian Conference on Combinatorial Mathematics 1972* (eds. J. Wallis and W. D. Wallis), TUNRA, Newcastle, New South Wales, 1972.

[19] R. J. Homel and J. Robinson. Nested partially balanced incomplete block designs, *Sankhyā, Series B*, 37:201–210, 1975.

[20] A. M. Houtman and T. P. Speed. Balance in designed experiments with orthogonal block structure, *Annals of Statistics*, 11:1069–1085, 1983.

[21] O. Kempthorne. Classificatory data structures and associated linear models, In: *Statistics and Probability: Essays in Honor of C. R. Rao* (eds. G. Kallianpur, P. R. Krishnaiah and J. K. Ghosh), North-Holland, Amsterdam, 1982.

[22] O. Kempthorne, G. Zyskind, S. Addelman, T. N. Throckmorton and R. F. White. *Analysis of Variance Procedures*, Aeronautical Research Laboratory, Ohio, Report No. 149, 1961.

[23] A. M. Kshirsagar. On balancing in designs in which heterogeneity is eliminated in two directions, *Calcutta Statistical Association Bulletin*, 7:469–476, 1957.

[24] K. R. Nair. Rectangular lattices and partially balanced incomplete block designs, *Biometrics*, 7:145–154, 1951.

[25] J. A. Nelder. The analysis of randomized experiments with orthogonal block structure. I. Block structure and the null analysis of variance, *Proceedings of the Royal Society of London, Series A*, 283:147–162, 1965.

[26] J. A. Nelder. The analysis of randomized experiments with orthogonal block structure. II. Treatment structure and the general analysis of variance, *Proceedings of the Royal Society of London, Series A*, 283:163–178, 1965.

[27] C. E. Praeger, C. A. Rowley and T. P. Speed. A note on generalised wreath product groups, *Journal of the Australian Mathematical Society, Series A*, 39:415–420, 1985.

[28] S. R. Searle, G. Casella and C. E. McCulloch. *Variance Components*, Wiley, New York, 1992.

[29] T. P. Speed. What is an analysis of variance?, *Annals of Statistics*, 15:885–910, 1987.

[30] T. P. Speed and R. A. Bailey. On a class of association schemes derived from lattices of equivalence relations, In: *Algebraic Structures and Applications* (eds. P. Schultz, C. E. Praeger and R. P. Sullivan), Marcel Dekker, New York, 1982.

[31] T. P. Speed and R. A. Bailey. Factorial dispersion models, *International Statistical Review*, 55:261–277, 1987.

[32] T. N. Throckmorton. *Structures of classification data*, Ph. D. thesis, Ames, Iowa, 1961.

[33] T. Tjur. Analysis of variance models in orthogonal designs, *International Statistical Review*, 52:33–81, 1984.

[34] F. Yates. A new method for arranging variety trials involving a large number of varieties, *Journal of Agricultural Science*, 26:424–455, 1936.

[35] F. Yates. A further note on the arrangement of variety trials: quasi-Latin squares, *Annals of Eugenics*, 7:319–332, 1937.

[36] G. Zyskind. On structure, relation, sigma, and expectation of mean squares, *Sankhyā, Series A*, 24:115–148, 1962.

Ordered Triple Designs and Wreath Products of Groups

Cheryl E. Praeger and Csaba Schneider

Abstract

We explore an interesting connection between a family of incidence structures and wreath products of finite groups.

Keywords: ordered triple designs; permutation groups; wreath products; product action; innately transitive groups; maximal subgroups of symmetric groups

1 Introduction

The problem discussed in this paper arose from a study in [2] of the set of primitive maximal subgroups of a finite symmetric group $\mathrm{Sym}\Omega$ containing a given subgroup of $\mathrm{Sym}\Omega$. Application of group theoretic results, depending on the classification of finite simple groups, reduced the problem of describing one family of such maximal subgroups to a problem concerning a certain kind of incidence structures. We chose this topic because of the unexpected links between several types of mathematical objects.

For a finite set Ω the maximal subgroups of $\mathrm{Sym}\Omega$ may be divided into several disjoint families: intransitive maximal subgroups, imprimitive maximal subgroups, and several families of primitive maximal subgroups; see [6]. A given permutation group G on Ω may be contained in many maximal subgroups of $\mathrm{Sym}\Omega$. The intransitive and imprimitive maximal overgroups of G may be determined from the G-orbits and the G-invariant partitions of Ω. However, determining the primitive overgroups of G is a difficult problem in general. It has been essentially solved in [6] and [9] in the case where G itself is primitive, and even this case required significant use of the finite simple group classification. In [2] we were concerned with a more general situation: the groups G of interest were innately transitive, in other words, they contain a minimal normal subgroup that is transitive. The maximal overgroups of G studied in [2] were wreath products in product action (see Section 3 for the definition of wreath products and product actions). Investigating such overgroups led to a study of certain incidence structures discussed in Section 2. Their connection with overgroups of innately transitive groups is described in more detail in Section 3, and a construction is given in Section 4.

2 Suitable ordered triple designs

Describing and constructing a certain family of overgroups of innately transitive groups required incidence structures of the type introduced in the following definition. Note that a permutation group $H \leqslant \mathrm{Sym}\,\Omega$ acts naturally on the set

$$\Omega^{(3)} = \{(\alpha_1, \alpha_2, \alpha_3) \mid \alpha_1,\ \alpha_2,\ \alpha_3 \text{ are distinct points of } \Omega\}$$

of triples of distinct points of Ω via $h : (\alpha_1, \alpha_2, \alpha_3) \mapsto (\alpha_1^h, \alpha_2^h, \alpha_3^h)$ for all $h \in H$ and $(\alpha_1, \alpha_2, \alpha_3) \in \Omega^{(3)}$. We denote by S_3 the symmetric group on a set of size 3.

Definition 1
(a) An *ordered triple design* $\mathcal{H}$ is a pair $(\Omega, \mathcal{T})$ in which Ω is a finite set, and $\mathcal{T}$ is a subset of $\Omega^{(3)}$, and for each $i \in \{1,2,3\}$ and each $\alpha \in \Omega$, the number of triples in $\mathcal{T}$ containing the point α in position i is independent of α, namely it is $|\mathcal{T}|/|\Omega|$.

(b) An ordered triple design $(\Omega, \mathcal{T})$ is said to be *suitable* if there exists $H \leqslant \mathrm{Sym}\,\Omega$ that leaves $\mathcal{T}$ invariant and is transitive on both Ω and $\mathcal{T}$. For such a group H, the subgroup A of S_3 induced on $\{\alpha_1, \alpha_2, \alpha_3\}$ by the setwise stabiliser $H_{\{\alpha_1,\alpha_2,\alpha_3\}}$ is the same (up to isomorphism) for all triples $(\alpha_1, \alpha_2, \alpha_3) \in \mathcal{T}$. Thus we also say that $(\Omega, \mathcal{T})$ is *A-suitable relative to H*.

(c) If $H \leqslant \mathrm{Sym}\,\Omega$ and $A \leqslant \mathsf{S}_3$, such that H is transitive on Ω, then an H-orbit $\mathcal{T}$ in $\Omega^{(3)}$ is said to be *A-suitable* if, for $(\alpha_1, \alpha_2, \alpha_3) \in \mathcal{T}$, the setwise stabiliser in H of $\{\alpha_1, \alpha_2, \alpha_3\}$ induces a permutation group isomorphic to A on $\{\alpha_1, \alpha_2, \alpha_3\}$.

Definition 1(c) enables us to characterise A-suitable ordered triple designs group theoretically. In Section 3 we explain how ordered triple designs arose in [2], while in Section 4 we show that each suitable ordered triple design arises in relation to our problem.

The proof of the following lemma is easy and is omitted.

Lemma 1
Let H be a transitive permutation group on a finite set Ω, let $\mathcal{T}$ be an H-orbit in $\Omega^{(3)}$, and let $A \leqslant \mathsf{S}_3$. Then $\mathcal{T}$ is A-suitable if and only if $(\Omega, \mathcal{T})$ is an A-suitable ordered triple design relative to H.

The concepts of generously 2-transitive and almost generously 2-transitive permutation groups were introduced by Neumann [7]. In our terminology, a permutation group H acting on Ω is generously 2-transitive if and only if every H-orbit in $\Omega^{(3)}$ is S_3-suitable; and H is almost generously 2-transitive if and only if every H-orbit in $\Omega^{(3)}$ is A_3-suitable or S_3-suitable. It was shown in [7] that each almost generously 2-transitive group is 2-transitive with the single exception of A_3. The classification of 2-transitive groups is a consequence of the finite simple group classification, so the generously and almost generously 2-transitive groups can be regarded as known.

In our construction of innately transitive groups in Section 4, part of the input data is an A-suitable ordered triple design relative to H. It turns out that the structure of

the group that is the result of our construction depends on A. We were interested in examples where A was either a cyclic group of order 2 or the trivial group. Hence the question arose as to how prevalent 1-suitable ordered triple designs might be. For some transitive permutation groups H on a set Ω, every H-orbit in $\Omega^{(3)}$ is 1-suitable. Such groups are characterised in Theorem 4 below. Here a permutation group G on a set Ω is said to be *semiregular* if the only element of G that fixes a point of Ω is the identity element.

Theorem 4
Let H be a transitive permutation group on a finite set Ω. Then all H-orbits in $\Omega^{(3)}$ are 1-suitable if and only if $|H|$ is not divisible by 3 and a Sylow 2-subgroup of H is semiregular.

PROOF. Let us assume that $|H|$ is not divisible by 3, and a Sylow 2-subgroup of H is semiregular. This implies that an element of H with even order has no fixed points in Ω. Let $\mathcal{T}$ be an H-orbit in $\Omega^{(3)}$ and $(\kappa_1, \kappa_2, \kappa_3) \in \mathcal{T}$. Suppose that $g \in H$ and g stabilises the set $\kappa = \{\kappa_1, \kappa_2, \kappa_3\}$, and consider the permutation g' induced by g on κ. Since the order $|g'|$ of g' divides the order of g, and hence divides $|H|$, we have that $|g'| \neq 3$. If $|g'| = 2$ then g has even order and g', and hence also g, fixes one element of the κ_i, which is a contradiction. Hence $g' = 1$ and it follows that $\mathcal{T}$ is 1-suitable.

Suppose now that every H-orbit in $\Omega^{(3)}$ is 1-suitable. If $|H|$ is divisible by 3, then there is an element $g \in H$ of order 3. If $\{\kappa_1, \kappa_2, \kappa_3\}$ is a $\langle g \rangle$-orbit of size 3, then $(\kappa_1, \kappa_2, \kappa_3)^H \subseteq \Omega^{(3)}$ is not 1-suitable. Hence $|H|$ is not divisible by 3. Suppose now that there is a non-identity 2-element g in H that fixes a point $\kappa_1 \in \Omega$. Then g^k is an involution, for some k, g^k fixes κ_1, and if $\{\kappa_2, \kappa_3\}$ is a $\langle g^k \rangle$-orbit in Ω with size 2, then $(\kappa_1, \kappa_2, \kappa_3)^H \subseteq \Omega^{(3)}$ is not 1-suitable. Hence a Sylow 2-subgroup of H is semiregular. □

The family of groups that satisfy the conditions of Theorem 4 contains some primitive and some insoluble examples, though most groups in this family are imprimitive and soluble.

Remark 1
Let H satisfy the conditions of Theorem 4.

(a) The only finite simple groups T for which 3 does not divide $|T|$ are the Suzuki groups $\mathrm{Sz}(q)$, where $q = 2^{2a+1} \geqslant 8$; see pages 8–9 of [5]. So if H is insoluble then the non-abelian composition factors of H are all isomorphic to $\mathrm{Sz}(q)$ for various q. There are certainly some insoluble examples H. For instance if $H = \mathrm{Sz}(q)$, and L is any subgroup of H such that $|L|$ is odd, then the action of H by right multiplication on $\{Lx \mid x \in H\}$ satisfies the conditions of Theorem 4.

(b) A transitive permutation group H is *primitive* on Ω if and only if the stabiliser H_α in H of a point $\alpha \in \Omega$ is a maximal subgroup. The conditions in Theorem 4 are equivalent to requiring $3 \nmid |H|$ and $2 \nmid |H_\alpha|$. Since all maximal subgroups of $\mathrm{Sz}(q)$ have even order (see [11]), none of the examples given in paragraph (a) are primitive.

Indeed, it is not difficult, using the O'Nan–Scott Theorem (see Theorem 4.1A of [4]), to show that if H is primitive and satisfies the conditions of Theorem 4, then H is a semidirect product $N \rtimes L$ where $N \cong \mathbb{Z}_p^d$, with p a prime, $p \neq 3$, and $L \leqslant \mathsf{GL}_d(p)$, such that $\gcd(6, |L|) = 1$ and L leaves no non-trivial, proper subspace of $\mathbb{Z}_p^d$ invariant. If $d = 1$ then any subgroup L with $\gcd(6, |L|) = 1$ gives an example. Also if $d \geqslant 2$ and $p^d - 1$ has a prime divisor $r \geqslant 5$ such that r does not divide $p^a - 1$ for any $a \leqslant d - 1$ then $\mathsf{GL}_d(p)$ contains a cyclic subgroup of order r that satisfies these conditions. Such a prime divisor always exists unless $p^d = 64$ or $d = 2$ and p is of the form $2^a 3^b - 1$ for some a, b; see [12].

Not every transitive group H gives rise to a 1-suitable orbit $\mathcal{T}$. If H is 3-transitive on Ω then the setwise stabiliser of each triple $\{\alpha, \beta, \gamma\}$ induces the symmetric group S_3 on $\{\alpha, \beta, \gamma\}$, and so 3-transitive groups have no 1-suitable orbits on triples. In [8] a transitive permutation group H on a set Ω was defined to be a *three-star group* if for all 3-subsets t of Ω the setwise stabiliser H_t does not fix t pointwise. Thus H is a three-star group if and only if it has no 1-suitable orbit in $\Omega^{(3)}$. Each 3-transitive group is a three-star group, and there are other examples, for example the group $H = \mathsf{PSL}_d(q)$ ($d \geqslant 3$ and q a prime-power) acting on the set Ω of 1-dimensional subspaces of the underlying d-dimensional vector space. An investigation of finite three-star groups by P. M. Neumann and the first author is in progress [8]. It has been shown in particular that primitive three-star groups have rank at most 3, but a complete classification of finite three-star groups is yet to be achieved.

3 Embedding permutation groups into wreath products

Let Γ be a finite set, $L \leqslant \mathsf{Sym}\,\Gamma$, $\ell \geqslant 2$ an integer, and $H \leqslant \mathsf{S}_\ell$. The *wreath product* $L\mathsf{wr}H$ is the semidirect product $L^\ell \rtimes H$ where for $(x_1, \ldots, x_\ell) \in L^\ell$ and $\sigma \in \mathsf{S}_\ell$, $(x_1, \ldots, x_\ell)^\sigma = (x_{1\sigma^{-1}}, \ldots, x_{\ell\sigma^{-1}})$. The product action of $L\mathsf{wr}H$ is the action of $L\mathsf{wr}H$ on Γ^ℓ defined by

$$(\gamma_1, \ldots, \gamma_\ell)^{(x_1, \ldots x_\ell)\sigma} = (\gamma_{1\sigma^{-1}}^{x_{1\sigma^{-1}}}, \ldots, \gamma_{\ell\sigma^{-1}}^{x_{\ell\sigma^{-1}}})$$

for all $(\gamma_1, \ldots, \gamma_\ell) \in \Gamma^\ell$, $(x_1, \ldots, x_\ell)\sigma \in L\mathsf{wr}H$.

The following couple of remarks give a summary of the elementary properties of wreath products and product actions. The interested reader will find the proofs of these comments in Section 2.7 of the book by Dixon and Mortimer [4]. If $\gamma \in \Gamma$ then $(\gamma, \ldots, \gamma) \in \Gamma^\ell$; set $\omega = (\gamma, \ldots, \gamma)$. The stabiliser $(L\mathsf{wr}H)_\omega$ in $L\mathsf{wr}H$ of ω is the subgroup $(L_\gamma)^\ell \rtimes H = L_\gamma \mathsf{wr} H$, where L_γ is the stabiliser of γ in L. (It is easy to see that H normalises $(L_\gamma)^\ell$, and so $(L_\gamma)^\ell \rtimes H$ is indeed a subgroup of $L\mathsf{wr}H$.) The subgroup L^ℓ is normal in $L\mathsf{wr}H$ and is transitive on Γ^ℓ if and only if L is transitive on Γ. Moreover no non-identity element of $L\mathsf{wr}H$ stabilises all points of Γ^ℓ. In other words, the product action of $L\mathsf{wr}H$ on Γ^ℓ is faithful. Therefore $L\mathsf{wr}H$ can be considered as a permutation group on Γ^ℓ.

If $|\Gamma| \geqslant 5$ then $\mathrm{Sym}\,\Gamma \mathrm{wr} \mathrm{S}_\ell$ is a maximal subgroup of $\mathrm{Sym}(\Gamma^\ell)$, and is primitive on Γ^ℓ. The subgroups of a finite symmetric group $\mathrm{Sym}\,\Omega$ of the form $\mathrm{Sym}\,\Gamma \mathrm{wr} \mathrm{S}_\ell$, where Ω can be identified with Γ^ℓ in such a way that $\mathrm{Sym}\,\Gamma \mathrm{wr} \mathrm{S}_\ell$ acts on Γ^ℓ as above, form one of several classes of primitive maximal subgroups of $\mathrm{Sym}\,\Omega$, identified by the O'Nan–Scott Theorem; see [6]. Thus an important part of classifying the primitive maximal subgroups of $\mathrm{Sym}\,\Omega$ containing a given (innately transitive) subgroup G is finding all ways of identifying Ω with a Cartesian product Γ^ℓ with $\ell \geqslant 2$ and $|\Gamma| \geqslant 5$, so that G acts as a subgroup of $\mathrm{Sym}\,\Gamma \mathrm{wr} \mathrm{S}_\ell$ in product action.

For the rest of this section suppose that G is an innately transitive group on a finite set Ω and that M is a non-abelian, transitive, minimal normal subgroup of G. Let $\mathcal{W}_G$ be the set of primitive maximal subgroups W of $\mathrm{Sym}\,\Omega$ such that W is a wreath product in product action and $G \leqslant W$.

Let $W \in \mathcal{W}_G$. Then $W \cong \mathrm{Sym}\,\Gamma \mathrm{wr} \mathrm{S}_\ell$ for some Γ and $\ell \geqslant 2$, and we can identify Ω with the Cartesian product Γ^ℓ. It was proved in [3] that $M \leqslant (\mathrm{Sym}\,\Gamma)^\ell$. Let ω be a fixed element of Ω, say $\omega = (\gamma_1, \ldots, \gamma_\ell)$, and for $i = 1, \ldots, \ell$ let $K_i = M_{\gamma_i}$. It was shown in [3] that the set $\mathcal{K}_\omega(W) = \{K_1, \ldots, K_\ell\}$ is invariant under conjugation by G_ω, the K_i have the same size,

$$\bigcap_{i=1}^{\ell} K_i = M_\omega \quad \text{and} \quad K_i \left(\bigcap_{j \neq i} K_j \right) = M. \tag{1}$$

In general we say that a set $\mathcal{K} = \{K_1, \ldots, K_\ell\}$ of subgroups of M is a *Cartesian system* of subgroups for M if $|K_i| = |K_j|$ for all $i, j \in \{1, \ldots, \ell\}$ and there is some $\omega \in \Omega$ such that (1) holds. Cartesian systems provide a way of identifying the subgroups in $\mathcal{W}_G$ from information internal to G.

Theorem 5 ([3])
For a fixed $\omega \in \Omega$ the map $W \mapsto \mathcal{K}_\omega(W)$ is a bijection between the set $\mathcal{W}_G$ and the set of G_ω-invariant Cartesian systems $\mathcal{K}$ of subgroups for M such that $\bigcap_{K \in \mathcal{K}} K = M_\omega$.

Fix $W \in \mathcal{W}_G$, say $W \cong \mathrm{Sym}\,\Gamma \mathrm{wr} \mathrm{S}_\ell$ for some Γ and $\ell \geqslant 2$, and let $\pi_W : W \to \mathrm{S}_\ell$ be the natural projection. Then $\pi_W(G)$ is a subgroup of S_ℓ. Moreover, since M is transitive on Ω, we have $G = MG_\omega$, and since $M \leqslant (\mathrm{Sym}\,\Gamma)^\ell = \ker \pi_W$, $\pi_W(G) = \pi_W(G_\omega)$. Thus π_W gives rise to an action of G_ω on $\{1, \ldots, \ell\}$. It was proved in [3] that the G_ω-actions on $\{1, \ldots, \ell\}$ and on the Cartesian system $\mathcal{K}_\omega(W)$ are equivalent. It can also be shown that $\pi_W(G)$ has at most 2 orbits in $\{1, \ldots, \ell\}$, and a description of the maximal subgroups $W \in \mathcal{W}_G$ for which $\pi_W(G)$ is intransitive on $\{1, \ldots, \ell\}$ is given in [2]. In this paper we are interested in the remaining case, namely in primitive maximal subgroups $W \in \mathcal{W}_G$ where $\pi_W(G)$ is transitive on $\{1, \ldots, \ell\}$. This is equivalent to requiring G_ω to be transitive on the corresponding Cartesian system $\mathcal{K}_\omega(W)$.

Suppose that $M = T^k$ where T is a finite, non-abelian, simple group and $k \geqslant 1$, and let $\sigma_i : M \to T$ denote the i-th coordinate projection map $(t_1, \ldots, t_k) \mapsto t_i$. Let $W \in \mathcal{W}_G$ and set $\mathcal{K}_\omega(W) = \{K_1, \ldots, K_\ell\}$. The properties of Cartesian systems imply that for all

$i \leqslant k$ and $j \leqslant \ell$

$$\sigma_i(K_j)\left(\bigcap_{j' \neq j} \sigma_i(K_{j'})\right) = T.$$

Thus it is important to understand the following sets of subgroups:

$$\mathcal{F}_i = \{\sigma_i(K_j) \mid j = 1, \ldots, \ell,\ \sigma_i(K_j) \neq T\}.$$

The set $\mathcal{F}_i$ is independent of i up to conjugation by G_ω, in the sense that for all i_1, $i_2 \in \{1, \ldots, k\}$ there is a $g \in G_\omega$ such that $\mathcal{F}_{i_2} = \mathcal{F}_{i_1}^g = \{L^g \mid L \in \mathcal{F}_{i_1}\}$. Moreover, using the finite simple group classification, it was shown in [1] that the number of indices j such that $\sigma_i(K_j) \in \mathcal{F}_i$ is at most 3. It is easy to see that this number is also independent of the choice of ω, and we denote this number by $c(G, W)$. In [2] the study of subgroups $W \in \mathcal{W}_G$ for which $\pi_W(G)$ is transitive is split into several cases corresponding to the value of $c(G, W) \in \{0, 1, 2, 3\}$ and to the group theoretical structure of the Cartesian system elements. In the case when $c(G, W) = 3$ we prove the following theorem.

Theorem 6

Suppose that G is an innately transitive permutation group with a non-abelian, transitive, minimal normal subgroup M, and suppose that $W \in \mathcal{W}_G$ such that $\pi_W(G)$ is transitive. Let $\{K_1, \ldots, K_\ell\}$ be the corresponding Cartesian system $\mathcal{K}_\omega(W)$ for a fixed $\omega \in \Omega$, and suppose that $c(G, W) = 3$. Then the following hold.

(a) The isomorphism type of the simple direct factor T of M and those of the subgroups A, B, and C in $\mathcal{F}_i$ are as in Table 1.

(b) For $j = 1, \ldots, \ell$, $\sigma_1(K_j)' \times \cdots \times \sigma_k(K_j)' \leqslant K_j$ and if T is as in row 1 or row 2 of Table 1 then $\sigma_1(K_j) \times \cdots \times \sigma_k(K_j) = K_j$.

(c) For $i = 1, \ldots, k$ let $a(i)$, $b(i)$, $c(i) \in \{1, \ldots, \ell\}$ be such that $\sigma_i(K_{a(i)}) \cong A$, $\sigma_i(K_{b(i)}) \cong B$, and $\sigma_i(K_{c(i)}) \cong C$, and set $\mathcal{T} = \{(K_{a(i)}, K_{b(i)}, K_{c(i)}) \mid i = 1, \ldots, k\}$. Then $(\mathcal{K}_\omega(W), \mathcal{T})$ is a suitable ordered triple design relative to the faithful action of $\pi_W(G)$ on $\mathcal{K}_\omega(W)$.

PROOF. Suppose that $M = T^k$ for some k and for $i = 1, \ldots, k$ let $\sigma_i : M \to T$ be the i-th coordinate projection defined by $\sigma_i(x_1, \ldots, x_k) = x_i$. Also for i_1, $i_2 \in \{1, \ldots, k\}$ we define $\sigma_{\{i_1, i_2\}} : M \to T \times T$ by $\sigma_{\{i_1, i_2\}}(x_1, \ldots, x_k) = (x_{i_1}, x_{i_2})$.

(a) By (1), $K_j\left(\bigcap_{m \neq j} K_m\right) = M$ for all j, and hence

$$\sigma_i(K_j)\left(\bigcap_{m \neq j} \sigma_i(K_m)\right) = \sigma_i(M) = T \quad \text{for} \quad i = 1, \ldots, k \text{ and } j = 1, \ldots, \ell.$$

	T	A	B	C
1	$\mathsf{Sp}_{4a}(2),\ a \geqslant 2$	$\mathsf{Sp}_{2a}(4).2$	$\mathsf{O}^-_{4a}(2)$	$\mathsf{O}^+_{4a}(2)$
2	$\mathsf{P}\Omega_8(3)$	$\Omega_7(3)$	$\mathbb{Z}_3^6 \rtimes \mathsf{PSL}_4(3)$	$\mathsf{P}\Omega_8(2)$
3	$\mathsf{Sp}_6(2)$	$\mathsf{G}_2(2)$	$\mathsf{O}^-_6(2)$	$\mathsf{O}^+_6(2)$
		$\mathsf{G}_2(2)'$	$\mathsf{O}^-_6(2)$	$\mathsf{O}^+_6(2)$
		$\mathsf{G}_2(2)$	$\mathsf{O}^-_6(2)'$	$\mathsf{O}^+_6(2)$
		$\mathsf{G}_2(2)$	$\mathsf{O}^-_6(2)$	$\mathsf{O}^+_6(2)'$

Table 1: Strong multiple factorisations $\{A,B,C\}$ of finite simple groups T

Thus if $\mathcal{F}_i = \{A,B,C\}$ for some i then the set $\{A,B,C\}$ is a strong multiple factorisation of T (see [1] for definitions), and, using [1, Table V], we obtain that T, A, B, and C must be as in one of the lines of Table 1.

(b) Suppose that $\mathcal{F}_1 = \{A,B,C\}$ for some subgroups A, B, and C of T. We see from Table 1 that A', B', and C' are perfect groups. Moreover for all $i \in \{1,\ldots,k\}$ we have that either $\sigma_i(K_j) = T$, or $\sigma_i(K_j)$ is isomorphic to one of A, B, and C. We show next that

$$\sigma_1(K_i)' \times \cdots \times \sigma_k(K_i)' \leqslant K_i \quad \text{for} \quad i = 1,\ldots,\ell.$$

Since $\sigma_i(K_j)'$ is a perfect group, for all i and j, it follows from [10, Lemma 3.2] that we only have to prove that

$$\sigma_{i_1}(K_j)' \times \sigma_{i_2}(K_j)' \leqslant \sigma_{\{i_1,i_2\}}(K_j) \quad \text{for} \quad j = 1,\ldots,\ell \text{ and } i_1,\ i_2 \in \{1,\ldots,k\}. \tag{2}$$

Suppose that $j \in \{1,\ldots,\ell\}$ and $i_1,\ i_2 \in \{1,\ldots,k\}$ are such that (2) does not hold. If $\sigma_{i_1}(K_j) = T$ or $\sigma_{i_2}(K_j) = T$ then [10, Lemma 2.2] implies that $\sigma_{\{i_1,i_2\}}(K_j)$ is a diagonal subgroup of $T \times T$ isomorphic to T. Suppose that this is the case. By assumption, there are $j_1,\ j_2,\ j_3,\ j_4 \in \{1,\ldots,\ell\} \setminus \{j\}$ such that $\sigma_{i_1}(K_{j_1}) \cong \sigma_{i_2}(K_{j_2}) \cong A$ and $\sigma_{i_1}(K_{j_3}) \cong \sigma_{i_2}(K_{j_4}) \cong B$. Now $K_j(K_{j_1} \cap K_{j_2} \cap K_{j_3} \cap K_{j_4}) = M$, and so applying the projection $\sigma_{\{i_1,i_2\}}$ gives

$$T \times T = \sigma_{\{i_1,i_2\}}(K_j)\sigma_{\{i_1,i_2\}}(K_{j_1} \cap K_{j_2} \cap K_{j_3} \cap K_{j_4}).$$

On the other hand,

$$\sigma_{\{i_1,i_2\}}(K_{j_1} \cap K_{j_2} \cap K_{j_3} \cap K_{j_4}) \leqslant (\sigma_{i_1}(K_{j_1}) \cap \sigma_{i_1}(K_{j_3})) \times (\sigma_{i_2}(K_{j_2}) \cap \sigma_{i_2}(K_{j_4}))$$

and so we obtain

$$T \times T = \sigma_{\{i_1,i_2\}}(K_j)\left((\sigma_{i_1}(K_{j_1}) \cap \sigma_{i_1}(K_{j_3})) \times (\sigma_{i_2}(K_{j_2}) \cap \sigma_{i_2}(K_{j_4}))\right). \tag{3}$$

Note that $(\sigma_{i_1}(K_{j_1}) \cap \sigma_{i_1}(K_{j_3})) \times (\sigma_{i_2}(K_{j_2}) \cap \sigma_{i_2}(K_{j_4})) \cong (A \cap B) \times (A \cap B)$. Therefore (3) is a factorisation of the characteristically simple group $T \times T$ in which one factor is a diagonal subgroup and the other factor is the direct product of two isomorphic subgroups. Therefore [10, Theorem 1.5] applies and we find that $\sigma_{i_1}(K_{j_1}) \cap \sigma_{i_1}(K_{j_3})$ has to be a

maximal subgroup of T. On the other hand, $K_{j_1}K_{j_3} = M$ implies $\sigma_{i_1}(K_{j_1})\sigma_{i_1}(K_{j_3}) = T$, and so $\sigma_{i_1}(K_{j_1}) \cap \sigma_{i_1}(K_{j_3})$ is properly contained in $\sigma_{i_1}(K_{j_1})$ and $\sigma_{i_1}(K_{j_3})$. This is a contradiction, and so each of $\sigma_{i_1}(K_j)$, $\sigma_{i_2}(K_j)$ is isomorphic to one of A, B, or C.

Suppose without loss of generality that $\sigma_{i_1}(K_j) \cong A$. Then there are indices j_1, j_2, j_3, $j_4 \in \{1,\dots,\ell\} \setminus \{j\}$ such that $\sigma_{i_1}(K_{j_1}) \cong \sigma_{i_2}(K_{j_2}) \cong B$ and $\sigma_{i_1}(K_{j_3}) \cong \sigma_{i_2}(K_{j_4}) \cong C$. The defining properties of Cartesian systems imply that

$$\{\sigma_{\{i_1,i_2\}}(K_j), \sigma_{i_1}(K_{j_1}) \times \sigma_{i_2}(K_{j_2}), \sigma_{i_1}(K_{j_3}) \times \sigma_{i_2}(K_{j_4})\}$$

is a strong multiple factorisation of $T \times T$, as defined in [10]. However, [10, Theorem 1.7] implies that (2) holds, and we assumed that this was not the case. Thus $\sigma_1(K_i)' \times \cdots \times \sigma_k(K_i)' \leqslant K_i$.

If T is as in row 2 then $\sigma_i(K_j)$ is a perfect group, for all $i \in \{1,\dots,k\}$ and $j \in \{1,\dots,\ell\}$, and so $K_i = \sigma_1(K_i) \times \cdots \times \sigma_k(K_i)$. Let us now suppose that T is as in row 1 and set $\bar{K}_i = \sigma_1(K_i) \times \cdots \times \sigma_k(K_i)$ for all i. Since $\bigcap_i K_i = \bigcap_i \bar{K}_i$ (see [1] page 181), it follows that $\bar{\mathcal{K}} = \{\bar{K}_1,\dots,\bar{K}_\ell\}$ is a Cartesian system of subgroups for M. Therefore

$$\prod_{i=1}^{\ell} |M : K_i| = \left|M : \bigcap_{i=1}^{\ell} K_i\right| = \left|M : \bigcap_{i=1}^{\ell} \bar{K}_i\right| = \prod_{i=1}^{\ell} |M : \bar{K}_i|,$$

which forces $|M : K_i| = |M : \bar{K}_i|$, and hence $K_i = \bar{K}_i$ for all i.

(c) Finally let $\mathcal{T}$ be as in (c), and let us show that $(\mathcal{K}_\omega(W), \mathcal{T})$ is a suitable ordered triple design. It is clear that $\mathcal{T} \subseteq \mathcal{K}_\omega(W)^{(3)}$. Note that G_ω is transitive on $\mathcal{K}_\omega(W)$. Suppose that $T_1,\dots,T_k$ are the simple direct factors of M. If $g \in G_\omega$ and i_1, $i_2 \in \{1,\dots,k\}$ such that $T_{i_1}^g = T_{i_2}$, then $A \cong \sigma_{i_1}\left(K_{a(i_1)}\right) \cong \sigma_{i_1}\left(K_{a(i_1)}\right)^g = \sigma_{i_2}\left(\left(K_{a(i_1)}\right)^g\right)$, and so $K_{a(i_2)} = \left(K_{a(i_1)}\right)^g$. The same argument shows that $K_{b(i_2)} = \left(K_{b(i_1)}\right)^g$ and $K_{c(i_2)} = \left(K_{c(i_1)}\right)^g$. Hence $T_{i_1}^g = T_{i_2}$ implies $\left(K_{a(i_1)}, K_{b(i_1)}, K_{c(i_1)}\right)^g = \left(K_{a(i_2)}, K_{b(i_2)}, K_{c(i_2)}\right)$. For each $t \in \mathcal{T}$ let $I_t = \{T_i \mid (K_{a(i)}, K_{b(i)}, K_{c(i)}) = t\}$. Then $\{I_t \mid t \in \mathcal{T}\}$ is a G_ω-invariant partition of $\{T_1,\dots,T_k\}$ such that the G_ω-actions on $\mathcal{T}$ and $\{I_t \mid t \in \mathcal{T}\}$ are equivalent. Since G_ω is transitive on $\{T_1,\dots,T_k\}$, we obtain that G_ω is transitive on $\mathcal{T}$. Thus $(\mathcal{K}_\omega(W), \mathcal{T})$ is a suitable ordered triple design relative to the group $\pi_W(G_\omega)$ induced by G_ω on $\mathcal{K}_\omega(W)$. Since $\pi_W(G) = \pi_W(G_\omega)$, the proof is complete. □

Thus each $W \in \mathcal{W}_G$ such that $\pi_W(G)$ is transitive and $c(G,W) = 3$ gives rise to a suitable ordered triple design $\mathcal{H}$ relative to $\pi_W(G)$. In addition, for this to occur the simple direct factor T of M and the three subgroups A, B, C such that $A \cap B \cap C = \sigma_i(M_\omega)$ are restricted to those given in one of the rows of Table 1. In the next section we give a construction for such groups G to demonstrate that each T, A, B, C given in Table 1 and each suitable ordered triple design $(\Omega, \mathcal{T})$ relative to a subgroup H of $\mathrm{Sym}\,\Omega$ can occur in Theorem 6. The groups will be wreath products as defined above.

4 The construction

Let T, A, B, C be as in one of the rows of Table 1, let $(\mathcal{K}, \mathcal{T})$ be a suitable ordered triple design relative to a subgroup H of $\mathrm{Sym}\,\mathcal{K}$, and set $\ell = |\mathcal{K}|$ and $k = |\mathcal{T}|$. We may assume without loss of generality that $\mathcal{K} = \{1, \ldots, \ell\}$. Set $\Delta = \{(A \cap B \cap C)x \mid x \in T\}$, so T acts transitively on Δ by right multiplication. It follows from Definition 1 that H acts transitively and faithfully on $\mathcal{T}$, and so we can view H as a subgroup of S_k. Consider the wreath product $G = T \,\mathrm{wr}\, H = T^k \rtimes H$ defined with respect to this action of H. Set $\Omega = \Delta^k$. Then G acts on Ω in its product action. Let M denote the normal subgroup T^k of G.

Now we use the properties of $T \,\mathrm{wr}\, H$ discussed in the second paragraph of Section 3. The group G acts faithfully and transitively on Ω, and M is a transitive normal subgroup of G. Moreover, since H is transitive on $\mathcal{T}$, H permutes the k coordinate subgroups of M transitively, and hence M is a minimal normal subgroup of G. Thus G is innately transitive and M is a transitive, minimal normal subgroup of G. Let γ denote the trivial coset $A \cap B \cap C$ in Δ, and set $\omega = (\gamma, \ldots, \gamma)$. Then $\omega \in \Omega$, $M_\omega = (A \cap B \cap C)^k$ and $G_\omega = M_\omega H$.

For each element $i \in \mathcal{K}$ set $K_i = \prod_{j \in \mathcal{T}} K_{ij}$ where

$$K_{ij} = \begin{cases} A & \text{if the first coordinate of } j \text{ is } i; \\ B & \text{if the second coordinate of } j \text{ is } i; \\ C & \text{if the third coordinate of } j \text{ is } i; \\ T & \text{otherwise.} \end{cases}$$

Let $\hat{\mathcal{K}} = \{K_1, \ldots, K_\ell\}$. We claim that $\hat{\mathcal{K}}$ is a G_ω-invariant Cartesian system for M and $\bigcap_i K_i = M_\omega$. Let $\sigma_i : M \to T$ denote the i-th coordinate projection mapping $(x_1, \ldots, x_k) \mapsto x_i$. First note that the K_i are direct products of their projections and for all i,

$$\sigma_i(K_1) \cap \cdots \cap \sigma_i(K_\ell) = A \cap B \cap C.$$

Therefore $K_1 \cap \cdots \cap K_\ell = (A \cap B \cap C)^k = M_\omega$. The choice of A, B, and C implies that for $i = 1, \ldots, k$ and $j = 1, \ldots, \ell$

$$\sigma_i(K_j)\left(\bigcap_{j' \neq j} \sigma_i(K_{j'})\right) = T.$$

Since $\sigma_i(K_j) \leqslant K_j$ for each i, j, it follows that $K_j\left(\bigcap_{j' \neq j} K_{j'}\right) = M$ for all j. Thus (1) holds and $\hat{\mathcal{K}}$ is a Cartesian system for M. Let us prove that the set $\hat{\mathcal{K}}$ is invariant under conjugation by H. Let $i \in \mathcal{K}$ and $g = H$. Then $K_i^g = \prod_{j \in \mathcal{T}} K_{ij}^g$. If $K_{ij} = A$ then $j = (i, i', i'')$ for some i', $i'' \in \mathcal{K}$ and $j^g = (i^g, i'^g, i''^g)$, and so $K_{i^g j^g} = A$. Similarly, if $K_{ij} = B$, C, T then $K_{i^g j^g} = B$, C, T, respectively. Therefore $K_i^g = K_{i^g} \in \hat{\mathcal{K}}$. Hence $\hat{\mathcal{K}}$ is H-invariant, and, since $M_\omega = (A \cap B \cap C)^k \leqslant K_i$ for all $i \in \{1, \ldots, \ell\}$, the set $\hat{\mathcal{K}}$ is invariant under conjugation by $G_\omega = M_\omega H$.

Therefore the conditions of Theorem 5 hold, and there is a wreath product $W \in \mathcal{W}_G$ such that $\hat{\mathcal{K}} = \mathcal{K}_\omega(W)$. It follows from the definition of $\hat{\mathcal{K}}$ that $c(G,W) = 3$. Finally, transitivity of H on $\mathcal{K}$ implies that G_ω is transitive on $\hat{\mathcal{K}} = \mathcal{K}_\omega(W)$, and it follows from our comments above that $\pi_W(G)$ is transitive. Thus all conditions of Theorem 6 hold.

The group G constructed above has the very interesting property that there are two different maximal subgroups in $\mathcal{W}_G$. The first is the subgroup W in the previous paragraph, and it is of the form $W = \mathsf{Sym}\Gamma \mathsf{wr} \mathsf{S}_\ell$ where $|\Gamma| = |M : K_i|$ $(1 \leqslant i \leqslant \ell)$. It follows from the definition of G that G is contained in $\mathsf{Sym}\Delta \mathsf{wr} \mathsf{S}_k$, and so also $\mathsf{Sym}\Delta \mathsf{wr} \mathsf{S}_k \in \mathcal{W}_G$. These are necessarily different subgroups, for example $c(G,W) = 3$, while $c(G, \mathsf{Sym}\Delta \mathsf{wr} \mathsf{S}_k) = 1$. Thus the set Ω can be identified with both Δ^k and Γ^ℓ, and G preserves both of these Cartesian decompositions of Ω.

Acknowledgements

We are grateful to Peter Neumann for his helpful comments on drafts of this paper, which much improved our exposition. We also acknowledge the support of an Australian Research Council large grant.

Dedication

Exploring links between different areas of mathematics is something the first author valued in her research collaboration with Terry Speed, and we both wish Terry a happy 60th birthday and many more years of productive mathematical research.

Cheryl E. Praeger, Department of Mathematics and Statistics, The University of Western Australia, 35 Stirling Highway, 6009 Crawley, Western Australia,
praeger@maths.uwa.edu.au

Csaba Schneider, Department of Mathematics and Statistics, The University of Western Australia, 35 Stirling Highway, 6009 Crawley, Western Australia,
csaba@maths.uwa.edu.au

References

[1] R. W. Baddeley and C. E. Praeger. On classifying all full factorisations and multiple-factorisations of the finite almost simple groups. *Journal of Algebra* 204:129–187, 1998.

[2] R. W. Baddeley, C. E. Praeger, and C. Schneider. Permutation groups and Cartesian decompositions. Version 2.0, 2001, in preparation.

[3] R. W. Baddeley, C. E. Praeger, and C. Schneider. Recognizing Cartesian decompositions for innately transitive groups. 2002, in preparation.

[4] J. D. Dixon and B. Mortimer. *Permutation groups*. Springer-Verlag, New York, 1996.

[5] D. Gorenstein, R. Lyons and R. Solomon. *The classification of the finite simple groups*. American Mathematical Society, Providence, RI, 1994.

[6] M. W. Liebeck, C. E. Praeger, and J. Saxl. A classification of the maximal subgroups of the finite alternating and symmetric groups. *Journal of Algebra* 111:365–383, 1987.

[7] P. M. Neumann. Generosity and characters of multiply transitive permutation groups. *Proceedings of the London Mathematical Society (3)* 31:457–481, 1975.

[8] P. M. Neumann and C. E. Praeger. Three-star permutation groups. Version of 17 February 2002, in progress.

[9] C. E. Praeger. The inclusion problem for finite primitive permutation groups. *Proceedings of the London Mathematical Society (3)* 60:68–88, 1990.

[10] C. E. Praeger and C. Schneider. Factorisations of characteristically simple groups. *Journal of Algebra*, to appear.

[11] M. Suzuki. On a class of doubly transitive groups. *Annals of Mathematics (2)*, 75:105–145, 1962.

[12] K. Zsigmondy. Zur Theorie der Potenzreste. *Monatshefte fur Mathematik und Physik*, 3:265–284, 1892.

Pearson's Goodness of Fit Statistic as a Score Test Statistic

Gordon K. Smyth

Abstract

For any generalized linear model, the Pearson goodness of fit statistic is the score test statistic for testing the current model against the saturated model. The relationship between the Pearson statistic and the residual deviance is therefore the relationship between the score test and the likelihood ratio test statistic, and this clarifies the role of the Pearson statistic in generalized linear models. The result is extended to cases in which there are multiple reponse observations for the same combination of explanatory variables.

Keywords: Pearson statistic; score test; chi-square statistic; generalized linear model; exponential family nonlinear model; saturated model

1 Introduction

Goodness of fit tests go back at least to Pearson's (1900) article establishing the asymptotic chi-square distribution for a goodness of fit statistic for the multinomial distribution. Pearson's chi-square statistic includes the test for independence in two-way contingency tables. It has been extended in generalized linear model theory to a test for the adequacy of the current fitted model. Given a generalized linear model with responses y_i, weights w_i, fitted means $\hat{\mu}_i$, variance function $v(\mu)$ and dispersion $\phi = 1$, the Pearson goodness of fit statistic is

$$X^2 = \sum \frac{w_i(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)}$$

[14]. If the fitted model is correct and the observations y_i are approximately normal, then X^2 is approximately distributed as χ^2 on the residual degrees of freedom for the model.

The Pearson goodness of fit statistic X^2 is one of two goodness of fit tests in routine use in generalized linear models, the other being the residual deviance. The residual deviance is the log-likelihood ratio statistic for testing the fitted model against the saturated model in which there is a regression coefficient for every observation. The Pearson statistic is a quadratic form alternative to the residual deviance, and is often preferred over the residual deviance because of its moment estimator character. The

expected value of the Pearson statistic depends only on the first two moments of the distribution of the y_i and in this sense the Pearson statistic is robust against misspecification of the response distribution.

The score test, like the likelihood ratio test, is a general asymptotic parametric test associated with the likelihood function [22]. Score tests are often simpler than likelihood ratio tests because the statistic requires parameter estimators to be obtained only under the null hypothesis. For this reason score tests have been proposed frequently in generalized linear model contexts to test for various sorts of model complications such as overdispersion [3, 5, 7, 13, 19, 24], zero inflation [8], adequacy of the link function [9, 20], or extra terms in the fitted model [1, 2, 4, 19, 21, 26].

While the residual deviance arises from a general inferential principle, namely the likelihood ratio test, the origin of the Pearson statistic has seemed more *ad hoc*. Several authors have noted that score tests for extra terms in the linear predictor give rise to chi-square statistics, but there has been no result for the residual Pearson statistic itself. Pregibon [21] shows, by using one-step estimators, that the score statistic for extra terms in the linear predictor can be expressed as a difference between two chi-square statistics, just as the likelihood ratio test can be obtained as the difference between two residual deviances. Cox and Hinkley [6, Examples 9.17 and 9.21] show that the simplest Pearson statistic, the goodness of fit statistic for the multinomial distribution, can be derived as a score statistic. This article shows that Cox and Hinkley's result for the multinomial extends to all generalized linear models. The Pearson goodness of fit statistic is itself a score test statistic, testing the current model against the saturated model. The relationship between the Pearson statistic and the residual deviance is therefore the relationship between the score test and the likelihood ratio test statistic, and this clarifies the role of the Pearson statistic in generalized linear models.

The result of this article extends to several more general situations. The result extends to data sets with multiple counts in categories and to generalizations of exponential families models, such as overdispersion models, for which there are extra parameters in the variance function. It includes for example as special cases the results on tests for independence in two-way contingency tables of Thall [26] and Paul and Banerjee [19]. The general proofs given here are simpler and more transparent than the special case proofs for contingency tables. Finally, the results given here do not require link-linearity as in generalized linear models, but apply to any exponential family non-linear regression model.

The theory of score tests is revised briefly in Section 2 and the background material required for generalized linear and non-linear models is stated briefly in Section 3. The main results of the article showing the relationship between score tests and goodness of fit are given in Section 4. Section 5 goes on to consider models with extra-dispersion and Section 6 considers estimation of the dispersion parameter.

2 Score tests

This section summarizes briefly the theory of likelihood score tests. Further background on score tests and likelihood ratio tests can be found in Rao [23, pages 417–418] and Cox and Hinkley [6, Section 9.3]. Let $\ell(\mathbf{y};\theta_1,\theta_2)$ be a log-likelihood function depending on a response vector $\mathbf{y}$ and parameter vectors θ_1 and θ_2. We wish to test the composite hypothesis $H_0 : \theta_2 = 0$ against the alternative that θ_2 is unrestricted. The components of θ_1 are so-called nuisance parameters because they are not of interest in the test but values must be estimated for them for a test statistic to be computed. The likelihood score vectors for θ_1 and θ_2 are the partial derivatives

$$\dot{\ell}_1 = \frac{\partial \ell}{\partial \theta_1}$$

and

$$\dot{\ell}_2 = \frac{\partial \ell}{\partial \theta_2}$$

respectively. The observed information matrix for the parameters is $-\ddot{\ell}$ with

$$\ddot{\ell} = \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_2^T} = \begin{pmatrix} \ddot{\ell}_{11} & \ddot{\ell}_{12} \\ \ddot{\ell}_{21} & \ddot{\ell}_{22} \end{pmatrix}.$$

The expected or Fisher information matrix is $I = E(-\ddot{\ell})$, which is partitioned conformally with $\ddot{\ell}$ as

$$I = \begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix}.$$

The score test statistic is based on the fact that the score vector $\dot{\ell}$ has mean zero and covariance matrix I. If the nuisance vector θ_1 is known, then the score test statistic of H_0 is

$$Z = I_{22}^{-1/2} \dot{\ell}_2,$$

where $I_{22}^{1/2}$ stands for any factor such that $I_{22}^{1/2} I_{22}^{T/2} = I_{22}$, or equivalently

$$S = Z^T Z = \dot{\ell}_2^T I_{22}^{-1} \dot{\ell}_2,$$

with ℓ_2 and I_{22} evaluated at $\theta_2 = 0$. The score vector $\dot{\ell}$ is a sum of terms corresponding to individual observations and so is asymptotically normal under standard regularity conditions. It follows that Z is asymptotically a standard normal p_2-vector under the null hypothesis H_0 and that S is asymptotically chi-square distributed on p_2 degrees of freedom, where p_2 is the dimension of θ_2.

If the nuisance parameters are not known, then the score test substitutes for them their maximum likelihood estimators $\hat{\theta}_1$ under the null hypothesis. Setting $\theta_1 = \hat{\theta}_1$ is equivalent to setting $\dot{\ell}_1 = 0$, so we need the asymptotic distribution of $\dot{\ell}_2$ conditional on $\dot{\ell}_1 = 0$, which is normal with mean zero and covariance matrix

$$I_{2.1} = I_{22} - I_{21} I_{11}^{-1} I_{12}.$$

The score test statistic becomes

$$S = \dot{\ell}_2^T I_{2.1}^{-1} \dot{\ell}_2,$$

with $\dot{\ell}_2$ and $I_{2.1}$ evaluated at $\theta_1 = \hat{\theta}_1$ and $\theta_2 = 0$. If $I_{12} = 0$ then θ_1 and θ_2 are said to be orthogonal. In that case, $\dot{\ell}_1$ and $\dot{\ell}_2$ are independent and $I_{2.1} = I_{22}$, meaning that the information matrix I_{22} does not need to be adjusted for estimation of θ_1,

Neyman [15] and Neyman and Scott [16] show that the asymptotic distribution and efficiency of the score statistic S is unchanged if an estimator other than the maximum likelihood estimator is used for the nuisance parameters, provided that the estimator is consistent with convergence rate at least $O(n^{-1/2})$, where n is the number of observations. They show that we can substitute into S any estimator $\tilde{\theta}_1$ of θ_1 for which $\sqrt{n}|\tilde{\theta}_1 - \theta_1|$ is bounded in probability as $n \to \infty$. In that case they rename the score statistic the $C(\alpha)$ test statistic.

3 Generalized Linear Models

Generalized linear models assume that observations are distributed according to a linear exponential family with an additional dispersion parameter. The density or probability mass function for each response is assumed to be of the form

$$f(y;\mu,\phi) = a(y,\phi)\exp[\{y\theta - \kappa(\theta)\}/\phi], \tag{1}$$

where a and κ are suitable known functions. The mean is $\mu = \dot{\kappa}(\theta)$ and the variance is $\phi\ddot{\kappa}(\theta)$. The mean μ and the canonical parameter θ are one-to-one functions of one another. We call ϕ the dispersion parameter and $v(\mu) = \ddot{\kappa}(\theta)$ the variance function.

Following Jørgensen [10, 12], we call the distribution described by (1) an *exponential dispersion model* and denote it $\mathrm{ED}(\mu,\phi)$. If the data $y_1,\ldots,y_n$ are independently distributed as $\mathrm{ED}(\mu,\phi)$, then the sample mean $\bar{y}$ is sufficient for μ and $\bar{y} \sim \mathrm{ED}(\mu,\phi/n)$. More generally, if $y_i \sim \mathrm{ED}(\mu,\phi/w_i)$ where the w_i are known weights, then the weighted sum $\bar{y}_w$ is sufficient for μ and

$$\bar{y}_w = \frac{\sum_{i=1}^n w_i y_i}{\sum_{i=1}^n w_i} \sim \mathrm{ED}\left(\mu, \frac{\phi}{\sum_{i=1}^n w_i}\right).$$

A generalized linear model assumes independent $y_1,\ldots,y_n$ with $y_i \sim \mathrm{ED}(\mu_i,\phi/w_i)$. The means μ_i are assumed to follow a link-linear model

$$g(\mu_i) = \mathbf{x}_i^T\beta, \tag{2}$$

where g is a known monotonic link function, $\mathbf{x}_i$ is a vector of covariates and β is an unknown vector of regression coefficients. Without loss of generality we will assume that the $n \times p$ matrix X with rows $\mathbf{x}_i$ is of full column rank and that $p < n$, where p is the dimension of β.

More generally, we consider generalized nonlinear models in which the mean vector $\mu = (\mu_1,\ldots,\mu_n)^T$ is a general n-dimensional function of the p-vector β. To ensure that

the parametrization is not degenerate, we assume that the gradient matrix $\partial\mu/\partial\beta$ is of full column rank, at least in a neighborhood containing the true value of β and the maximum likelihood estimate $\hat{\beta}$.

This article mainly considers models in which the dispersion is known, $\phi = 1$ say. Most models with discrete responses have known dispersion.

4 Goodness of Fit Tests

Let Ω be the locus of possible values for μ, $\Omega = \{\mu(\beta) : \beta \in \mathbb{R}^p\}$. Let H_0 be the null hypothesis that μ belongs to Ω and let H_a be the alternative hypothesis that μ is unrestricted. The goodness of fit test for the current model tests H_0 against H_a. For a generalized linear model, H_0 is the hypothesis that the μ_i are described by the link-linear model (2).

Theorem 7
The score statistic for the goodness of fit test of a generalized nonlinear model with unit dispersion is the Pearson chi-square statistic

$$S = \sum_{i=1}^{n} w_i(y_i - \hat{\mu}_i)^2/v(\hat{\mu}_i),$$

where $\hat{\mu}_i$ is the expected value μ_i evaluated at the maximum likelihood estimator $\hat{\beta}$.

Proof. There exists a parameter vector β_2 of dimension $n - p$ such that (β, β_2) is a one-to-one transformation of μ in the neighborhood of interest and such that $\beta_2 = 0$ if and only if $\mu \in \Omega$. The goodness of fit test consists of testing $H_0 : \beta_2 = 0$ against the alternative that β_2 is unrestricted. The components of the original parameter vector β are the nuisance parameters for this test. In the generalized linear model case, the implicit parameter vector β_2 can be constructed by finding an $n \times (n - p)$ matrix X_2 such that (X, X_2) is of full rank. Then H_a is the saturated model that $g(\mu_i) = X\beta + X_2\beta_2$ for some β and some β_2.

Let $\dot{\ell}_1$ and $\dot{\ell}_2$ be the score vectors for β and β_2 respectively, and let I be the Fisher information matrix, partitioned into I_{11}, I_{12} and I_{22} as in Section 2. The Fisher information for β_2 adjusted for estimation of β is $I_{2.1}$ and the score statistic for testing H_0 is

$$S = \dot{\ell}_2^T I_{2.1}^{-1} \dot{\ell}_2,$$

where $\dot{\ell}_2$ and $I_{2.1}$ are evaluated at $\beta = \hat{\beta}$ and $\beta_2 = 0$.

Let $V = \text{diag}\{v(\mu_i)/w_i\}$ and write

$$\mathbf{e} = V^{-1/2}(\mathbf{y} - \mu)$$

for the vector of Pearson residuals. Also write

$$U_1 = V^{-1/2}\frac{\partial\mu}{\partial\beta}$$

and

$$U_2 = V^{-1/2}\frac{\partial \mu}{\partial \beta_2}.$$

It is straightforward to show that the score vectors are given by

$$\dot{\ell}_j = U_j^T \mathbf{e}$$

for $j = 1,2$ and the information matrices are given by

$$I_{jk} = U_j^T U_k$$

for $j,k = 1,2$ [25] [27].

Write P_1 for the matrix $P_1 = U_1(U_1^T U_1)^{-1}U_1^T$ of the orthogonal projection onto the column space of U_1. Also write

$$U_{2.1} = (I - P_1)U_2$$

and $P_{2.1}$ for the matrix

$$P_{2.1} = U_{2.1}(U_{2.1}^T U_{2.1})^{-1}U_{2.1}^T$$

of the orthogonal projection onto the column space of $U_{2.1}$. Then P_1 and $P_{2.1}$ project onto orthogonal subspaces and $P_1 + P_{2.1} = I$ since the dimensions of the subspaces add to n.

We can rewrite

$$I_{2.1} = U_2^T U_2 - U_2^T U_1(U_1^T U_1)^{-1}U_1^T U_2 = U_2^T(I - P_1)U_2 = U_{2.1}^T U_{2.1}.$$

We can also rewrite

$$\dot{\ell}_2 = (U_2^T - U_2^T P_1)\mathbf{e} = U_{2.1}^T \mathbf{e},$$

because evaluating at $\beta = \hat{\beta}$ ensures that $U_1^T \mathbf{e} = 0$ and hence $P_1 \mathbf{e} = 0$. This shows that the score statistic is

$$S = \mathbf{e}^T U_{2.1}(U_{2.1}^T U_{2.1})^{-1}U_{2.1}^T \mathbf{e} = \mathbf{e}^T P_{2.1}\mathbf{e} = \mathbf{e}^T(P_1 + P_{2.1})\mathbf{e} = \mathbf{e}^T \mathbf{e}$$

which is the Pearson statistic. □

Example. Theorem 1 shows that the chi-square test for independence in a two-way contingency table is a score statistic, based on the assumption that the counts are independent and Poisson distributed. For multiway contingency tables, Theorem 1 shows that the score test of the hypothesis that any chosen subset of the pairs of faces are independent yields a Pearson statistic.

We now consider the case where there are multiple observations for some or all of the covariate combinations. In such cases it is usually more natural to associate the saturated alternative with unique combinations of the explanatory variables rather than to allow every μ_i to be different. The following corollary to Theorem 1 shows that the score test statistic in such cases is naturally expressed in terms of the mean response for each covariate-combination group. The score statistic in the corollary is the Pearson goodness of fit statistic when the data has been reduced to sufficient statistics for each covariate-combination group.

Corollary 1

Suppose that $y_{ij} \sim \text{ED}(\mu_i, 1/w_{ij})$, $i = 1, \ldots, n$, $j = 1, \ldots, n_i$, *are independent. The score test statistic of* H_0, *that the* μ_i *are functions of* β, *against the alternative* H_a *that they are unrestricted, is given by*

$$S = \sum_{i=1}^{n} w_{i\cdot}(\bar{y}_{wi} - \hat{\mu}_i)^2 / v(\hat{\mu}_i),$$

where $\hat{\mu}_i$ *is the maximum likelihood estimator of* μ_i, $w_{i\cdot}$ *is the sum of weights*

$$w_{i\cdot} = \sum_{j=1}^{n_i} w_{ij}$$

and $\bar{y}_{wi}$ *is the weighted mean*

$$\bar{y}_{wi} = \frac{1}{w_{i\cdot}} \sum_{j=1}^{n_i} w_{ij} y_{ij}.$$

Proof. The weighted means $\bar{y}_{wi}$ are sufficient for the μ_i, and $\bar{y}_i \sim \text{ED}(\mu_i, 1/w_{i\cdot})$. The $\bar{y}_{wi}$ are distributed as for the y_i but with weights $w_{i\cdot}$, so the result follows immediately from Theorem 1. □

Example. Suppose that the y_{ij} are binary responses and that $w_{ij} = 1$ for all i and j. Then

$$S = \sum_{i=1}^{n} n_i(r_i - \hat{p}_i)^2 / v(\hat{p}_i),$$

where r_i is the empirical proportion for the ith covariate-combination group, $\hat{p}_i$ is the estimated probability that $y_{ij} = 1$, and $v(\hat{p}_i) = \hat{p}_i(1 - \hat{p}_i)$. If $y_{i\cdot} = \sum_{j=1}^{n_i} y_{ij}$ is the number of successes for the ith group, then the $y_{i\cdot}$ are binomial(n_i, p_i) and

$$S = \sum_{i=1}^{n} (y_{i\cdot} - \hat{\mu}_i)^2 / v_i(\hat{\mu}_i),$$

with $\mu_i = np_i$ and $v_i(\mu_i) = \mu_i(n_i - \mu_i)/n_i$. This is the Pearson goodness of fit statistic for the data summarized in the usual generalized linear model way as binomial counts for each covariate-combination group.

Example. Paul and Banerjee [19] derive the score test for interaction in a two-way contingency table with multiple counts in each cell. Corollary 1 includes Paul and Banerjee's Theorem 1 as a special case.

5 Extra Parameters in the Variance

Suppose now that there are extra parameters which affect the variance of the y_i, but not the mean, and which are outside the exponential dispersion model framework. Let γ be the vector of extra parameters and let G be the parameter space for γ. Suppose that for each fixed value of γ, the y_i follow a generalized nonlinear model with variance function $\mu \to v(\mu;\gamma)$. The values of γ effectively index a class of generalized nonlinear models. This setup arises frequently when extra parameters are introduced to accommodate overdispersion in generalized linear models [1, 2, 7, 19].

It is straightforward to show that γ and β are orthogonal parameters. This follows because

$$\frac{\partial \ell}{\partial \beta} = \frac{\partial \mu}{\partial \beta} V^{-1}(\mathbf{y} - \mu)$$

and μ does not depend on γ. Therefore, the cross derivative $\partial^2 \ell / \partial \beta \partial \gamma$ will be linear in $\mathbf{y} - \mu$ and will have expectation zero.

Orthogonality of γ and β implies that estimation of γ does not affect the form of the score statistics for goodness of fit. According to the theory of $C(\alpha)$ tests, γ may be replaced in the score test statistics by any estimator which is $O(n^{-1/2})$ consistent without changing the distributional properties of S to first order. This gives the following theorem.

Theorem 8
Suppose that for each $\gamma \in G$, $y_1, \ldots, y_n \sim \mathrm{ED}(\mu_i, 1/w_i)$ *are independent with variance function* $v(\mu;\gamma)$. *The* $C(\alpha)$ *goodness of fit statistic is*

$$S = \sum_{i=1}^{n} w_i (y_i - \hat{\mu}_i)^2 / v(\hat{\mu}_i; \tilde{\gamma}),$$

where $\tilde{\gamma}$ *is any* $\sqrt{n}$*-consistent estimator of* γ *and* $\hat{\mu}_i$ *is the maximum likelihood estimator of* μ_i *given* $\gamma = \tilde{\gamma}$.

Corollary 2
Suppose that for each $\gamma \in G$, $y_{ij} \sim \mathrm{ED}(\mu_i, 1/w_{ij})$, $i = 1, \ldots, n$, $j = 1, \ldots, n_i$, *are independent with variance function* $v(\mu;\gamma)$. *The* $C(\alpha)$ *goodness of fit statistic is*

$$S = \sum_{i=1}^{n} w_{i\cdot} (\bar{y}_{wi} - \hat{\mu}_i)^2 / v(\hat{\mu}_i; \tilde{\gamma}),$$

where $\tilde{\gamma}$ *is any* $\sqrt{n}$*-consistent estimator of* γ, $\hat{\mu}_i$ *is the maximum likelihood estimator of* μ_i *given* $\gamma = \tilde{\gamma}$, *the* $w_{i\cdot}$ *are sums of weights and the* $\bar{y}_{wi}$ *are weighted means.*

The proofs of Theorem 2 and the corollary are similar to the proofs in Section 2.

Example. Suppose that y_{ij} follows a negative binomial distribution with mean μ_i and variance function $V(\mu;c) = \mu + c\mu^2$, $i = 1, \ldots, n$, $j = 1, \ldots, n_i$ for each $c \geq 0$. Suppose

that the μ_i are a function of a vector β of regression parameters. For fixed values of c, the means $\bar{y}_i$ are sufficient for the μ_i and are negative binomial with the same variance function and weights n_i. The $C(\alpha)$ goodness of fit statistic therefore is

$$S = \sum_{i=1}^{n} \frac{n_i(\bar{y}_i - \hat{\mu}_i)^2}{\hat{\mu}_i + \tilde{c}\hat{\mu}_i^2},$$

where $\tilde{c}$ is a $\sqrt{n}$-consistent estimator of c and $\hat{\mu}_i$ is the maximum likelihood estimator of μ_i with $c = \tilde{c}$. This includes Theorem 3 of Paul and Banerjee (1998).

One possible estimator for γ is the maximum likelihood estimator. An alternative estimation method is to solve $S = n - p$ with respect to γ. This estimator is often preferred in overdispersion contexts because it is evidently a consistent estimator based only on the first and second moments of the y_i and therefore has a quasi-likelihood flavor (Breslow, 1990). Obviously, the score statistic S is no longer useful as a goodness of fit statistic if γ is estimated by either of the above methods.

If there are repeat observations for covariate combinations, then an estimate of γ may be obtained from the 'pure error' or within-covariate combination variability. In this approach, γ can be estimated by solving

$$\sum_{i=1}^{n}\sum_{j=1}^{n_i} \frac{w_{ij}(y_{ij} - \bar{y}_{wi})^2}{v(\bar{y}_{wi};\gamma)} = \sum_{i=1}^{n}(n_i - 1).$$

With such an estimator for γ, S still has meaning as a goodness of fit statistic.

6 Unknown Dispersion Parameter

All the above results have assumed that $\phi = 1$. If ϕ is unknown, then both $\dot{\ell}$ and I are divided by ϕ and the score statistic for goodness of fit for a generalized nonlinear model becomes

$$S = \sum_{i=1}^{n} \frac{w_i(y_i - \hat{\mu}_i)^2}{\phi v(\hat{\mu}_i)}.$$

The appearance of the unknown scale parameter ϕ in S means that the statistic is no longer useful for judging goodness of fit. The statistic leads instead, by equating S to its expectation, to the so-called Pearson estimator of ϕ,

$$\tilde{\phi} = \frac{1}{n-p}\sum_{i=1}^{n} \frac{w_i(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)},$$

which is the default estimator of ϕ in generalized linear model functions in the statistical programs Splus and R. Other estimators of ϕ are discussed by Jørgensen [11].

When there are repeat observations, the difference between the full version of the score statistic in Theorem 1 and the reduced form in Corollary 1 can be used to define a 'pure error' estimate of the dispersion parameter ϕ,

$$\tilde{\phi}_{\text{pure}} = \frac{1}{\sum(n_i - 1)} \sum_{i=1}^{n} \sum_{j=1}^{n_i} \frac{w_{ij}(y_{ij} - \bar{y}_{wi})^2}{v(\bar{y}_{wi})}.$$

In the case of normal linear regression, this is the well known 'pure error' estimator of the variance. With the use of this this estimator, the score statistic recovers its use as a goodness of fit statistic, but now as a generalized F-statistic rather than chi-square. Substituting the pure error estimator into the score test for the reduced data gives

$$F = \frac{\sum(n_i - 1)}{n - p} \sum_{i=1}^{n} \frac{w_{i\cdot}(\bar{y}_{wi} - \hat{\mu}_i)^2}{\tilde{\phi}_{\text{pure}} v(\hat{\mu}_i)}.$$

If the y_{ij} are approximately normal, then under the null hypothesis F follows approximately an F-distribution on $n - p$ and $\sum(n_i - 1)$ degrees of freedom. This is asymptotically true for example as the weights $w_{ij} \to \infty$ or the dispersion $\phi \to 0$, because any exponential dispersion model $\text{ED}(\mu, \phi)$ tends to normality as $\phi \to 0$ [11, 12]. The F statistic above is a generalization of the normal theory equivalents, described for example by Weisberg [28, Section 4.3].

Dedication

This article is in honor of Terry Speed, from whom I learned generalized linear models while an undergraduate student in Perth, Western Australia. Terry's enthusiasm for statistics and science was and remains infectious. The topic of this article partly arises from a more recent conversation with Terry.

Gordon K. Smyth, Division of Genetics and Bioinformatics, Walter and Eliza Hall Institute of Medical Research, Melbourne, Australia, `smyth@wehi.edu.au`

References

[1] N. E. Breslow. Score tests in overdispersed generalized linear models. In A. Decarli, B. J. Francis, R. Gilchrist, and G. U. H. Seeber, editors, *Proceedings of GLIM 89 and the 4th International Workshop on Statistical Modelling*, pages 64–74. Springer-Verlag: New York, 1989.

[2] N. E. Breslow. Tests of hypotheses in overdispersed Poisson regression and other quasi-likelihood models. *Journal of the American Statistical Association*, 85:565–571, 1990.

[3] A. Cameron and P. Trivedi. Regression-based tests for overdispersion in the Poisson model. *Journal of Econometrics*, 46:347–364, 1990.

[4] C.-F. Chen. Score tests for regression models. *Journal of the American Statistical Association*, 78:158–161, 1983.

[5] D. R. Cox. Some remarks on overdispersion. *Biometrika*, 70:269–274, 1983.

[6] D. R. Cox and D. V. Hinkley. *Theoretical Statistics*. Chapman and Hall: London, 1974.

[7] C. B. Dean. Testing for overdispersion in Poisson and binomial regression models. *Journal of the American Statistical Association*, 87:451–457, 1992.

[8] D. Deng and S. R. Paul. Score tests for zero inflation in generalized linear models. *Canadian Journal of Statistics*, 28:563–570, 2000.

[9] F. C. Genter and V. T. Farewell. Goodness-of-link testing in ordinal regression models. *Canadian Journal of Statistics*, 13:37–44, 1985.

[10] B. Jørgensen. Exponential dispersion models (with discussion). *Journal of the Royal Statistical Society Series* B, 49:127–162, 1987.

[11] B. Jørgensen. The theory of exponential dispersion models and analysis of deviance. Monografias de Matemática No. 51, Instituto de Mathemática pura e Aplicada, Rio de Janeiro, 1992.

[12] B. Jørgensen. *Theory of Dispersion Models*. Chapman and Hall: London, 1997.

[13] W.-S. Lu. Score tests for overdispersion in Poisson regression models. *Journal of Statistical Computation and Simulation*, 56:213–228, 1997.

[14] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall: London, 1989.

[15] J. Neyman. Optimal asymptotic tests of composite hypotheses. In V. Grenander, editor, *Probability and Statistics: The Harold Cramér Volume*, pages 213–234. Wiley: New York, 1959.

[16] J. Neyman and E. L. Scott. On the use of $C(\alpha)$ optimal tests of composite hypotheses. *Bulletin of the International Statistical Institute, Proceedings of the 35th Session*, 41:477–497, 1966.

[17] K. Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5*, 50:157–175, 1900.

[18] S. R. Paul and D. Deng. Goodness of fit of generalized linear models to sparse data. *Journal of the Royal Statistical Society Series* B, 62:323–333, 2000.

[19] S. R. Paul and T. Banerjee. Analysis of two-way layout of count data involving multiple counts in each cell. *Journal of the American Statistical Association*, 93:1419–1429, 1998.

[20] D. Pregibon. Goodness of link tests for generalized linear models. *Applied Statistics*, 29:15–24, 1980.

[21] D. Pregibon. Score tests in GLIM with applications. In R. Gilchrist, editor, *GLIM 82: Proceedings of the International Conference on Generalised Linear Models*, pages 87–97. Springer-Verlag: New York, 1982.

[22] C. R. Rao. Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Proceedings of the Cambridge Philosophical Society*, 44:50–57, 1947.

[23] C. R. Rao. *Linear Statistical Inference and its Applications*, Second Edition. Wiley: New York, 1973.

[24] P. J. Smith and D. F. Heitjan. Testing and adjusting for departures from nominal dispersion in generalized linear models. *Applied Statistics*, 42:31-41, 1993.

[25] G. K. Smyth. Exponential dispersion models and the Gauss-Newton algorithm. *Australian Journal of Statistics*, 33:57–64, 1991.

[26] P. F. Thall. Score tests in two-way layouts of counts. *Communications in Statistics Part A—Theory and Methods*, 21:3017–3036, 1992.

[27] B.-C. Wei. *Exponential Family Nonlinear Models*. Springer-Verlag: Singapore, 1998.

[28] S. Weisberg. *Applied linear regression*. Wiley: New York, 1985.

A Bayesian Approach to Variable Selection when the Number of Variables is Very Large

Harri T. Kiiveri

Abstract

In this paper, we present a rapid Bayesian variable selection technique which can be used when the number of variables is *much* greater than the number of samples. The method can handle tens of thousands of variables, such as might be measured using biological array technologies. A general formulation is first given, followed by specific details for the class of generalised linear models.

Keywords: Bayesian; Jeffreys hyperprior; posterior; variable selection; EM algorithm; generalised linear models; survival analysis

1 Introduction

Traditional methods of variable selection for statistical models include backward and forward stepwise procedures, and all subsets calculations using branch and bound algorithms, see for example [19]. Typically some criterion such as LAIC or BICE is used to guide the selection process. These stepwise methods have also been implemented in software packages such as R and Splus for more general models than linear regression, *e.g.* generalised linear models.

These traditional methods were implicitly designed for situations where the number of variables is less than the number of observations, and the number of variables was at most of the order of hundreds. Unfortunately, these methods do not cope well with large numbers of variables, say of the order of ten thousand, or when the number of observations is less than the number of variables. In these circumstance they either fail completely, or, even if they can be modified to work, require such a huge computational effort that they are impractical to use.

More recently, Bayesian variable selection methods based on Markov chain Monte Carlo methods have been developed, see for example [4, 13, 21, 22]. These have some attractive properties; however, aside from other issues, these methods are computationally intensive and do not scale up well to problems with ten thousand variables or more.

With the advent of microarray technologies, variable selection problems with ten thousand variables and hundreds of observations are becoming quite common, with the likelihood that the problem sizes will scale up at least one order of magnitude in the near future. Clearly, new methods are required to handle these large problems.

With this background in mind, we present here an automated method for eliminating redundant parameters from statistical models. The general method is presented first, followed by the special case when parameter elimination corresponds to variable selection in generalised linear models. This method can be applied when the number of parameters is much greater than the number of observations as well as in the usual case when the number of parameters is less than the number of observations.

In Section 2 we describe the general algorithm for the situation when there are two sets of parameters β and φ. In this case there is a prior expectation that many components of β are zero but not those of φ. For example, the β might be a large set of parameters such as might occur in a matrix factorisation and the φ might be a scale parameter or a shape parameter.

In Section 3 we consider an important special case of the algorithm, namely generalised linear models, in which a response, discrete or continuous, is explained by a set of covariates. In this case, eliminating (setting to zero) components of β corresponds to selecting relevant covariates or components and discarding the rest.

One application is to biological array data, where each biological array has a response associated with it, such as disease class or a continuous measurement of response to treatment. We seek to find (a small number of) components of the biological array data which explain or predict the response. Another application area is in spectroscopy, where spectra are measured over a large number of wavelengths and it is desired to predict sample properties of interest from the observed spectrum.

In the following, N denotes the number of samples, and vectors such as y, z and μ have components y_i, z_i and μ_i for $i = 1, \ldots, N$. Vector multiplication and division is defined component-wise and $\Delta(\cdot)$ denotes a diagonal matrix whose diagonals are equal to the argument. We also use $||\cdot||$ to denote Euclidean norm.

2 General algorithm for parameter selection

Consider a likelihood for some data y which is a function of a $p \times 1$ parameter vector β, many components of which are *a priori* expected to be zero, and a $q \times 1$ vector of parameters φ (not expected to be zero); note that q could be zero. We want a sparse model representation with as many components of β zero as possible.

The work of Figueiredo [10, 11] can be extended to handle this general problem. Basically, Figueiredo formulated a hierarchical prior for the regression parameters in the standard regression model as well as for the probit regression model for binary data. This prior had a Jeffreys hyperprior and strongly favoured regression parameters being zero. By using the trick of introducing a latent variable, he was able to construct an efficient EM algorithm for maximising the "posterior" distribution of the regression parameters. This posterior had discontinuous derivatives at any point where a component of beta was zero and would have caused problems in maximising the posterior directly. A natural by product of the maximisation was the elimination of redundant variables.

Following Figueiredo, we specify a prior for the parameters β by introducing a $p \times 1$

vector of hyperparameters ν^2. This prior is of the form

$$p(\beta) = \int_{\nu^2} p(\beta|\nu^2)\, p(\nu^2)\, d\nu^2, \tag{1}$$

where $p(\beta|\nu^2)$ is $N(0, \mathrm{diag}\{\nu^2\})$ and $p(\nu^2) \propto \prod_{i=1}^{p} 1/\nu_i^2$ is a Jeffreys prior for ν^2, [16]. We choose an uninformative prior for φ, although the following can be easily modified to include an informative prior. Writing $L(y|\beta\varphi)$ for the likelihood function, in this Bayesian framework the posterior distribution of β, φ and ν given y is

$$p(\beta, \varphi, \nu^2|y) \propto L(y|\beta\varphi)p(\beta|\nu^2)p(\nu^2). \tag{2}$$

By treating ν^2 as a vector of missing data, the EM algorithm [6] may be used to maximise (2) to produce maximum *a posteriori* estimates of β and φ. The prior above is such that the maximum *a posteriori* estimates will tend to be sparse; *i.e.* if a large number of parameters are redundant, many components of β will be zero. The algorithm is stated below.

2.1 EM algorithm for the general problem

To implement the EM Algorithm, we need to perform the so-called *E step* and *M step*. In the following, we start by initialising the algorithm, then perform the *E step*, which provides a function to maximise in the *M step*. Newton-Raphson iterations are used to carry out the *M step*, see [17]. After the *M step*, current values of φ are updated. Parameter values which fall below a threshold during the iterations are eliminated from the model, *i.e.* are fixed at zero.

1. Set $n = 0$, $S_0 = \{1, 2, \ldots, p\}$, initialise $\varphi^{(0)}$, β^* and put $\varepsilon = 10^{-5}$ (say)

2. Define

$$\beta_i^{(n)} = \begin{cases} \beta_i^*, & i \in S_n \\ 0 & \text{otherwise,} \end{cases}$$

and at iteration n, define P_n to be a matrix of zeroes and ones defined from the identity matrix of the same dimension as β by deleting columns corresponding to components of β which are zero. It is easy to see that

$$\begin{aligned} \gamma &= P_n^T \beta \qquad & \beta &= P_n \gamma \\ \gamma^{(n)} &= P_n^T \beta^{(n)} \qquad & \beta^{(n)} &= P_n \gamma^{(n)}, \end{aligned} \tag{3}$$

where the nonzero elements of $\beta^{(n)}$ are $\gamma^{(n)}$.

3. Perform the *E step* by calculating

$$\begin{aligned} Q(\beta|\beta^{(n)},\varphi^{(n)}) &= E\{\log p(\beta,\varphi,\nu|y)|y,\beta^{(n)},\varphi^{(n)}\} \\ &= L(y|\beta,\varphi^{(n)}) - 0.5(||\beta/\beta^{(n)}||^2), \end{aligned} \tag{4}$$

where L is the log likelihood function of y. The expectation is over ν^2. Using $\beta = P_n\gamma$ and $\beta^{(n)} = P_n\gamma^{(n)}$, Equation (4) can be written as

$$Q(\gamma|\gamma^{(n)},\varphi^{(n)}) = L(y|P_n\gamma,\varphi^{(n)}) - 0.5(||\gamma/\gamma^{(n)}||^2). \tag{5}$$

4. Perform the *M step*, which involves finding the maximum of (5) over γ. This can be done with Newton-Raphson iterations as follows. Set $\gamma_0 = \gamma^{(n)}$ and for $r = 0,1,2,\ldots,\gamma_{r+1} = \gamma_r + \alpha_r\delta_r$, where α_r is chosen by a line search algorithm to ensure that $Q(\gamma_{r+1}|\gamma^{(n)},\varphi^{(n)}) > Q(\gamma_r|\gamma^{(n)},\varphi^{(n)})$, and

$$\delta_r = \Delta(\gamma^{(n)})\left[-\Delta(\gamma^{(n)})\frac{\partial^2 L}{\partial^2\gamma_r}\Delta(\gamma^{(n)}) + I\right]^{-1}\left(\Delta(\gamma^{(n)})\frac{\partial L}{\partial\gamma_r} - \frac{\gamma_r}{\gamma^{(n)}}\right), \tag{6}$$

where $\partial L/\partial\gamma_r = P_n'\partial L/\partial\beta_r$, $\partial^2 L/\partial^2\gamma_r = P_n'\partial^2 L/\partial^2\beta_r P_n = P_n'\partial^2 L/\partial^2\beta_r P_n$. Equation (6) is simply the Newton-Raphson algorithm involving the first and second derivatives of (5) with respect to γ after some algebraic manipulation. Note the regularisation of the second derivative matrix induced by the prior.

5. Maximise (5) as a function of φ given the current estimate of β. Let γ^* be the value of γ_r when some convergence criterion is satisfied, *e.g.* $||\gamma_r - \gamma_{r+1}|| < \varepsilon$ (for example 10^{-5}). Define $\beta^* = P_n\gamma^*$, $S_{n+1} = \{i : |\beta_i^*| > \max_j(|\beta_j^*|\varepsilon_1)\}$ where ε_1 is a small constant, say 10^{-5}. The set S_{n+1} identifies variables which are still in the model. Now set $n = n+1$ and choose $\varphi^{(n+1)} = \varphi^{(n)} + \kappa_n(\varphi^* - \varphi^{(n)})$, where φ^* is a (local) maximum which satisfies $\partial/\partial\varphi L(y|P_n\gamma^*,\varphi) = 0$ and κ_n is a damping factor such that $0 < \kappa_n \leq 1$.

6. Check convergence. If $||\gamma^* - \gamma^{(n)}|| < \varepsilon_2$ where ε_2 is suitably small, then stop; otherwise, go to step 2 above.

For the general case, modifications are required if the regularised matrix in (6) is indefinite. The term $\partial^2 L/\partial^2\gamma_r$ in step 4 above can also be replaced by its expectation $E[\partial^2 L/\partial^2\gamma_r]$; we do this in Section 3 below.

2.2 Variable selection in generalised linear models

An important special case of the model and algorithm described above is generalised linear models (GLMs, see [20]). In the notation in the 1985 GLIM System Release manual, a GLM has likelihood function

$$L = \log p(y|\beta,\varphi) = \sum_{i=1}^{N} \left\{ \frac{y_i\theta_i - b(\theta_i)}{a_i(\varphi)} + c(y_i,\varphi) \right\}, \tag{7}$$

where $y = (y_1,\ldots,y_n)^T$ and $a_i(\varphi) = \varphi/w_i$, with the w_i being a fixed set of known weights and φ a single scale parameter. We also have

$$E\{y_i\} = b'(\theta_i) \tag{8}$$

$$\mathrm{Var}\{y_i\} = b''(\theta_i)a(\varphi) = \tau_i^2 a_i(\varphi). \tag{9}$$

Each observation has a set of covariates x_i and a linear predictor $\eta_i = x_i^T\beta$. The relationship between the mean of the i^{th} observation μ_i and its linear predictor is given by the link function $\eta_i = g(\mu_i) = g(b'(\theta_i))$. The inverse of the link is denoted by h, *i.e.* $\mu_i = b'(\theta_i) = h(\eta_i)$. In summary, in addition to the scale parameter, a GLM can be specified by four components:

- the likelihood or (scaled) deviance function
- the link function
- the derivative of the link function
- the variance function.

Some common and well known examples of GLMs are given in table 1.

Table 1: Some examples of common GLMs

Distribution	Link function $g(\mu)$	Derivative of link function	Variance function	Scale parameter
Gaussian	μ	1	1	yes
Binomial	$\log(\mu/(1-\mu))$	$1/(\mu(1-\mu))$	$\mu(1-\mu)/n$	no
Poisson	$\log(\mu)$	$1/\mu$	μ	no
Gamma	$1/\mu$	$-1/\mu^2$	μ^2	yes
Inverse Gaussian	$1/\mu^2$	$-2/\mu^3$	μ^3	yes

For generalised linear models, it can be shown that

$$\frac{\partial L}{\partial \beta} = X^t \left\{ \Delta \left(\frac{1}{\tau_i^2} \frac{\partial \mu_i}{\partial \eta_i} \right) \left(\frac{y_i - \mu_i}{a_i(\varphi)} \right) \right\}, \tag{10}$$

where X is the N by p matrix with i^{th} row x_i^T and

$$E\left\{\frac{\partial^2 L}{\partial\beta^2}\right\} = -E\left\{\frac{\partial L}{\partial\beta}\frac{\partial L}{\partial\beta}^T\right\}. \tag{11}$$

This can be written as

$$\frac{\partial L}{\partial\beta} = X^T V^{-1}(\frac{\partial\eta}{\partial\mu})(y-u) \tag{12}$$

$$E\left\{\frac{\partial^2 L}{\partial\beta^2}\right\} = -X^t V^{-1} X, \tag{13}$$

where $V = \Delta(a_i(\varphi)\tau_i^2(\partial\eta_i/\partial\mu_i)^2)$.

3 EM algorithm for variable selection in GLMs

A description of the EM algorithm follows for the special case of generalized linear models. The algorithm is of the same form as in Section 2, however we give more details regarding the choice of initial value and the calculation of first and second derivatives.

1. Set $n = 0, S_0 = \{1,2,\ldots,p\}$, $\varphi^{(0)}$, and $\varepsilon = 10^{-5}$ (say). If $p \leq N$ compute initial values of β^* by

$$\beta^* = (X^t X + \lambda I)^{-1} X^t g(y+\zeta); \tag{14}$$

if instead $p > N$, then compute initial values of β^* by

$$\beta^* = \frac{1}{\lambda}(I - X^t(X^t X + \lambda I)^{-1})X^t g(y+\xi), \tag{15}$$

where the ridge parameter λ satisfies $0 < \lambda \leq 1$ (say) and ζ is small and chosen so that the link function is well-defined at $y+\zeta$. Cross-validation [14] could be used to estimate λ.

2. Define

$$\beta_i^{(n)} = \begin{cases} \beta_i^*, & i \in S_n \\ 0 & \text{otherwise} \end{cases}$$

and let P_n be a matrix of zeroes and ones such that the nonzero elements $\gamma^{(n)}$ of $\beta^{(n)}$ satisfy

$$\gamma^{(n)} = P_n^T \beta^{(n)}, \qquad \beta^{(n)} = P_n \gamma^{(n)}$$
$$\gamma = P_n^T \beta, \qquad \beta = P_n \gamma.$$

3. Perform the *E step* by calculating

$$\begin{aligned} Q(\beta|\beta^{(n)},\varphi^{(n)}) &= E\{\log p(\beta,\varphi,\nu|y)|y,\beta^{(n)},\varphi^{(n)}\} \\ &= L(y|\beta,\varphi^{(n)}) - 0.5(||\beta/\beta^{(n)}||^2), \end{aligned} \tag{16}$$

where L is the GLM log likelihood function of y. Since $\beta = P_n\gamma$ and $\beta^{(n)} = P_n\gamma^{(n)}$, Equation (16) can be written as

$$Q(\gamma|\gamma^{(n)},\varphi^{(n)}) = L(y|P_n\gamma,\varphi^{(n)}) - 0.5(||\gamma/\gamma^{(n)}||^2) \tag{17}$$

4. Perform the *M step*. This can be done with Newton-Raphson iterations as follows. Set $\gamma_0 = \gamma^{(n)}$; for $r = 0,1,2,\ldots$, $\gamma_{r+1} = \gamma_r + \alpha_r\delta_r$, where α_r is chosen by a line search algorithm to ensure $Q(\gamma_{r+1}|\gamma^{(n)},\varphi^{(n)}) > Q(\gamma_r|\gamma^{(n)},\varphi^{(n)})$. For $p \leq N$, use

$$\delta_r = \Delta(\gamma^{(n)})[Y_n^T V^{-1} Y_n + I]^{-1}(Y_n^T V^{-1} z_r - \frac{\gamma_r}{\gamma^{(n)}}), \tag{18}$$

where

$$\begin{aligned} Y^T &= \Delta(\gamma^{(n)})P_n^T X^T \\ V &= \Delta\left(a_i(\varphi)\tau_i^2(\frac{\partial\eta_i}{\partial\mu_i})^2\right) \\ z &= \frac{\partial\eta}{\partial\mu}(y-\mu) \end{aligned}$$

and the subscript r denotes that these quantities are evaluated at $\mu_r = h(XP_n\gamma_r)$. For $p > N$, use

$$\delta_r = \Delta(\gamma^{(n)})\left[I - Y_n^T(Y_nY_n^T + V_r)^{-1}Y_n\right](Y_n^T V_r^{-1} z_r - \frac{\gamma_r}{\gamma^{(n)}}), \tag{19}$$

with V_r and z_r defined as before.

5. Let γ^* be the value of γ_r when some convergence criterion is satisfied, for example $||\gamma_r - \gamma_{r+1}|| < \varepsilon$ (*e.g.* 10^{-5}). Define $\beta^* = P_n\gamma^*$, $S_{n+1} = \{i : |\beta_i^*| > \max_j(|\beta_j^*|\varepsilon_1)$, where ε_1 is a small constant, say 10^{-5}. Set $n = n+1$ and choose $\varphi^{n+1} = \varphi^n + \kappa_n(\varphi^* - \varphi^n)$, where φ^* satisfies $\partial/\partial\varphi L(y|P_n\gamma^*,\varphi) = 0$ and κ_n is a damping factor such that $0 < \kappa_n \leq 1$. In some cases the scale parameter may be known, or this equation can be solved explicitly to get an updating equation for φ.

6. Check convergence. If $||\gamma^* - \gamma^{(n)}|| < \varepsilon_2$ for ε_2 suitably small, then stop; otherwise, go to step 2 above.

4 Remarks

1. The algorithm can be implemented to be $O(\min(N^3, p^3))$. Differentiation of (5) with respect to γ gives

$$\frac{\partial Q}{\partial \gamma} = \frac{\partial L}{\partial \gamma} - \frac{\gamma}{(\gamma^{(n)})^2}. \tag{20}$$

By the definition of the algorithm in Section 2, $\gamma^{(n+1)}$ is defined so that the left hand side of (20) is zero. Hence, if the sequence $(\gamma^{(n)}, \varphi^{(n)})$ converges, then from

$$(\gamma^{(n)})^2 \left(\frac{\partial L}{\partial \gamma^{(n+1)}} \right) = \gamma^{(n+1)}$$

we can see that redundant parameters which are still in the model but have yet to cross the threshold for omission approach zero at a quadratic rate. This observation is due to Dr. Frank De Hoog (personal communication) and is mirrored in the observed performance of the algorithm.

2. The selection of initial values is important, as values too close to zero can result in the solution $\beta = 0$. It also appears that multiple local maxima exist. The initial value is chosen so as to get a perfect fit to the training data if possible. The algorithm can then be viewed as sequentially throwing out variables which do not affect the fit, or cause the least degradation to the fit.

3. Integrating the prior in (2) over ν we obtain

$$p(\beta) \propto \prod_{j=1}^{p} |\beta_j|^{-1}.$$

Hence, if the likelihood evaluated at $\beta = 0$ is positive, the posterior will be improper. Use of Markov chain Monte Carlo (MCMC) to simulate from such a posterior requires caution, see for example [12].

4. Figueiredo [11] shows that replacing the Jeffreys prior in (1) by the prior

$$p(\nu_i^2|\gamma) = \exp(-\nu_i^2/\gamma)/\gamma$$

gives

$$p(\beta_i|\gamma) \propto \exp(-|\beta_i|\sqrt{2/\gamma}),$$

which is the prior used in the Lasso technique [23]. The algorithms described above have a simple modification to implement this model. Instead of using

$$E\{\nu_i^{-2}|\beta_i\} = \frac{1}{|\beta_i|^2}$$

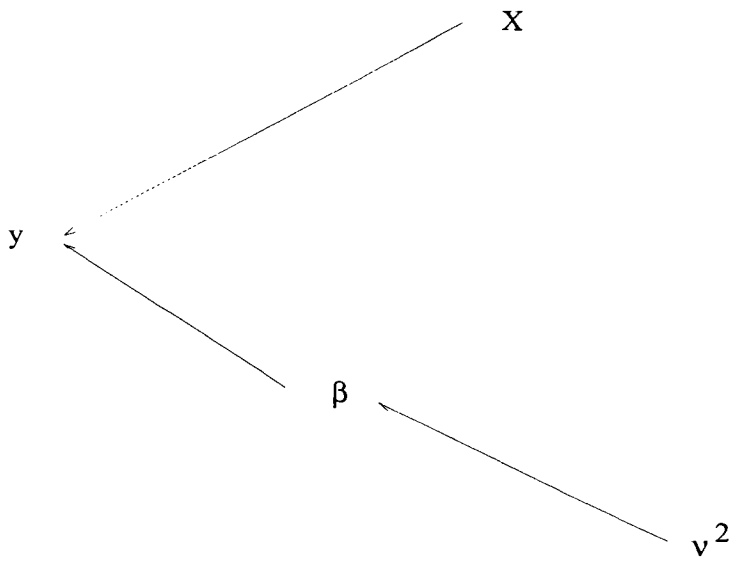


Figure 1: Graphical representation of factorisation of joint density in GLMs

in the *E step* at (4) and (16), use

$$E\{\nu_i^{-2}|\beta_i\} = \frac{1}{|\beta_i|}\sqrt{2/\gamma}.$$

The modifications to the Q functions in (4) and (15) should be clear. This requires the specification of a hyperparameter γ, something which is not required for the Jeffreys prior on ν. It is possible to give a general class of proper priors which includes as a special case the Lasso prior and as a limiting case the model (1) (in preparation).

5. The joint density for the GLMs can be represented graphically as in Figure 1. The *E step* in the EM algorithms described above does not involve y because of the conditional independence of ν^2 and y given β. This means that the algorithm can be applied for a wide variety of different likelihoods.

Another variation is to treat β as missing. With appropriate choice of hyperprior and likelihoods, this treatment gives algorithms for relevance vector machines, see [24]. However, approximations are usually required to do the *E step* since this now depends on y.

6. The algorithm in Section 3 can also be used for quasi-likelihood methods as described in [26] and [18].

7. The matrix X of covariates can be replaced by a matrix K with ij^{th} element k_{ij} and $k_{ij} = \kappa(x_i - x_j)$ for some kernel function κ. This matrix can also be augmented

with a vector of ones. Some possible kernels, including radial basis function kernels, are described in [9]. This treatment opens the possibility of fitting general smooth, as opposed to merely linear, functions of the covariates.

8. Our experience with the algorithm suggests that it is sometimes a little over-enthusiastic in throwing out variables. It is useful to keep a history of variables included in the model as iterations proceed, and to consider sets of variables one or two sets back from the final solution as well. The algorithm can also be used to perform an initial screening of variables for some other procedure by stopping iterations when some subset size is approached *e.g.* 50 variables or when the initial "perfect" fit degrades significantly.

9. By projecting variables not chosen onto the space spanned by a set of chosen variables and then clustering, equivalence classes of important variables can be identified. Alternative solutions can be explored by using a sequence of runs in which the variables chosen in the previous run and those equivalent to them are omitted from consideration in the next run.

5 Examples

In this section, we present examples of the use of these algorithms for some common GLMs and for survival analysis. In each case, we use the version of the algorithm in Section 3.1 with Jeffreys hyperprior (no hyperparameters required). Execution time was typically less than one minute when run in R on a computer with a Pentium III 500 MHz processor and 256 Mb of RAM.

5.1 Standard linear regression model

The algorithm for linear regression is described in [11]. We include it here as an example of a generalised linear model. Consider the sugars data analysed in [3]. The data consist of 125 training observations, where each observation consists of a (transformed) spectrum measured at 700 wavelengths. There is a validation set of 21 observations. The "responses" to be predicted are the percentage composition of three sugars, sucrose, glucose and fructose, in water. We analyse each sugar separately for illustrative purposes here.

The standard regression model is well-known to be a generalised linear model with

- Link function: $g(\mu) = \mu$
- Derivative of link function: $\frac{\partial \eta}{\partial \mu} = 1$
- Variance function: $\tau^2 = 1$
- Scale parameter $\varphi = \sigma^2$
- Deviance (likelihood function): $-\frac{N}{2}\log(\sigma^2) - 0.5\sum_{i=1}^{N}(y_i - \mu_i)^2$

- Updating formula for σ^2 given by

$$(\sigma^2)^{(n+1)} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu_i^*)^2,$$

where μ_i^* is the mean evaluated at β^* in step 5 of the algorithm.

For the linear regression model, we substitute the deviance function defined above for L in (16). Using the above, we can evaluate the terms in (18) as

$$\begin{aligned} Y_n^T &= \Delta(\gamma^{(n)})P_n^T X^T \\ V &= \sigma_n^2 I \\ z_r &= (y - \mu_r) \end{aligned}$$

$$\delta_r = \Delta(\gamma^{(n)})[Y_n^T Y_n + \sigma_n^2 I]^{-1}\left(Y_n^T(y-\mu_r) - \sigma_n^2\frac{\gamma_r}{\gamma^{(n)}}\right). \tag{21}$$

The iterations (21) are basically ridge regressions. An expression for the case when p is greater than N, which involves inversion of a smaller matrix, can be obtained from (19).

For sucrose and glucose, the algorithm in Section 3.1 selected 9 variables (wavelengths), including a constant term. For fructose, the algorithm selected 5 wavelengths with no constant term. The chosen wavelengths in nanometres are given below.

Sucrose 1896 1904 1908 1960 1968 2248 2250 2284

Glucose 1882 1908 1950 1958 1968 2008 2280 2332

Fructose 1908 2082 2254 2256 2330

Results for mean square error (MSE) are given in Table 2.

Table 2: Results on training and validation data

Sugar	Training MSE	Validation MSE
Sucrose	0.10	2.34
Glucose	0.09	0.36
Fructose	0.13	0.38

The mean square error for sucrose on the validation set is much larger than that of the other two sugars. A look at the data suggests that there is a bias in the validation set in the water absorption region of the spectrum as compared to the training data. Deleting

the corresponding wavelengths (1748 to 2498 inclusive) and re-running the algorithm for sucrose reduced the mean square error on the validation set to 1.11 and produced a model with 5 wavelengths, namely 1406, 1756, 1772, 1792, and 2316. Although the validation mean square errors are somewhat larger than those reported in [3], the predictions are quite good and make use of smaller sets of wavelengths than those chosen by the selection method in [3].

5.2 Logistic regression example

We illustrate logistic regression with the data set of [2]. There are $p = 4026$ genes and $N = 36$ samples. In the following, DLBCL refers to *diffuse large B-cell lymphoma*. The samples have been classified into two disease types: GC B-like DLBCL (21 samples) and Activated B-like DLBCL (15 samples). We use this set to illustrate how the above methodology may be used for rapidly identifying genes which are potentially diagnostic of different disease types. The data have been used to define the classes, see [2]; however, we simply use the data set to illustrate the method here.

Logistic regression is a generalised linear model with response y here being the disease class labelled 0 or 1. We also have

- Link function: $g(\mu) = \log(\mu/(1-\mu))$
- Derivative of link function: $1/(\mu(1-\mu))$
- Variance function: $\mu(1-\mu)$
- Scale parameter $\varphi = 1$
- Deviance (likelihood function): $\sum_{i=1}^{N}\{y_i \log(\mu_i) + (1-y_i)\log(1-\mu_i)\}$
- No updating formula is required for the scale parameter.

For the logistic regression model, we substitute the Deviance function defined above for L in (16). Using the above, we can evaluate the terms in (18) as

$$\begin{aligned} Y_n^T &= \Delta(\gamma^{(n)})P_n^T X^T \\ V &= \Delta(\mu_r^{-1}(1-\mu_r)^{-1}) \\ z_r &= \mu_r^{-1}(1-\mu_r)^{-1}(y-\mu_r) \end{aligned}$$

and

$$\delta_r = \Delta(\gamma^{(n)})[Y_n^T \Delta(\mu_r(1-\mu_r)))Y_n + I]^{-1}(Y_n^T(y-\mu_r) - \frac{\gamma_r}{\gamma^{(n)}}). \tag{22}$$

The iterations (22) are once again basically ridge regressions. The algorithm identified 3 relevant genes. The classification accuracy on the training data is given below. This is a much smaller set of genes than the set used by Alizadeh *et al.* [2] to construct the classes.

Table 3: Classification accuracy for the 3 gene logistic regression model

	Predicted class 1	Predicted class 2
True class 1	20	1
True class 2	2	13

5.3 Another logistic regression example

The dataset for this example [15] is available from `http://www-genome.wi.mit.edu/MPR/data_set_ALL_AML.html`. The training data consist of 38 observations with 7129 variables (genes). The validation set contains 34 observations. The response variable is the leukemia class: acute myeloid leukemia (AML) or acute lymphoblastic leukemia (ALL). The data set available over the web includes more genes than those used in the original analysis of [15]. The dataset contains some controls; however, to test the algorithm we include all the data in the data set.

The algorithm retains 4 genes out of the 7129 considered. These are

- 1763 Thymosin beta-4 mRNA
- 1779 MPO Myeloperoxidase
- 2402 Azurocidin gene
- 6201 Interleukin-8 precursor.

This set of genes gives perfect separation of the classes in the training data. An analysis of equivalent sets suggests that gene 6201 can be interchanged with gene 6200, namely Interleukin 8 (IL8) gene. These genes are biologically meaningful in this context.

Table 4 shows results for the selected model applied to the validation set.

Table 4: Validation accuracy for the logistic regression model with 4 selected genes

	Predicted ALL	Predicted AML
True ALL	20	0
True AML	3	11

We also performed an analysis similar to [7] whereby the data was randomly divided into training and test sets in the ratio 2:1. For comparison purposes, we used the 3157 genes used in [7]. The variable selection was run for each training set, and predictions were made for the corresponding test set in a total of 150 runs. All 3157 genes were considered in each run, there was no preselection of genes. The median number of misclassifications observed was 2 with a maximum of 5 and minimum of 0. When

we used a more general prior somewhat between the lasso and (1), the median number of misclassifications was 1.5 with maximum 3 and minimum 0. The mean number of variables chosen was 3. For details see [8].

5.4 Poisson regression example

We use the data set in Section 5.2 to also illustrate gene selection in Poisson regression.

We artificially created a new gene (gene number 1) and a Poisson response for each array with mean given by the expression value of gene 1. This new gene was added to the previous data matrix. Hence, there are 4027 "genes" and 36 samples in this case. The response has a Poisson distribution.

Poisson regression is a generalised linear model with

- Link function: $g(\mu) = log(\mu)$
- Derivative of link function: $1/\mu$
- Variance function: $\tau^2 = \mu$
- Scale parameter $\varphi = 1$
- Deviance (likelihood function): $\sum_{i=1}^{N} \{y_i \log(\mu_i) - \mu_i\}$
- No updating formula is required for the scale parameter.

The algorithm required 5 iterations to correctly identify "gene" 1 as the relevant gene.

5.5 Cox proportional hazards model

We apply a version of the general algorithm in Section 2 to the survival data of Alizadeh *et al.* [2] (available at `http://llmpp.nih.gov/lymphoma/data.shtml`). A parametric version of this can also be fitted as a GLM using a Poisson model, see [1]. In this application, two observations (patients DLCL-0051 and DLCL-0052) are omitted because there is no survival information available for them. The data consist of cDNA microarray measurements on 4026 genes from 40 patients, survival times for each patient and a censoring indicator.

A Cox proportional hazards model [5] is fitted with an initial 4026 explanatory variables (*i.e.* genes) that are rapidly whittled down by the algorithm to just three explanatory genes. The explanatory genes identified by the algorithm are GENE3797X, GENE3302X and GENE356X. These are

- Immunoglobulin heavy chain V(H)5 pseudogene L2-9 transcript
- adenosine deaminase – this is a target for some drugs used to treat lymphoma
- AIM2 – involved with interferon induction and cell fate.

The selected genes are biologically meaningful. More details about this analysis and a simple prognostic indicator can be found in [25].

6 Conclusion

The algorithms described above seem promising in situations where there is little prior knowledge concerning the relationship of a large number of variables to a response of interest. They are fast and can be scaled up to handle *large* numbers of variables. They can also provide a useful screening tool to weed out apparently unimportant parameters or variables prior to an analysis by some other method.

A concern in this context is the production of results which are purely artifacts due to the large number of variables to choose from. Another concern is the influence of individual high dimensional observations when the number of samples is relatively small. As regards the former, permutation tests and the use of validation data sets have confirmed that the results so far are unlikely to be artifacts. In limited testing to date with biological arrays, the algorithms have produced biologically meaningful and apparently new results. A key feature is the consistent identification of smaller sets of variables with performance similar to the larger sets reported by other analyses. A similar statement can be made for spectroscopic data. Concerning the stability of the models selected, leave one out cross-validation calculations have so far demonstrated a high degree of stability in the chosen models. However, more work is required to test these ideas.

We are currently exploring other applications, such as logistic multi-class classification models.

The algorithms and analysis methods described here are protected by patents which are owned by CSIRO.

Acknowledgments

I would like to thank Dr. Frank De Hoog and Professor Phillip Brown for helpful insights and discussions.

Harri T. Kiiveri, CSIRO Mathematical and Information Sciences, The Leeuwin Centre, Floreat, Western Australia, harri.kiiveri@csiro.au

References

[1] M. Aitkin and C. Clayton. The fitting of exponential, Weibull and extreme value distributions to complex censored survival data using GLIM. *Applied Statistics*, 29:156–163, 1980.

[2] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson Jr, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–511, 2000.

[3] P. J. Brown. Measurement, regression and calibration. *Oxford University Press*, 1993.

[4] P. J. Brown, M. Vanucci, and T. Fearn. Bayes model averaging with selection of regressors. *Journal of the Royal Statistical Society B*, 64:519–536, 2002.

[5] D. R. Cox and D. Oakes. Analysis of survival data. *Chapman and Hall, London*, 1984.

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39:1–21, 1977.

[7] S. Dudoit, J. Fridyland, and T. P. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97:77–87, 2002.

[8] R. Dunne. Classification of genes and arrays for microarray data. Technical report, Internal CMIS report, 2001.

[9] T. Evgeniou, M. Pontil, and T. Poggio. A unified framework for regularization networks and support vector machines. *MIT AI memo 1654*, 1999.

[10] M. Figueiredo. Adaptive sparseness using Jeffreys prior. *Neural Information Processing Systems - NIPS ' 2001, Vancouver, December 2001*, 2001.

[11] M. Figueiredo. Unsupervised sparse regression. In *MSRI Workshop on Nonlinear Estimation and Classification, Berkeley, CA*, March 2001. to appear.

[12] I. E. Gelfand and S. K. Sahu. Identifiability, improper priors, and Gibbs sampling for generalised linear models. *Journal of the American Statistical Association*, 94:247–253, 1999.

[13] E. I. George and R. E. McCulloch. Stochastic search variable selection. In *Markov Chain Monte Carlo in Practice*. Chapman and Hall, New York, 1996.

[14] G. H. Golub, M. Heath, and G. Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21:215–223, 1979.

[15] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

[16] S. Kotz and N. L. Johnson. Encyclopedia of statistical sciences. *Wiley, New York*, 4:639, 1983.

[17] D. G. Luenberger. Introduction to linear and nonlinear programming. *Addison-Wesley, Reading, Massachusetts*, 1973.

[18] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, second edition, 1989.

[19] A. J. Miller. *Subset selection in Regression*. Chapman and Hall, New York, 1990.

[20] J. A. Nelder and R. W. M. Weddeburn. Generalised linear models. *Journal of the Royal Statistical Society A*, 135:370–384, 1972.

[21] D. B. Phillips and A. F. M. Smith. Bayesian model comparison via jump diffusions. In *Markov Chain Monte Carlo in Practice*. Chapman and Hall, New York, 1996.

[22] N. Sha, M. Vanucci, and P. J. Brown. Bayesian variable selection in multinomial probit models with application to spectral data and DNA microarrays. *Submitted for publication*, 2002.

[23] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, 58:267–288, 1996.

[24] M. Tipping. The relevance vector machine. In *Advances in Neural Information Processing Systems*, volume 12, pages 652–658, 2000.

[25] A. C. Trajstman and H. T. Kiiveri. Applications of a rapid variable selection technique to microarray output and survival data generated from a study of B-cell lymphoma: gene discovery and survival prognosis. *CSIRO CMIS internal report*, 2002.

[26] R. W. M. Wedderburn. Quasi-likelihood functions, generalised linear models, and the Gauss-Newton method. *Biometrika*, 64:439–447, 1974.

Minimum Description Length Model Selection Criteria for Generalized Linear Models

Mark H. Hansen and Bin Yu

Abstract

This paper derives several model selection criteria for generalized linear models (GLMs) following the principle of Minimum Description Length (MDL). We focus our attention on the mixture form of MDL. Normal or normal-inverse gamma distributions are used to construct the mixtures, depending on whether or not we choose to account for possible over-dispersion in the data. In the latter case, we apply Efron's [6] double exponential family characterization of GLMs. Standard Laplace approximations are then employed to derive computationally tractable selection rules. Each constructed criterion has adaptive penalties on model complexity, either explicitly or implicitly. Theoretical results for the normal linear model, and a set of simulations for logistic regression, illustrate that mixture MDL can "bridge" the selection "extremes" AIC and BIC in the sense that it can mimic the performance of either criterion, depending on which is best for the situation at hand.

Keywords: AIC; Bayesian methods; BIC; code length; information theory; minimum description length; model selection; generalized linear models

1 Introduction

Statistical model selection attempts to decide between competing model classes for a data set. As a principle, maximum likelihood is not well suited for this problem as it suggests choosing the largest model under consideration. Following this strategy, we tend to overfit the data and choose models that have poor predictive power. Model selection emerged as a field in the 1970s, introducing procedures that "corrected" the maximum likelihood approach. The most famous and widely used criteria are *An Information Criterion* (AIC) of Akaike [1, 2] and the *Bayesian Information Criterion* (BIC) of Schwarz [15]. They both take the form of a penalized maximized likelihood, but with different penalties: AIC adds 1 for each additional variable included in a model, while BIC adds $\log n/2$, where n is the sample size. Theoretical and simulation studies (*cf.* Shibata [16], Speed and Yu [18], and references therein), mostly in the regression case, have revealed that when the underlying model is finite-dimensional (specified by a finite number of parameters), BIC is preferred; but when it is infinite-dimensional, AIC performs best. Unfortunately, in practical applications we rarely have this level of

information about how the data were generated, and it is desirable to have selection criteria which perform well independent of the form of the underlying model. That is, we seek criteria which adapt automatically to the situation at hand. In this paper, we derive such adaptive model selection criteria for generalized linear models (GLMs) under the Minimum Description Length (MDL) framework. With MDL we find several generic prescriptions or "forms" for constructing such selection criteria. In this paper, we focus on one MDL form that is based on mixtures.

The MDL approach began with Kolmogorov's theory of algorithmic complexity, matured in the literature on information theory, and has recently received renewed interest within the statistics community. By viewing statistical modeling as a means of generating *descriptions* of observed data, the MDL framework (*cf.* Rissanen [13], Barron *et al.* [3], and Hansen and Yu [8]) discriminates between competing model classes based on the *complexity* of each description. Precisely, the Minimum Description Length (MDL) Principle recommends that we

Choose the model that gives the shortest description of data.

While there are many kinds of descriptions and many ways to evaluate their complexity, we follow Rissanen [13] and use a *code length* formulation based on the candidate model.

To make this more precise, we first recall that for each probability distribution Q on a finite set $\mathcal{A}$ there is an associated *code* that prepares elements of $\mathcal{A}$ for transmission across some (noiseless) communication channel. We consider binary codes, meaning that each codeword is a string of 0's and 1's. It is possible to find a code so that the number of bits (the number of 0's and 1's in a codeword) used to encode each symbol of $a \in \mathcal{A}$ is essentially $-\log_2 Q(a)$; that is, $-\log_2 Q$ can be thought of as a *code length function.* Huffman's algorithm [5] takes a distribution Q and produces a so-called *prefix code* with the right length function.[1] Conversely, any integer-valued function L corresponds to the code length of some binary prefix code if and only if it satisfies Kraft's inequality

$$\sum_{a \in \mathcal{A}} 2^{-L(a)} \leqslant 1, \tag{1}$$

see Cover and Thomas [5] for a proof. Therefore, given a prefix code on $\mathcal{A}$ with length function L, we can define a distribution on $\mathcal{A}$ as follows:

$$Q(a) = \frac{2^{-L(a)}}{\sum_{z \in \mathcal{A}} 2^{-L(z)}} \quad \text{for any } a \in \mathcal{A}.$$

With Kraft's inequality, we find a correspondence between codes and probability distributions. In what follows, we work with natural logs and take $-\log Q$ to be an idealized code length.

[1] While the details are beyond the scope of this short paper, the interested reader is referred to Hansen and Yu [8] and Cover and Thomas [5].

One of the early problems in information theory involves transmitting symbols that are randomly generated from a probability distribution P defined on $\mathcal{A}$. Let $A \in \mathcal{A}$ denote a random variable with this distribution. Now, from the discussion in the previous paragraph, any code defined on $\mathcal{A}$ can be associated with an idealized code length function $-\log Q$ for some distribution Q. With this setup, the expected code length for symbols generated from P is given by $-E \log Q(A) = -\sum_a P(a) \log Q(a)$. By Jensen's inequality, we see that the shortest code length is achieved by a code that has $-\log P$ as its idealized length function. That is, the expected code length is bounded from below by $-E \log P(a) = -\sum_a P(a) \log P(a)$, the entropy of P. In the literature on information theory, this fact is known as Shannon's Inequality.

In this paper, we focus on descriptions of data that consist of probability models, and compare them based on the efficiency of the corresponding code in terms of improvements in code length relative to the entropy of the data generating process. When the competing models are members of a parametric family, using MDL to select a model, or rather, to estimate a parameter, is equivalent to maximum likelihood estimation (when the cost of transmitting the parameter estimate is fixed). To compare different model classes, different parametric families, or carry out model selection from among several candidate model classes, efficient codes for each class need to *fairly represent* its members. We do not elaborate on this idea, but instead comment that it is possible to demonstrate rigorously that several coding schemes achieve this fairness and hence provide valid selection criteria (for, say, i.i.d. or time series observations). We refer readers to Barron *et al.* [3] and Hansen and Yu [8].

Among the schemes that yield valid selection criteria, the best known is the so-called *two-stage code*, in which we first encode the maximum likelihood estimate (MLE) of the parameters in the model, and then use the model with the MLE to encode the data (say, via Huffman's algorithm described above). Hence this form is a penalized likelihood, and to first order is exactly the same as BIC. Other forms of MDL include predictive, mixture and normalized maximum likelihood (NML). The predictive form makes the most sense when the data come in sequentially and has a close connection to prequential inference; the mixture codes are described in more detail in the next section; the NML form is new and evolving, and code length expressions are known only in a few special cases, including the binomial model and Gaussian linear regression (*cf.* Rissanen [14], Barron *et al.* [3], and Hansen and Yu [8]).

The rest of the paper is organized as follows. Section 2 gives the details of a mixture code in the context of regression-type models. Section 3 covers the gMDL model selection criterion (so named because of its use of Zellner's g-prior [20]) from Hansen and Yu [8] in the variance known and unknown cases to prepare the reader for the new results in Section 4. The criterion when σ^2 is known appeared originally in George and Foster [7] in the context of a Bayesian analysis. Section3.3 contains a new theorem to show the bridging effect of the gMDL criterion between AIC and BIC in a normal linear regression model.

Section 4 derives a version of the mixture form gMDL for GLMs. In this case,

normal or normal-inverse gamma distributions are used to construct a mixture model, depending on whether or not we choose to account for possible over-dispersion in the data. When the dispersion parameter is known, the resulting criterion appeared first in Peterson [11] in the context of a Bayesian analysis. To account for dispersion effects, we use Efron's [6] double exponential family characterization of GLMs as the likelihood. Standard Laplace approximations are employed to derive computationally tractable selection rules. Each constructed criterion has adaptive penalties on model complexity, either explicitly or implicitly. The last section of the paper contains a set of simulations for logistic regression to illustrate that mixture MDL can "bridge" AIC and BIC in the sense that it can mimic the performance of either criterion, depending on which is best for the situation at hand. The performance measures include the probability of selecting the correct model and test-error based on a selected model. The latter is found to be much less sensitive to the model selection criterion than the former due to the robustness of 0-1 loss in classification.

2 Mixture MDL

In this paper, we consider regression-type models; that is, we would like to characterize the dependence of a random variable $Y \in \mathcal{Y} \subset \mathbb{R}$ on a vector of potential covariates $(X_1, \ldots, X_K) \in \mathbb{R}^K$. We consider various parametric model classes (or conditional densities) for Y, indexed by a 0-1 binary vector $\gamma = (\gamma_1, \ldots, \gamma_K)$; each model depends on a subset of the covariates corresponding to 1's in the model index vector γ. Generically, we let $\mathcal{M}_\gamma$ denote a simple model class with dimension $k_\gamma = \sum_{j=1}^K \gamma_j$, which depends on the predictors $(X_1, \ldots, X_K)$ through the linear combination

$$\sum_{j:\gamma_j=1} \beta_j X_j, \tag{2}$$

where $\beta_\gamma = (\beta_j)_{\{j:\gamma_j=1\}}$ is a vector of parameters. To fit this relationship, our basic data are observations of the form (Y_i, X_i), $i = 1, \ldots, n$, where $X_i = (X_{i1}, \ldots, X_{iK})$. In observational studies it makes sense to consider X_i as being random, whereas in designed experiments the values of the covariates are specified. Let $Y = (Y_1, \ldots, Y_n) \in \mathcal{Y}^n$ denote the vector of responses and let X_K be the $n \times K$ full design matrix, $[X_K]_{ij} = X_{ij}$. By X_γ we mean a submatrix of X consisting of those columns j for which $\gamma_j = 1$. We connect the data to the model (2) via the conditional density functions $f_{\theta_\gamma}(y|X_\gamma)$, $y \in \mathcal{Y}^n$, for some set of parameters $\theta_\gamma \in \Theta$. (Typically, θ_γ will include regression parameters β_γ and possibly a dispersion effect.) In order to assess the suitability of $\mathcal{M}_\gamma$, we derive a description length for Y based on $\mathcal{M}_\gamma$.

For simplicity, we now drop the subscript γ except in places where a reminder seems necessary. The reader should interpret the model class $\mathcal{M}$, its dimension k, the design matrix X, and the parameters θ and β as all depending on some subset of the available predictors. We then judge the appropriateness of this model based on the so-called *mixture form* of MDL. As its name suggests, this criterion starts with a mixture distribution

that combines all the members in the class $\mathcal{M}$

$$m(y|X) = \int f_\theta(y|X)w(\theta|X)d\theta, \quad y \in \mathcal{Y}^n, \tag{3}$$

where w is a probability density function on θ. This integral has a closed form expression when $f_\theta(\cdot|X)$ is an exponential family and w is a conjugate distribution.

If $\mathcal{Y}$ is a finite set of values, we can use the distribution (3) to directly form a *mixture code* for strings $y \in \mathcal{Y}^n$. In this setting, we assume that both sender and receiver know about the covariates X, and we only have to transmit y. As an example, suppose $\mathcal{Y} = \{0,1\}$ so that y is a binary string of length n. We use the model class $\mathcal{M}$ and the distribution (3) to construct a mixture code for all 2^n strings $y \in \{0,1\}^n$. From the discussion in Section 1, we can apply Huffman's algorithm to build a code that has the (idealized) length function $L(y) = -\log_2 m(y|X)$ for all $y \in \mathcal{Y}^n$. This means that the number of bits required to transmit any $y \in \{0,1\}^n$ is essentially $-\log_2 m(y|X)$. The MDL principle then distinguishes between candidate model classes based on the associated length function $L(Y)$, the number of bits required to transmit the observed data Y. As mentioned earlier, we have chosen to use base e in the log for our derivations.

In Section 1, we only considered building codes for finite sets of symbols. When $Y_i \in \mathcal{Y} \subset \mathbb{R}$, $i = 1,\ldots,n$, is a continuous response, we form an approximate length function by first discretizing the set $\mathcal{Y}$. That is, given a precision δ we obtain the description length

$$-\log \int f_\theta(y|X)w(\theta|X)d\theta + n\log\delta. \tag{4}$$

Assuming that the precision used for this approximation is the same regardless of model class $\mathcal{M}$, we again arrive at the expression

$$-\log \int f_\theta(y|X)w(\theta|X)d\theta \tag{5}$$

as a suitable length function. In the next section, we present a brief review of mixture MDL for the simple linear model. A full derivation of these results can be found in Hansen and Yu [8].

When choosing between two model classes, the mixture form of MDL (with fixed hyperparameters) is equivalent to a Bayes factor (Kass and Raftery [9]) based on the same distributions on the parameters spaces. As we see in the next section, MDL allows for a natural, principled mechanism for dealing with hyperparameters that distinguishes it from classical Bayesian analysis. Also, keep in mind that w is not introduced as a prior in the Bayesian sense, but rather as a device for creating a distribution for the data Y from $\mathcal{M}$. This distinction also allows more freedom in choosing w, and has spawned a number of novel applications in engineering.

3 Regression

We begin with the simplest GLM, namely the normal linear model $\mathcal{M}_\gamma$:

$$Y_i = \sum_{j:\gamma_j=1} \beta_j X_{ij} + \varepsilon_i \,. \tag{6}$$

where the ε_i are normally distributed with mean zero and variance σ^2. To remind the reader that our basic model classes will consist of various subsets of the predictor variables $(X_1, \ldots, X_K)$, we restored the γ notation in the above equation. For simplicity, however, from this point on, we drop it and consider derivations with respect to a single model class, a single choice of γ. Technically, we do not need to assume that the relationship in (6) holds for some collection of predictors X_K, but instead we entertain model classes because they are capable of capturing the major features observed in the observed data string Y. For comparison with more general GLMs later, we treat separately the case in which σ^2 is known and unknown. In the former case, the parameter vector θ in the mixture (3) consists only of the coefficients β; while in the latter, θ involves both β and σ^2.

We review this material because relatively straightforward, direct analysis yields the MDL selection criteria. When we tackle the complete class of GLMs, the derivation becomes more difficult, but the final forms are reminiscent of those derived in this section.

3.1 Known error variance σ^2

Here, we take $\theta = \beta$ and let $w(\beta|X)$ be a normal distribution with mean zero and variance-covariance matrix $\sigma^2 V$. As $\mathcal{Y} = \mathbb{R}$, we have to appeal to the discretized form of MDL (4). By using a conjugate distribution, we are able to perform the integration in (5) exactly. This leads to a code length of the form

$$\begin{aligned} L(y|V) &= -\log m(y|X,V) \\ &= \frac{1}{2}\log|V^{-1}+X^tX| + \frac{1}{2}\log|V| \\ &\quad + \frac{1}{2\sigma^2}\left(y^ty - y^tX\left(V^{-1}+X^tX\right)^{-1}X^ty\right), \end{aligned} \tag{7}$$

where we have dropped terms that depend only on n. We have also made explicit the dependence of the length function on the variance-covariance matrix V. Clearly, we can simplify this expression by taking $V = c(X^tX)^{-1}$ so that

$$-\log m(y|X,c) = \frac{k}{2}\log(1+c) + \frac{1}{2\sigma^2}\left(y^ty - \frac{c}{1+c}FSS\right), \tag{8}$$

where $FSS = y^tX(X^tX)^{-1}X^ty$ is the usual fitted sum of squares corresponding to the OLS estimate $\widehat{\beta} = (X^tX)^{-1}X^tY$. This particular choice of distribution is often attributed

to Zellner [20] who christened it the g-prior. Because the mixture form reduces to a relatively simple expression, the g-prior has been used extensively to derive Bayesian model selection procedures for the normal linear model. Under this prior, it is not hard to show that the posterior mean of β is $\frac{c}{1+c}\widehat{\beta}$.

In (8) we have highlighted the dependence of the mixture on the scaling parameter c. George and Foster [7] studied various approaches to setting c, establishing that certain values lead to well-known selection criteria like AIC and BIC.[2] Ultimately, they propose an empirical Bayes approach, selecting an estimate $\hat{c}$ via maximum likelihood. Hansen and Yu [8] take a similar approach to the hyperparameter c, but motivate it from a coding perspective. We review this approach here. Essentially, each choice of c produces a different mixture distribution and hence a different code. Therefore, to let c depend on the data, both sender and receiver need to agree on which value of c to use. Hansen and Yu [8] take a two-stage approach to hyperparameters like c; that is, c is transmitted first and then once each side knows which code to use, the data are sent. Of course, communicating c in this way adds to the code length, a charge that we make explicit by writing

$$L(y) = L(y|c) + L(c) = -\log m(y|X,c) + L(c). \tag{9}$$

Following Rissanen [13], the cost $L(c)$ is taken to be $\frac{1}{2}\log n$.[3] Minimizing (9) with respect to c gives

$$\hat{c} = \max\left(\frac{FSS}{k\sigma^2} - 1, 0\right),$$

and substituting into (8) yields a code length (9) of the form

$$L(y) = \begin{cases} \frac{y'y - FSS}{2\sigma^2} + \frac{k}{2}\left[1 + \log\left(\frac{FSS}{\sigma^2 k}\right)\right] + \frac{1}{2}\log n & \text{for } FSS > k\sigma^2 \\ \\ \frac{y'y}{2\sigma^2} & \text{otherwise.} \end{cases} \tag{10}$$

When the minimizing value of c is zero, the prior on β becomes a point mass at zero, effectively producing the "null" model corresponding to all effects being zero. This accounts for the second case in the above expression. We should note that the extra $\frac{1}{2}\log n$ penalty is essential to guarantee consistency of the selection method when the null model is true. The Bayesian criterion of George and Foster [7] is basically the same, but leaves off this extra term.

[2] A similar calibration between Bayesian methods and well-known selection criteria can also be found in Smith and Spielgelhalter [17].

[3] The cost $\frac{1}{2}\log n$ can be motivated as follows: for regular parametric families, an unknown parameter can be estimated at rate $1/\sqrt{n}$. Hence there is no need to code such a parameter with a precision finer than $1/\sqrt{n}$. Coding c with precision $1/\sqrt{n}$ gives a cost to the first order $-\log[1/\sqrt{n}] = \log n/2$.

3.2 Unknown error variance

We now consider the regression model (6) when σ^2 is unknown. George and Foster [7] advocate estimating σ^2 then applying the form (10); however, we prefer to assign a distribution to σ^2 and incorporate it into the mixture. Following Hansen and Yu [8], we employ a conjugate normal-inverse gamma distribution to form the mixture code; that is, $1/\sigma^2$ has a gamma distribution with shape parameter a; and given σ^2, β is normal with mean zero and variance $\sigma^2 V$. Setting $\tau = \sigma^2$, these densities are given by

$$w(\beta,\tau) \propto \tau^{\frac{-(k+3)}{2}} \exp\left[\frac{-\beta' V^{-1}\beta + a}{2\tau}\right], \tag{11}$$

where a and V are hyperparameters. Under this class of priors, the mixture distribution (3) has the form

$$\begin{aligned} -\log m(y|X,a,V) &= \frac{1}{2}\log|V^{-1}+X'X| + \frac{1}{2}\log|V| - \frac{1}{2}\log a \\ &\quad + \frac{n+1}{2}\log\left(a + y'y - y'X\left(V^{-1}+X'X\right)^{-1}X'y\right) \end{aligned} \tag{12}$$

where we have ignored terms that do not depend on our particular choice of model. The derivation of $m(y|X,a,V)$, the marginal or predictive distribution of y, is standard and can be found in O'Hagan [10].

Our approach to handling the hyperparameter a is the same as that in the previous section. Minimizing (12) with respect to a we find that $\hat{a} = (y'y - y'X(V^{-1} + X'X)^{-1}X'y)/n$ which leaves

$$\begin{aligned} -\log m(y|X,\hat{a},V) &= \frac{1}{2}\log|V^{-1}+X'X| + \frac{1}{2}\log|V| \\ &\quad + \frac{n}{2}\log\left(y'y - y'X\left(V^{-1}+X'X\right)^{-1}X'y\right). \end{aligned} \tag{13}$$

As in the known-variance case, we can achieve a simplification in computing the mixture distribution if we again make Zellner's choice of $V = c(X'X)^{-1}$. This leaves

$$-\log m(y|X,\hat{a},c) = \frac{k}{2}\log(1+c) + \frac{n}{2}\log\left(y'y - \frac{c}{1+c}FSS\right), \tag{14}$$

To settle the hyperparameter c, we again minimize the overall code length to find

$$\hat{c} = \max(F-1,0) \quad \text{with} \quad F = \frac{FSS}{kS}, \tag{15}$$

where F is the usual F-ratio for testing the hypothesis that each element of β is zero, and $S = RSS/(n-k)$. The truncation at zero in (15) rules out negative values of the prior variance. Rewriting (15), we find that $\hat{c}$ is zero unless $R^2 > k/n$, where R^2 is the usual squared multiple correlation coefficient. When the value of $\hat{c}$ is zero, the prior

on β becomes a point mass at zero, effectively producing the "null" mixture model[4] corresponding to zero regression effects. Substituting the optimal value of $\hat{c}$ into (14), we arrive at a final mixture form

$$\text{gMDL} = \begin{cases} \frac{n}{2}\log S + \frac{k}{2}\log F + \log n, & R^2 \geqslant k/n \\ \frac{n}{2}\log\left(\frac{y'y}{n}\right) + \frac{1}{2}\log n, & \text{otherwise.} \end{cases} \tag{16}$$

Note that we have added the cost to code the hyperparameters a and c, producing an extra $\log n$ and $(1/2)\log n$ in the upper and lower expressions, respectively.

3.3 Comparison

As alluded to in the introduction, two widely used model selection criteria are AIC and BIC. In the case of regression with an unknown variance, they take forms

$$\text{AIC} = \frac{n}{2}\log RSS + k \quad \text{and} \quad \text{BIC} = \frac{n}{2}\log RSS + \frac{k}{2}\log n. \tag{17}$$

Comparing these with (16), we see that the essential difference is in the penalty. Both AIC and BIC have data independent penalties, while gMDL has a data-dependent $\log F/2$ for each additional dimension.

By charging less for each new variable, AIC tends to include more terms. When the underlying model consists of many effects, or more precisely the model is infinite-dimensional, AIC tends to perform better. If we take the figure of merit to be prediction error, then AIC has been shown both through theory and simulation studies to be optimal in this setting. When the true, data generating mechanism is finite-dimensional (and is included among the candidates being compared), the stronger penalty of BIC tends to perform better. For this kind of problem, selection criteria may also be judged based on consistency (which leads to prediction optimality); that is, whether or not they ultimately select the correct model as the number of samples tends to infinity. BIC has been shown to perform optimally in this setting.

We now demonstrate that gMDL with its adaptive penalty enjoys the advantages of both AIC and BIC in the regression context. We focus on the simple linear model because the expressions are easy to work with, although we expect the same kind of result will hold for GLMs. To simplify our analysis, we assume the regressors are ordered as $X_{i1}, X_{i2}, \ldots$. Following Breiman and Freedman [4], we assume that $X_i = (X_{i1}, X_{i2}, \ldots, X_{ij}, \ldots)$ are Gaussian, zero-mean random vectors and let

$$\sigma_k^2 = \text{var}\left(\sum_{j=k+1}^{\infty} \beta_j X_{ij} \,\middle|\, X_{i1}, \ldots, X_{ik} \right).$$

[4]The null model is a scale mixture of normals, each $N(0,\tau)$ and τ having an inverse-gamma prior.

Then the finite-dimensional model assumption implies that $\sigma^2_{k_0} = 0$ for some $k_0 > 0$. Using similar arguments as those used to prove Theorem 1.4 in Breiman and Freedman [4] and the fact that $\frac{1}{n}||y||^2 = (\sigma^2 + \sigma_0^2)(1 + o_p(1))$, it is straightforward to establish the following two results.

Theorem 9
The quantity F in (15) satisfies $F = [\frac{n}{k}\frac{\sigma_0^2 - \sigma_k^2}{\sigma_k^2 + \sigma^2} + 1](1 + o_p(1))$ where $o_p(1) \to 0$ in probability uniformly over $0 \leqslant k \leqslant n/2$.

Corollary 3
If the model is finite-dimensional and the maximum dimension of the models examined $K = K_n = o(n)$, then gMDL is consistent and is also prediction-optimal.

The above theorem presents an expansion of the data dependent-penalty of gMDL, and the corollary establishes that gMDL enjoys the same optimality as BIC when the model is finite-dimensional. When $\sigma_k^2 > 0$ for all k, the underlying model is infinite-dimensional. In this case, the quantity F/n can be viewed as the average signal to noise ratio for the fitted model. Adjusting the penalty with $(k/2)\log F/n$, gMDL is able to adapt to perform well in terms of prediction in both domains, finite- or infinite-dimensional. The simulation studies in Hansen and Yu (2001) support this adaptivity of gMDL, since there gMDL has an overall prediction performance better than AIC or BIC.

In the next section, we show that the newly derived MDL-based criteria for GLMs are also adaptive.

4 Generalized Linear Models

The characterization of a GLM starts with an exponential family of the form

$$f(y) = \exp\left(\frac{y\psi - b_1(\psi)}{b_2(\phi)} + b_3(y, \phi)\right), \quad y \in \mathcal{Y}, \tag{18}$$

where b_1, b_2 and b_3 are known functions. We refer to ψ as the canonical parameter for the family. Typically, we take $b_2(\phi) = \phi$, and refer to ϕ as the dispersion parameter. It plays the role of the noise-variance in the ordinary regression setup of the previous section. The family (18) contains many practically important cases, including the normal, binomial, Poisson, exponential, gamma and inverse Gaussian distributions. With this model, it is not hard to show that if Y has distribution (18),

$$\begin{array}{lcll} E(Y) & = & \mu & = & b_1'(\psi) \\ \mathrm{var}(Y) & = & \sigma^2 & = & b_1''(\psi) b_3(\phi). \end{array} \tag{19}$$

As with the normal case above, the GLM framework allows us to study the dependence of a response variable $Y \in \mathcal{Y}$ on a vector of covariates $(X_1, \ldots, X_K)$. Each model class

corresponds to some value of the binary vector $\gamma = (\gamma_1, \ldots, \gamma_K)$, and we relate the mean μ of Y to a subset of the covariates via the linear predictor

$$\eta = g(\mu), \qquad \text{for } \eta = \sum_{j:\gamma_j=1} \beta_j X_j, \tag{20}$$

where g is a one-to-one, continuously differentiable transformation known as the link function. Using (19) and (20) we see that $\eta = g(b'(\psi))$.[5] Again we let $\beta_\gamma = (\beta_j)_{j:\gamma_j=1}$ denote the vector of regression coefficients and $k_\gamma = \sum \gamma_j$ its dimension. The unknown parameters associated with this model are denoted θ_γ and include both β_γ as well as a possible dispersion effect ϕ. We observe data of the form (Y_i, X_i) for $i = 1, \ldots, n$ where $X_i = (X_{i1}, \ldots, X_{iK})$ and again X_K is the $n \times K$ full design matrix $[X_K]_{ij} = X_{ij}$. We let X_γ refer to a submatrix of X_K consisting of only those columns j for which $\gamma_j = 1$. Let $f_{\theta_\gamma}(y|X_\gamma)$ denote the density for Y based on model class γ.

As with our treatment of the regression context, maintaining the model index γ needlessly complicates our derivations. From this point on, we again drop it, reminding the reader that terms like $\mathcal{M}$, X, k, and β all refer to a specific subset of covariates. For all the GLM cases, we begin with a Laplace approximation to the mixture form which will be exact for the normal linear model. That is, we start with

$$m(y|X) \approx (2\pi)^{\frac{k}{2}} |-H^{-1}(\tilde{\beta})|^{\frac{1}{2}} f_{\tilde{\beta}}(y|X) w(\tilde{\beta}), \tag{21}$$

where H is the Hessian of $h(\beta) = \log f_\beta(y|X) + \log w(\beta)$ and $\tilde{\beta}$ is the posterior mode of β. In working with this form, we repeatedly make use of the Fisher information matrix $I(\beta) = X^t W(\beta) X$, where W is a diagonal weight matrix. Note that for GLMs, the observed Fisher information is the same as the Fisher information when we use the canonical parameterization.

Form (21) is still difficult to work with in practice because there is typically no closed-form expression for the posterior mode. We now consider several criteria that make sensible choices for f and w that lead to computationally tractable criteria.

4.1 Direct approach

In this section, we derive a criterion that first appeared in Peterson [11]. As with the regression context, the original motivation for this form was not MDL, but rather an approximation to a full Bayesian approach. Our analysis follows closely the case of σ^2 known for regression. Let β be the MLE of β, and assume that the prior $w(\beta)$ is normal with mean zero and variance-covariance V. Then, we can approximate $\tilde{\beta}$ via a single Newton step

$$\begin{aligned} \tilde{\beta} &\approx \beta - H(\beta)^{-1} h'(\beta) \\ &\approx \beta + (I(\beta) + V^{-1})^{-1} V \beta \end{aligned}$$

[5] Taking $b' = g^{-1}$ means that the canonical parameter ψ and the linear predictor η are the same. This choice of g is known as the canonical link function.

using the fact that $H(\beta) = -(I(\beta) + V^{-1})$, where $I(\widehat{\beta})$ is the Fisher information evaluated at $\widehat{\beta}$. We now focus on the case where the prior variance-covariance matrix for β is simply $cI(\widehat{\beta})^{-1}$. For the normal linear model, this leads us to Zellner's g-prior. Unfortunately, for the other important members of this family, the prior variance-covariance matrix will depend on $\widehat{\beta}$. From a strict coding perspective this is hard to accept; it would imply that sender and receiver both know the coefficient $\widehat{\beta}$ (or at least $I(\widehat{\beta})$). Nonetheless, it is instructive to follow this line of analysis and compare it with the results of the previous section. For $V = cI(\widehat{\beta})^{-1}$ we find that the one-step Newton-Raphson iteration gives

$$\tilde{\beta} \approx \frac{c}{1+c}\widehat{\beta}$$

which agrees with our regression form of MDL when σ^2 is known.

Continuing with the expression (21), we find that

$$\begin{aligned} \log w(\tilde{\beta}) &\approx \log w\left(\frac{c}{1+c}\widehat{\beta}\right) \\ &= -\frac{k}{2}\log 2\pi + \log|cI(\widehat{\beta})| - \frac{1}{2}\frac{c}{(1+c)^2}\widehat{\beta}^t I(\widehat{\beta})\widehat{\beta} \end{aligned} \tag{22}$$

and that after a Taylor expansion of $\log f_\beta(y|X)$ around $\widehat{\beta}$

$$\begin{aligned} \log f_{\tilde{\beta}}(y|X) &\approx \log f_{\widehat{\beta}}(y|X) - \frac{1}{2}(\tilde{\beta}-\widehat{\beta})^t I(\widehat{\beta})(\tilde{\beta}-\widehat{\beta}) \\ &= \log f_{\widehat{\beta}}(y|X) - \frac{1}{2}\frac{1}{(1+c)^2}\widehat{\beta}^t I(\widehat{\beta})\widehat{\beta}. \end{aligned} \tag{23}$$

Combining (22) and (23) we arrive at the expression

$$\begin{aligned} \log f_{\tilde{\beta}}(y|X) + \log w(\tilde{\beta}) &\approx \log f_{\widehat{\beta}}(y|X) - \frac{1}{2}\frac{1}{1+c}\widehat{\beta}^t I(\widehat{\beta})\widehat{\beta} \\ &\quad -\frac{k}{2}\log 2\pi + \frac{k}{2}\log c + \frac{1}{2}\log|I(\widehat{\beta})|. \end{aligned} \tag{24}$$

Finally, collecting terms in (21) we find an expression for the code length given c

$$-\log m(y|c,X) \approx \frac{k}{2}\log(1+c) + \frac{1}{2}\frac{1}{1+c}\widehat{\beta}^t I(\widehat{\beta})\widehat{\beta} - \log f_{\widehat{\beta}}(y|X).$$

We then eliminate the hyperparameter c using the same minimization approach in (9). This yields

$$\hat{c} = \max\left(\frac{\widehat{\beta}^t I(\widehat{\beta})\widehat{\beta}}{k} - 1, 0\right).$$

Substituting this in the mixture form, we find the final MDL criterion

$$L(y) = \begin{cases} -\log f_{\widehat{\beta}}(y|X,\widehat{\beta}) + \frac{k}{2}\left[1 + \log\left(\frac{\widehat{\beta}' I(\widehat{\beta})\widehat{\beta}}{k}\right)\right] + \frac{1}{2}\log n & \text{for } \widehat{\beta}' I(\widehat{\beta})\widehat{\beta} > k \\ -\log f_0(y|X) & \text{otherwise.} \end{cases}$$

The function $f_0(y|X)$ represents the log-likelihood when all regression effects are zero. Again, we have added an extra $\frac{1}{2}\log n$ term to the top expression to account for the coding cost of c. This corresponds exactly to the regression context when σ^2 is known.

4.2 Accounting for over-dispersion

In many families, like the Poisson and binomial models, the dispersion parameter is fixed $\phi = 1$. However, in practice it is often the case that the data do not support this value, forcing consideration of over-dispersed models. There are several ways to introduce extra variability into the form (18), many of which are primarily meant as computational devices. Efron [6] constructs a family to explicitly account for over-dispersion that admits an analysis for GLMs similar to that for ordinary regression in the σ^2-unknown case. A related technique was independently derived by West [19].

To understand this form, we have to first rewrite the log-likelihood for a GLM in terms of its mean vector $l(y|\mu)$, where $\mu = (\mu_1, \ldots, \mu_n)$. Now, using this notation, without the restriction (20) on the mean, the maximum value of the log-likelihood is simply $l(y|y)$. We then define the deviance as the difference

$$D(y|\beta) = 2l(y|y) - 2l(y|\mu),$$

where β is the vector of regression coefficients that yield μ through (20). To incorporate a dispersion parameter, Efron [6] motivates the use of

$$\tau^{-n/2} e^{l(y|\mu)/\tau + (1-1/\tau)l(y|y)} \tag{25}$$

as an (approximate) likelihood. Technically, this expression should include a normalizing constant $C(\tau, \beta)$. Following Efron [6], however, it can be shown that $C(\tau, \beta) = 1 + O(n^{-1})$, and hence can be ignored for reasonable sample sizes. Rewriting (25), we work with

$$\tau^{-n/2} e^{-\frac{(2l(y|y) - 2l(y|\mu))}{2\tau}} e^{l(y|y)} = \tau^{-n/2} e^{-\frac{D(y|\beta)}{2\tau}} e^{l(y|y)}. \tag{26}$$

Then, arguing as we did for the σ^2 unknown case in regression, we use a normal-inverse gamma prior with variance-covariance matrix τV. The joint probability of β, τ and y is given by

$$\tau^{-1-(n+1)/2} \frac{e^{-a/2\tau}}{\sqrt{\pi}} \sqrt{\frac{a}{2}} (2\pi\tau)^{-k/2} |V^{-1}|^{1/2} e^{\frac{-\beta' V^{-1}\beta - D(y|\beta)}{2\tau}}. \tag{27}$$

To integrate out β, we use the Laplace method again which this time yields

$$\frac{\tau^{-1-(n+1)/2}}{\sqrt{\pi}}\sqrt{\frac{a}{2}}|V^{-1}|^{1/2}|V^{-1}+I(\widehat{\beta})|^{-1/2}e^{\frac{-a-\tilde{\beta}'V^{-1}\tilde{\beta}-D(y|\tilde{\beta})}{2\tau}}, \tag{28}$$

where $I(\tilde{\beta})$ is the Fisher information matrix evaluated at the posterior mode $\tilde{\beta}$. Integrating with respect to τ then yields

$$\begin{aligned} -\log m(y|a,V) &= \frac{n+1}{2}\log\left(a+\tilde{\beta}'V^{-1}\tilde{\beta}+D(y|\tilde{\beta})\right) \\ &\quad -\frac{1}{2}\log a+\frac{1}{2}\log|V|+\frac{1}{2}\log|V^{-1}+I(\tilde{\beta})|. \end{aligned} \tag{29}$$

Following the prescription in the regression context, we eliminate the hyperparameter a by minimizing the overall code length. In this case, we easily find that

$$\begin{aligned} -\log m(y|\hat{a},V) &= \frac{n}{2}\log\left(\tilde{\beta}'V^{-1}\tilde{\beta}+D(y|\tilde{\beta})\right) \\ &\quad +\frac{1}{2}\log|V|+\frac{1}{2}\log|V^{-1}+I(\tilde{\beta})|. \end{aligned}$$

We have now obtained a usable criterion for model selection. Specifying V, we can compute $\tilde{\beta}$ with simple Newton-Raphson iterations. In the regression analysis, we used Zellner's g-prior for β which led to a closed-form selection criterion. The analog in this case is $V = cI^{-1}(\widehat{\beta})$. For a GLM, this choice is somewhat unsettling because $I(\widehat{\beta})$ is computed at the MLE. If we were to adhere to a strict MDL setting, it would not make sense; from a coding perspective, both sender and receiver would have to know about $\widehat{\beta}$, or at least $I(\widehat{\beta})$. Recall that for a GLM, the Fisher information matrix takes the form $X^tW(\widehat{\beta})X$ where W is a diagonal weight matrix. One simple alternative is to take $V = c(X^tX)^{-1}$, or $V = c1$, where 1 is the identity matrix. In each of these cases, we must either approximate the $\tilde{\beta}$ or iterate to find it. We consider both kinds of selection criteria.

Following the approximation route, if we choose $V = cI^{-1}(\widehat{\beta})$, we get

$$\tilde{\beta} \approx \frac{c}{1+c}\widehat{\beta} \tag{30}$$

and

$$\frac{k}{2}\log(1+c)+\frac{n}{2}\log\left(\frac{1}{1+c}\widehat{\beta}^tI(\widehat{\beta})\widehat{\beta}+D(y|\widehat{\beta})\right). \tag{31}$$

Here we have substituted in the one-step Newton-Raphson approximation for $\tilde{\beta}$ and have approximated the deviance $D(y|\tilde{\beta})$ by a Taylor expansion around $\widehat{\beta}$ and used a relation from Raftery [12]. Maximizing with respect to c yields

$$\hat{c} = \max(F-1,0) \tag{32}$$

where

$$F = \frac{(n-k)\widehat{\beta}' I(\widehat{\beta})\widehat{\beta}}{kD(y|\widehat{\beta})}.$$

This then gives the form

$$L(y) = \begin{cases} \frac{n}{2}\log\frac{D(y|\widehat{\beta})}{n-k} + \frac{k}{2}\log F + \log n & \text{if } F > 1 \\ \frac{n}{2}\log\frac{D(y|0)}{n} + \frac{1}{2}\log n & \text{otherwise,} \end{cases}$$

where $D(y|0)$ represents the deviance calculated under a model with zero regression effects.

For the other choices of $V^{-1} = c\Sigma$, there is not a closed-form expression for the maximizing c. Instead, we can perform a search, but this is best done in conjunction with finding $\tilde{\beta}$. It is also possible to use the approximate $\tilde{\beta}$ (30) to derive a simple iteration to find c. In this case, we find

$$c = \frac{kR_c}{n\widehat{\beta}' I(\widehat{\beta})\left(I(\widehat{\beta}) + c\Sigma\right)^{-1}\Sigma\left(I(\widehat{\beta}) + c\Sigma\right)^{-1} I(\widehat{\beta})\widehat{\beta} + R_c \text{trace}\left(c1 + I(\widehat{\beta})\Sigma^{-1}\right)} \tag{33}$$

where

$$R_c = D(y|\widehat{\beta}) + c\widehat{\beta}'\Sigma\widehat{\beta} - c^2\widehat{\beta}'\Sigma(I(\widehat{\beta}) + c\Sigma)^{-1}\Sigma\widehat{\beta}. \tag{34}$$

Convergence of this algorithm is usually fairly fast, although as we will see, it can depend on the starting values.

5 Simulations

We have chosen 8 different simulation setups to compare AIC and BIC with the new MDL-based criteria derived in this section. We focus on logistic regression, and consider $K = 5$ potential covariates. We specify two distributions on X. In the first, each column consists of $n = 100$ observations from a standard normal distribution and the different columns are independent. In the second case, we again use normal covariates, but now we consider a correlation structure of the form

$$\text{cov}(X_i, X_j) = \rho^{|i-j|}, \text{ for } i, j = 1, \ldots, 5.$$

Here, we took $\rho = 0.75$. Then Y was generated by the standard logistic GLM using one of 8 different coefficient vectors. All $2^5 = 32$ possible models were fit and compared using the various selection criteria. Table 1 gives the classification error rate for each procedure: Column 4 corresponds to mixture MDL with a normal-inverse gamma mixing distribution to capture dispersion effects and $V^{-1} = c^{-1}I(\widehat{\beta})$ (Section 4.2); Column 5 corresponds to mixture MDL with a fixed dispersion parameter ϕ and hence a normal mixing distribution again with $V^{-1} = c^{-1}I(\widehat{\beta})$ (Section 4.1); Columns 6 and

Table 1: Classification errors for the different selection criteria.

Coefficients	ρ	Bayes Rate	Mix ϕ	$\phi = 1$	BIC	AIC	1 Iter	X^tX Iter	1 search	X^tX search
3 2 2 0 0	0	0.125	0.138	0.138	0.135	0.137	0.137	0.138	0.137	0.137
	0.75	0.087	0.101	0.101	0.104	0.101	0.100	0.104	0.100	0.101
5 1 1 1 1	0	0.098	0.115	0.116	0.128	0.118	0.120	0.118	0.118	0.118
	0.75	0.072	0.087	0.087	0.092	0.089	0.087	0.087	0.087	0.089
2 2 2 2 2	0	0.115	0.130	0.130	0.131	0.130	0.130	0.131	0.130	0.130
	0.75	0.067	0.081	0.083	0.095	0.086	0.081	0.081	0.081	0.087
3.0 1.5 1.5 0.5 0.0	0	0.137	0.153	0.152	0.154	0.152	0.153	0.155	0.153	0.153
	0.75	0.093	0.108	0.108	0.112	0.109	0.108	0.111	0.108	0.109
5 0 0 0 0	0	0.103	0.116	0.114	0.110	0.114	0.113	0.113	0.113	0.113
	0.75	0.104	0.116	0.115	0.109	0.114	0.114	0.112	0.114	0.113
2 0 0 0 0	0	0.220	0.233	0.233	0.228	0.233	0.230	0.230	0.230	0.230
	0.75	0.221	0.231	0.231	0.227	0.232	0.230	0.229	0.230	0.229
1.5 1.5 1.5 1.5 0.0	0	0.163	0.177	0.177	0.177	0.177	0.177	0.180	0.177	0.177
	0.75	0.101	0.119	0.120	0.129	0.121	0.118	0.124	0.118	0.122
8 4 2 1 0	0	0.060	0.074	0.074	0.077	0.074	0.075	0.074	0.074	0.074
	0.75	0.040	0.057	0.058	0.060	0.058	0.057	0.057	0.057	0.058

7 are BIC and AIC (17). Columns 8 through 11 also make use of the normal-inverse gamma distribution but with different choices of the variance-covariance matrix V^{-1}; $c^{-1}1$ for 8 and 10, and $c^{-1}X^tX$ for 9 and 11. Columns 8 and 10 differ only in how we estimate $\tilde{\beta}$ and $\hat{c}$; in the first case the iteration (33) is used, while in the second a full search is performed to identify both $\tilde{\beta}$ and the appropriate value of c. The same holds for Columns 9 and 11, but with the different variance-covariance matrix.

Table 1 shows that most of the selection criteria behave the same, at least in terms of classification error; this 0-1 error is very robust. In Table 2 we illustrate the types of models selected by each scheme. The first column identifies the simulations from Table 1. The second column presents a model summary of the form $x-y$ where x denotes the number of variables correctly included in the model and y denotes the number of excess variables. So, for the first panel of Table 2, the true model $(2,0,0,0,0)$ consists of only one effect. The heading "1-0" represents the correct model and is marked with a "*", while the column "1-1" means that one extra term was included. From this table, we see that the three MDL criteria (Columns 9, 11 and 12) adapt to either AIC or BIC depending on which performs better in all 8 set-ups. Column 10 seemed to have some problems, and we believe this is because the iterations (33) failed to converge properly (possibly due to the approximations used to generate the form). Finally, we see that the columns using $I(\widehat{\beta})$ can perform poorly (those denoted Mixture ϕ and $\phi = 1$). Recall that we derived these forms even though their reliance on $\widehat{\beta}$ violates the basic coding ideas behind MDL.

We consider the cases in more depth, starting with the first panel of Table 2. Here "truth" is a small model, $(2,0,0,0,0)$, an ideal case for BIC. Clearly, BIC selects the right model more often than the other procedures. The mixture MDL procedures that use variance-covariance matrices other than $I(\widehat{\beta})$ also perform quite well. In terms of

Table 2: Summarizing the number of times different sized models were selected for a sample of simulation runs given in Table 1.

Coefficients		Model Summary	Mix ϕ	$\phi = 1$	BIC	AIC	1 Iter	X^tX Iter	1 search	X^tX search
$\beta = (2,0,0,0,0)$		0-1	0	0	0	0	0	0	0	0
		0-2	0	0	0	0	0	0	0	0
		0-3	0	0	0	0	0	0	0	0
		0-4	0	0	0	0	0	0	0	0
	*	1-0	131	134	215	121	176	179	183	179
		1-1	70	72	31	85	55	53	51	51
		1-2	37	37	4	40	17	18	15	19
		1-3	12	7	0	4	2	0	1	1
		1-4	0	0	0	0	0	0	0	0
$\beta = (3,2,2,0,0)$		0-1	0	0	0	0	0	0	0	0
		0-2	0	0	0	0	0	0	0	0
		1-0	0	0	0	0	0	0	0	0
		1-1	0	0	0	0	0	0	0	0
		1-2	0	0	0	0	0	0	0	0
		2-0	0	0	0	0	0	3	0	0
		2-1	0	0	1	0	0	0	0	0
		2-2	0	0	0	0	0	0	0	0
	*	3-0	111	134	227	173	176	71	184	180
		3-1	103	93	20	71	49	31	61	64
		3-2	36	23	2	6	25	145	5	6
$\beta = (5,1,1,1,1)$		1-0	0	0	4	0	0	1	0	0
		2-0	0	1	35	2	4	19	3	3
		3-0	9	12	64	24	33	13	23	23
		4-0	56	70	89	94	79	18	86	89
	*	5-0	185	167	58	130	134	199	138	135

test error, each procedure is about the same. Overall, we can recommend the MDL-based criteria in terms of their ability to adapt and select concise models.

In the second panel of Table 2, the coefficient vector is $(3,2,2,0,0)$, a middle-ground case. The $I(\widehat{\beta})$ criteria perform rather poorly, as does the X^tX case with iterations (33) to find $\hat{c}$. In the latter case, the poor performance is even reflected in the prediction error. We intend to examine whether the approximation that led to (33) caused the problem, or if it was poor starting values for the iterations.

Finally, in the last panel of Table 2, we consider a "full" model with coefficient vector $(5,1,1,1,1)$, an ideal situation for AIC. Here we see that BIC fails to capture the correct model form, and the test error is slightly worse as a result. All the MDL criteria outperform even AIC in terms of identifying the correct model, although this does not translate into significant test error improvements.

Acknowledgments

B. Yu's research is partially supported by grants from NSF (FD01–12731) and ARO (DAAD19–01–1–0643). M. Hansen would like to thank John Chambers, Diane Lambert, Daryl Pregibon and Duncan Temple Lang for helpful discussions.

Dedication

This paper is dedicated to Terry Speed, who introduced B. Yu to MDL 15 years ago. By encouraging the highest academic standards, Terry is an inspiration to all of Berkeley's students.

Mark H. Hansen, Statistics Research, Bell Laboratories, Lucent Technologies,
cocteau@bell-labs.com

Bin Yu, Department of Statistics, University of California, Berkeley,
binyu@stat.berkeley.edu

References

[1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.

[2] H. Akaike. An objective use of Bayesian models. *Annals of the Institute of Statistical Mathematics*, 29:9–20, 1977.

[3] A. Barron, J. Rissanen, and B. Yu. Minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory. (Special Commemorative Issue: Information Theory: 1948-1998)*, 44:2743–2760, 1998.

[4] L. Breiman and D. Freedman. How many variables should be entered in a regression equation? *Journal of the American Statistical Association*, 78:131–136, 1983.

[5] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, New York, 1991.

[6] B. Efron. Double exponential families and their use in generalized linear regression. *Journal of the American Statistical Association*, 81:709–721, 1986.

[7] E. I. George and D. P. Foster. Calibration and empirical Bayes variable selection. *Biometrika*, 87:731–747, 2000.

[8] M. Hansen and B. Yu. Model selection and minimum description length principle. *Journal of the American Statistical Association*, 96:746–774, 2001.

[9] R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90:773–795, 1995.

[10] A. O'Hagan. *Kendall's Advanced Theory of Statistics: Bayesian Inference. Vol 2B*. John Wiley & Sons, New York, 1994.

[11] J. J. Peterson. A note on some model selection criteria. *Statistics and Probability Letters*, 4:227–230, 1986.

[12] A. E. Raftery. Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika*, 83:251–266, 1996.

[13] J. Rissanen. *Stochastic complexity and statistical inquiry*. World Scientific, Singapore, 1989.

[14] J. Rissanen. Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, 42:48–54, 1996.

[15] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:4611–464, 1978.

[16] R. Shibata. An optimal selection of regression variables. *Biometrika*, 68:45–54, 1981.

[17] A. F. M. Smith and D. J. Spielgelhalter. Bayes factors and choice criteria for linear models. *Journal of the Royal Statistical Society, Series B*, 42:213–220, 1980.

[18] T. Speed and B. Yu. Model selection and prediction: normal regression. *Journal of the Institute of Statistical Mathematics*, 45:35–54, 1993.

[19] M. West. Generalized linear models: scale parameters, outlier accomodation and prior distributions. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, editors, *Bayesian Statistics 2*, pages 531–558, Amsterdam, 1985. North-Holland.

[20] A. Zellner. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In P.K. Goel and A. Zellner, editors, *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, pages 233–243, Amsterdam, 1986. North-Holland.

Risk Assessment: a Forest Fire Example

David R. Brillinger, Haiganoush K. Preisler and John W. Benoit

Abstract

The concern of this paper is obtaining baseline values for the number of forest fires as a function of time and location and other explanatories. A model is developed and applied to a large data set from Federal lands in the state of Oregon. To proceed the data are grouped into small spatial-temporal cells (voxels). Fires are rare so there are many of these voxels with no fires. In fact there are so many such cells that in the analyses presented a sample is taken to make the work manageable. The paper sets down a likelihood for the sampled data and fits a generalized additive model involving location, elevation and day of the year as explanatories.

Keywords: Forest fires; generalized additive model; Oregon; risk analysis; sampled data; wildfires

1 Introduction

Forest fires represent a problem of considerable societal importance. We mention the following report that appeared in the San Francisco Chronicle of 7/16/2002,

> *... Nearly two weeks ago, the Forest Service used up the entire $321 million budgeted for firefighting in 2002. It is expected to spend another $645 million by the end of the year. ...Wildfires have already burned more than 3.3 million acres this year, more than twice the yearly average over the last decade.*

The concern here is the development of a risk model for use in estimating the probability of a forest fire taking place at a particular location and time as a function of those and other explanatory variables. The work is implemented for the case of a fine grid of cells and an accompanying large data set. An analysis is carried out for a region surrounding the state of Oregon, henceforth referred to as Oregon, and employing: location, elevation and day of year as explanatories. The elements of the approach are:

1. a spatial-temporal point process and associated covariates,
2. likelihood-based inferential methods developed for such processes,
3. approximation of the point process by a 0-1 valued process on a lattice,
4. a sampling of the 0; *i.e.* no-fire cells,
5. generalized additive model technology.

Oregon was chosen for this pilot study because of its approximate rectangular geographic shape and its high rate of fires. The work presented here has as its goals model development and the estimation of baseline values for future work. The model contains but a few explanatory variables. Employing a sample of the no-fire cases rather than all the no-fire cases available is an unusual aspect of the approach presented.

The sections of the paper are: Introduction, Risk assessment, Previous statistical work, Model and analysis development, The data, Results and Discussion.

Risk assessment is familiar to Terry Speed. His papers include: reviews of the procedures that have been employed in the nuclear industry [24], [27], an analysis concerning a ship following a specified sea route [26], and an analysis of risk to levees from adverse weather conditions [25]. The first mentioned contains the following wonderful exchange with the Director of Technology, U.S. National Transportation Safety Board,

> *Dear Professor Speed,*
>
> *In response to your aerogramme of April 5, 1977, the Chairman's statement concerning the chances of two jumbo jets colliding (6 million to one) has no statistical validity nor was it intended to be a rigorous or precise probability statement. The statement was made to emphasize the intuitive feeling that such an occurrence has a very remote but not impossible chance of happening.*
>
> *Thank you for your interest in this regard.*
>
> *Sincerely yours ...*

From these papers and personal observation it is clear that Terry knows a lot about taking risks. May he continue to have fun doing so for many years.

2 Risk assessment

Probabilistic risk assessment can be defined as the process of estimating, for some class, the probabilities of hazardous events taking place within a specified time period and in a specified context. Such an assessment often proceeds by reducing a particular complex system to its simpler components. This is followed by the fitting and validation of stochastic models associated with the components. Typically large doses of substantive subject matter are required in such modeling and data analysis projects.

This paper is concerned with the case of forest fires. Wildfires are a natural disturbance in virtually all the world's ecosystems and the annual losses are staggering, see the *Chronicle* report above. It seems clear that fire occurrence depends on local conditions such as: location, elevation, wind velocity, precipitation, temperature, air humidity, topography, litter type, level of suppression amongst other explanatories. In our work fire ignition will be viewed as a random phenomenon. A pertinent conceptual model is: when the temperature (or some related latent variable) at a given location exceeds a threshold, depending on the local conditions, a fire breaks out. The latent

variable will depend on explanatories such as those listed above. In the work below the logit transform will be employed. It corresponds to a latent variable with a logistic distribution.

3 Previous statistical work

Often a Poisson model has been employed for the number of fires, see Dayananda [3], Poulin-Costello [17], Mandallaz and Ye [8, 9]. In other cases it is a logistic: Chou *et al.* [2], Poulin-Costello [17], Martell *et al.* [11]. In another approach McKenzie *et al.* [12] use multiple regression and regression trees. Markov chain models are employed in Martell [10]. Peng and Schoenberg's works, [22], [16], relate wildfire incidence to temperature, precipitation, fuel moisture and fire history for Los Angeles County. These researchers find that time expired since a location has burned previously appears important. Roads *et al.* [20] use a regression model for fire occurrence with 6 fire danger indices by fuel type as explanatories. There is also spatial autocorrelation. [18] provide a review.

4 Model and analysis development

4.1 The spatial-temporal conditional intensity function

To begin suppose that the space-time domain is broken up into voxels $(x,x+dx] \times (y,y+dy] \times (t,t+dt]$. Consider the spatial-temporal point process, N, with conditional intensity function assumed to exist and defined by

$$\lambda(x,y,t) = Prob\{dN(x,y,t) = 1|H_t\}/dxdydt$$

where $dN(x,y,t) = N(dx,dy,dt)$ counts the number of fires in the voxel where H_t is the history of the process N up to and including time t. Supposing that λ contains a parameter θ the log-likelihood function will be written

$$L(\theta) = \int_0^T \int_x \int_y log[\lambda(x,y,t|\theta)]dN(x,y,t) - \int_0^T \int_x \int_y \lambda(x,y,t|\theta)dxdydt \quad (1)$$

see Fishman and Snyder [5]. Explanatories may appear in λ but are presently suppressed in the notation.

Asymptotics of maximum likelihood estimates based on point process likelihoods have been developed in Ogata [15], Sagalovsky [21], Rathbun [19], Schoenberg [23]. The last two papers focus on the spatial-temporal case.

Consider next practical approaches to using the log-likelihood (1) in practice.

4.1.1 Approach 1

The second term of (1) is the awkward one in our case because it covers such a large area in space-time. It will be approximated. One way to do this is by sampling points from an independent of N spatial Poisson process, $M(dx,dy,dt)$, of rate π on the space $\mathcal{R} \times [0,T]$. Then the log-likelihood (1) may be approximated by

$$\int_0^T \int_x \int_y log[\lambda(x,y,t|\theta)]dN(x,y,t) \; - \; \int_0^T \int_x \int_y \lambda(x,y,t|\theta)dM(x,y,t)/\pi. \qquad (2)$$

Averaging over M, but holding the process N fixed, the expected value of (2) is (1). The expected number of points in the approximating sum of the second term is π times the volume of $\mathcal{R} \times [0,T]$.

4.1.2 Approach 2

A model that is often simpler to deal with follows. It is an approximation to Approach 1. Replace the spatial-temporal point process, $N(dx,dy,dt)$, by a 0-1 valued process $N_{x,y,t}$ on a lattice with $N_{x,y,t} = 1$ if there is a fire in the corresponding voxel and by 0 otherwise for (x,y) *in* $\mathcal{R}$; $t = 0,...,T-1$. (This idea was used to advantage in [1].) Suppose then that

$$Prob\{N_{x,y,t} = 1|H_{t-1}\} \; = \; \lambda_{x,y,t}$$

with H_t the history up to and including t. A Bernoulli approximation to the log likelihood (1) is now

$$\sum_{x,y,t} N_{x,y,t} \; log(\lambda_{x,y,t}) \; + \; \sum_{x,y,t} (1 - N_{x,y,t}) \; log \; (1 \; - \; \lambda_{x,y,t}) \qquad (3)$$

In the present case there are many voxels for which $N_{x,y,t}$ is 0 and so the second sum in (3) contains many terms. To deal with this, randomly select 0-voxels with probability π and include them alone in the analysis. (In what follows there will actually be two-stage sampling with $\pi = \pi_1\pi_2$. This is to possibly obtain more efficient estimates. The estimation procedure then involves two types of uncertainty; one from the fire process itself and a second from the sampling of data points. The latter component will decrease with increasing π.)

To simplify the notation for the moment, index the voxels by k rather than x,y,t. Let S denote the collection of the voxels that had a fire and the sample of those that did not. In what follows it will be assumed that the N_k are independent given the explanatories. This condition will be relaxed in future work. Using the identity

$$Prob\{A|B\} \; = \; Prob\{B|A\}Prob\{A\}/Prob\{B\},$$

one has the conditional probability

$$Prob\{N_k = 1|H_{t-1}, \; k \; in \; S\} \; = \; \gamma_k \; = \; \lambda_k/(\lambda_k + (1 - \lambda_k)\pi)$$

with the complimentary probability for the event $N_k = 0$. Now by elementary algebra

$$logit\ \gamma_k\ =\ logit\ \lambda_k\ +\ log\ 1/\pi.$$

Using the indicated assumption of independence the log-likelihood based on the N_k for k in S is

$$\sum_{k\ in\ S} [N_k\ log(\ \gamma_k)\ +\ (1-N_k)\ log(1-\gamma_k)], \tag{4}$$

i.e. a Bernoulli likelihood. Appropriate uncertainty measures are often available for estimates obtained by maximizing (4), see the Appendix.

In the analyses $\eta\ =\ logit\ \lambda$ will be a "linear predictor" based on explanatories. To obtain estimates one can use a generalized linear model program, such as *glm*() of Splus, with an offset of $log(1/\pi)$ to carry out an analysis. In fact the generalized additive model program *gam*(), [6], will be employed below.

Above an assumption of independence was made. For the present this will be made reasonable by the inclusion of explanatories. For example, location is meant to handle the similarity of nearby values.

4.1.3 Approach 3

Another method begins, again, by approximating the log-likelihood (1) by one based on 0-1 variates. Further suppose once again the 0-voxels have been sampled. Let $\{\delta_{x,y,t}\}$ denote independent Bernoulli variates corresponding to the sampling with parameter π. They are to be independent of the fire process, N. Consider the approximate log-likelihood

$$\sum_{x,y,t} [\ N_{x,y,t}\ log(\ \lambda_{x,y,t})\ +\ (1-N_{x,y,t})\ log(1\ -\ \lambda_{x,y,t})\ \delta_{x,y,t}\ /\ \pi]$$

$$=\ \sum_{k\ in\ S} w_k[N_k\ log\ \lambda_k\ +\ (1-N_k)\ log(1-\lambda_k)]. \tag{5}$$

The w_k are weights that equal 1 at locations with fires and equal $1/\pi$ at the sampled locations with no fire. If $\pi\ =\ 1$ this reduces to the log-likelihood (3).

4.2 Assessing fit

Suppose a particular link function has been employed, *e.g.* the logit, and one wishes to assess it. The fitted linear predictor values may be assigned to the cells of a histogram. For each cell some of the corresponding N's will be 0 and some will be 1. The "number of 1's" divided by "the number of 1's" plus "the number of 0's" weighted up by $1/\pi$ provides a nonparametric estimate of the function $\lambda(\eta)\ =\ Prob\{N = 1|\eta = 1\}$ with η representing the linear predictor of the generalized additive employed model. This estimate may be compared to $\lambda(\eta|\hat{\theta})$ where $\hat{\theta}$ is the maximum likelihood estimate. An example is provided below, see Figure 6.

4.3 Predictions

Quantities of substantial interest to the Forest Service managers, for planning purposes, include probabilities such as

$$\lambda_{x,y,t+1} = Prob\{N_{x,y,t+1} = 1|H_t\}.$$

The fitted linear predictor and its statistics may be used to compute pertinent estimates. Provided some explanatories in the model are leading variables, and they can be reasonably forecast, one would have useful predictions.

The work presented here has in mind obtaining baseline values using models involving location, elevation and day of year only. The next stage of research will study improvements obtained when leading variables, *e.g.* based on meteorology, are included.

5 The data

In this pilot study fire occurrence data from Federal lands in the state of Oregon were used. The data consisted of locations and dates of every fire greater than 0.1 acre that occurred between April 26, 1989 and December 31, 1996. (There were 15,786 such fires.) The date refers to the day the fire started. The source of the fire location data was the USDA Forest Service, National Fire Occurrence Data Base [28]. Figure 1 provides an elevation map of Oregon with federal fire locations for 1990. The dark winding horizontal line at the top of the figure is the Columbia River. It provides the border of Oregon with Washington. Figure 2 shows the so-called "Federal Mask", *i.e.* the Federal lands in Oregon. Their area accounts for approximately 56% of the state.

Figure 3 is a time series plot of the square roots of fortnightly counts of fires over the time period of the study. A clear annual effect corresponding to the fire season may be seen. There is no serious suggestion of a trend.

Data on response variables and explanatories were selected as follows. The fires were those inside Federal lands, delineated by Figure 2. To get the quantities for beginning data analyses a two-stage sampling procedure was employed for the voxels without a fire because of the data management problem. In the first stage of the sampling a collection of days was selected, with each day having probability $\pi_1 = 0.1$ of being picked. In the second stage a proportion $\pi_2 = 0.0012$ of cells inside the Federal mask was picked. This resulted in a total of 73880 cases of which 15,786 were fire occurrences N_k of (4) equals 1 and 58094 were with N_k equal 0. The overall rate of fires experienced for the period of the study was 15786/(2760*209490) = .0000273 per km^2 per day. (There were 2760 days and a region of area 209490 km^2 in the study.)

To obtain the estimates, the function *gam*() of Splus, [6], was used to fit a Bernoulli model with an offset of $log\ 1/\pi$, a linear predictor

$$\eta_k = logit\ \lambda_k = g_1(x_k, y_k) + g_2(d_k) + g_3(e_k) \qquad (6a)$$

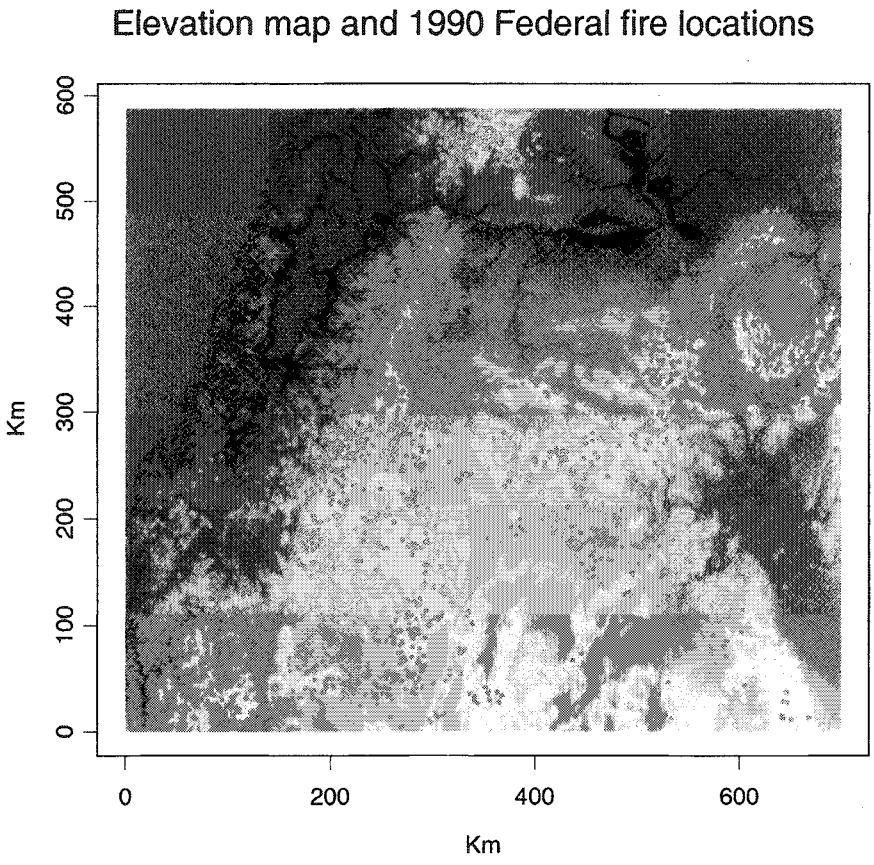


Figure 1: Federal fire locations in 1990. Dark blue corresponds to water and red circles to fire locations.

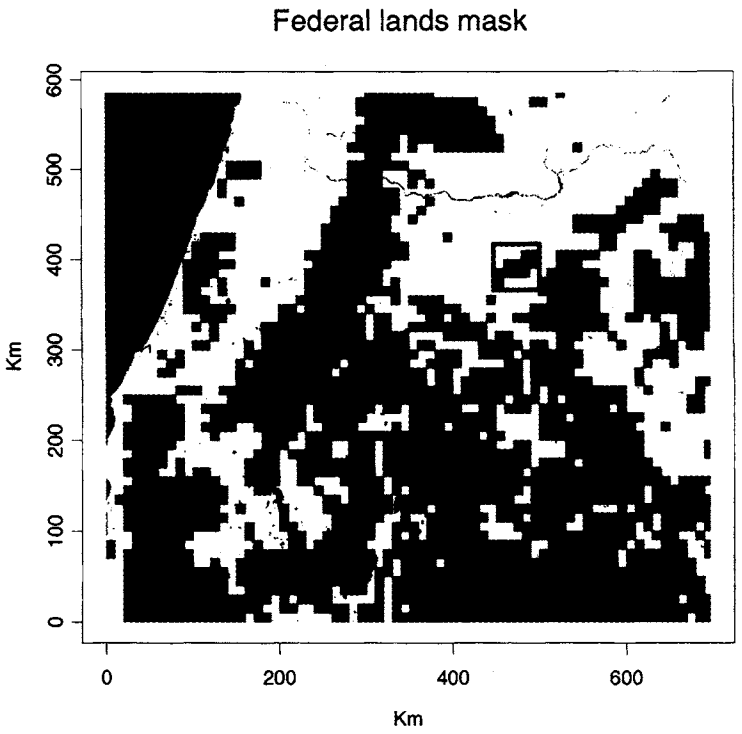


Figure 2: The Federal lands in Oregon. The box refers to the region taken for an example in Section 6. It is referred to as Region B.

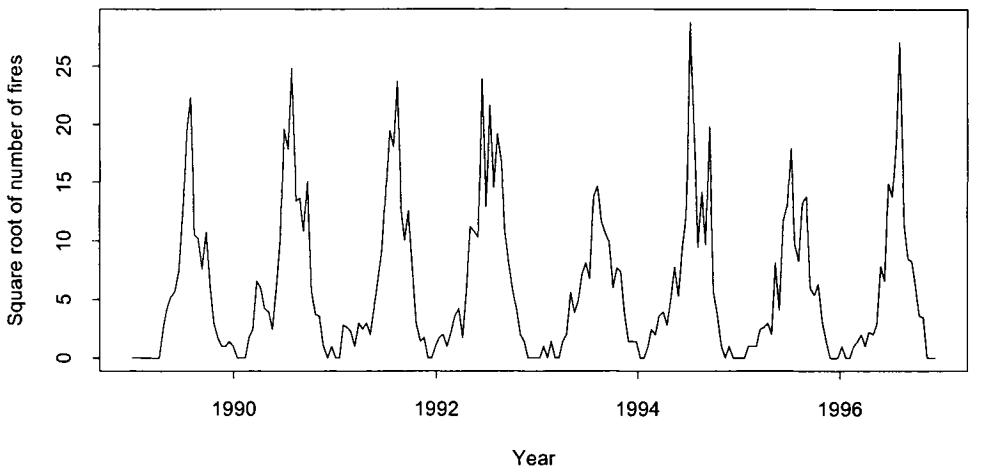


Figure 3: Square root of fortnightly count of Oregon fires in Federal lands 1989-1996.

and a probability mass function of

$$Prob\{N = 1|\eta\} = exp\{\eta\}/(\pi + (1-\pi)exp\{\eta\}). \quad (6b)$$

Here (x_k, y_k) corresponds to the locations (in meters) of coordinates of the k-th response, d_k corresponds to day in year and e_k refers to elevation at the location. The $g()$ are (nonparametric) smooth functions to be estimated. The functions g_2 and g_3 were splines with g_2 having period 1 year, [13]. The term g_1 was estimated via the scatterplot smoother *lo*.

One of the interesting aspects of the work is that the data set is quite large and awkward. Maps for the state of Oregon were extracted from the CD's, converted to ASCII files and then a sample was selected in two stages. Locations of fires were recorded in units of latitude and longitude. These were converted to metric units. Even with a small fraction (π=0.00012) of the available voxels, each gam run required approximately 10 minutes to run on a 900MHz computer. This meant that it was not easy to do exploratory data analysis.

6 Results

6.1 The fit

Figure 4 provides the estimated spatial effect for model (6a,b). It was obtained by the second approach described above. The results are presented in both perspective and contour form. Some general features apparent in this map are the low intensity level in the south-eastern region of Oregon, contour levels -2 and -3. This region is mostly high desert with few forests. Regions with the historically highest level, contour level 1, of fire appear to be in the Cascade Mountains.

Figure 5 provides plots of the estimated effect of elevation and day of the year. Unsurprisingly, the latter indicates amongst other things more fires in the summer as was evident in Figure 3. The elevation effect increases for those less than 2000m. The approximate 95% bounds are obtained by the jackknife procedure as discussed in the Appendix. Generally the bounds are small as might be expected given the large amount of data involved. The uncertainty bounds for the higher elevations are wide because there are not many high locations in the data set.

Figure 6 implements the method of Section 4.2 for assessing the appropriateness of the model (6a,b). It is a plot of the empirical relative frequencies of fires as points. These are computed after grouping the data into classes based on the linear predictor. Also provided, as a solid line, are the estimated probabilities from model (6a,b) and, as dashed lines, approximate binomial confidence bounds. The fit appears reasonable, but one does notice points above the curve on the right. This departure may disappear when other explanatories are included in the model.

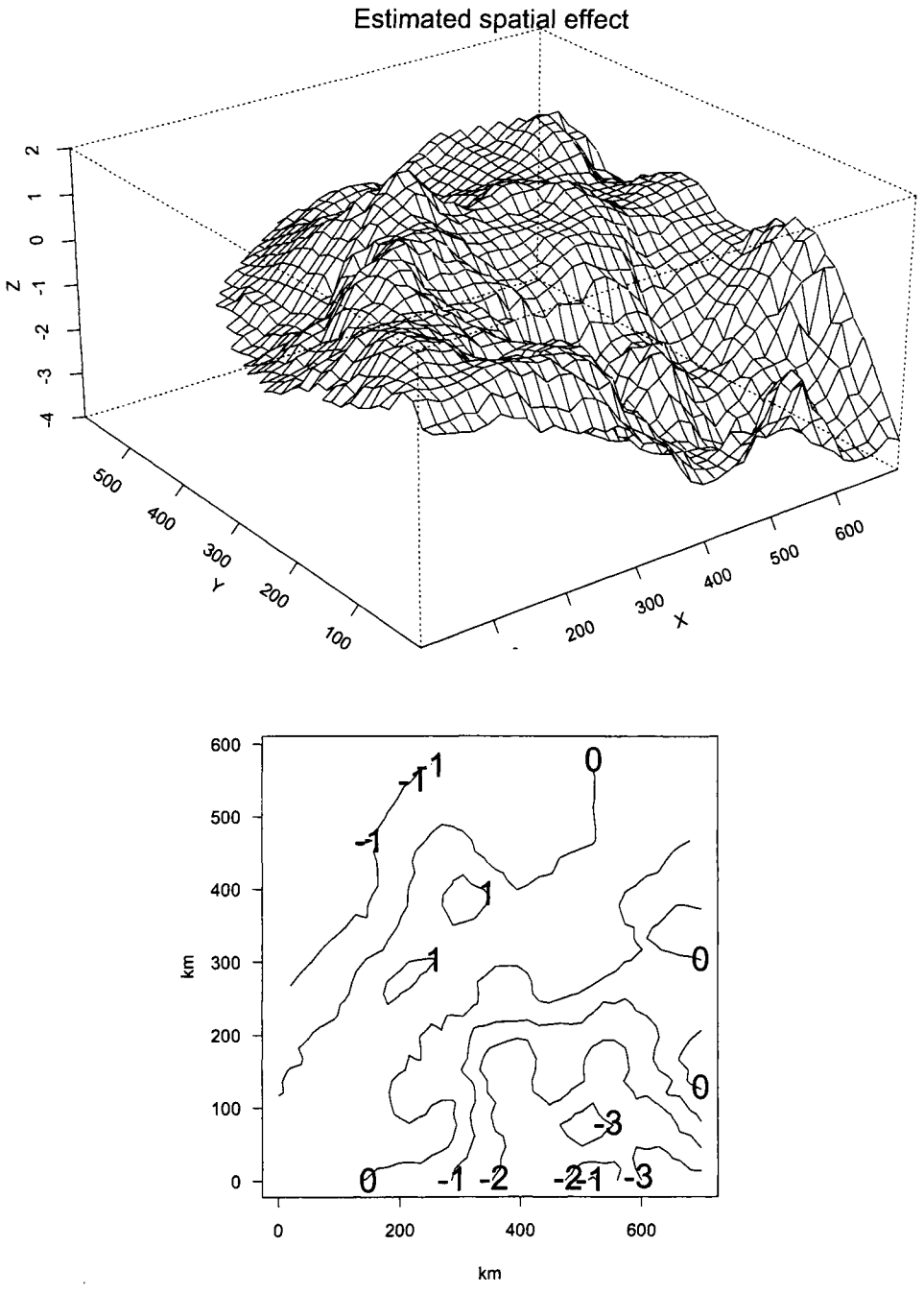


Figure 4: Estimated spatial effect, $\hat{g}_1$, for model (6a,b).

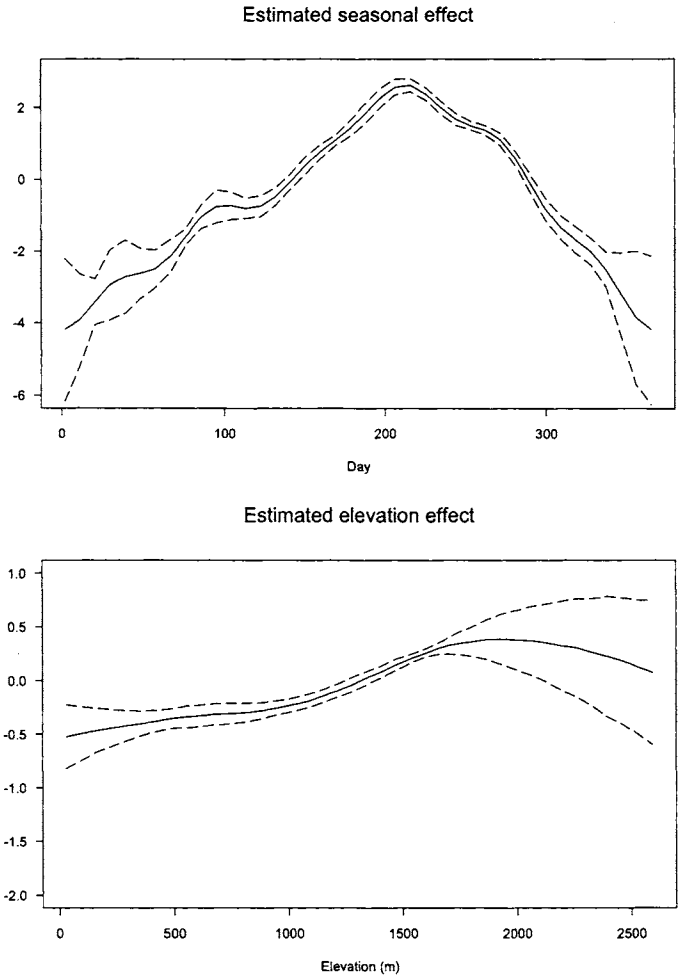


Figure 5: Estimated effects $\hat{g}_2$ of day in year and $\hat{g}_3$ of elevation for the model (6a,b). The dashed lines provide approximate marginal 95% bounds computed by a jackknife procedure.

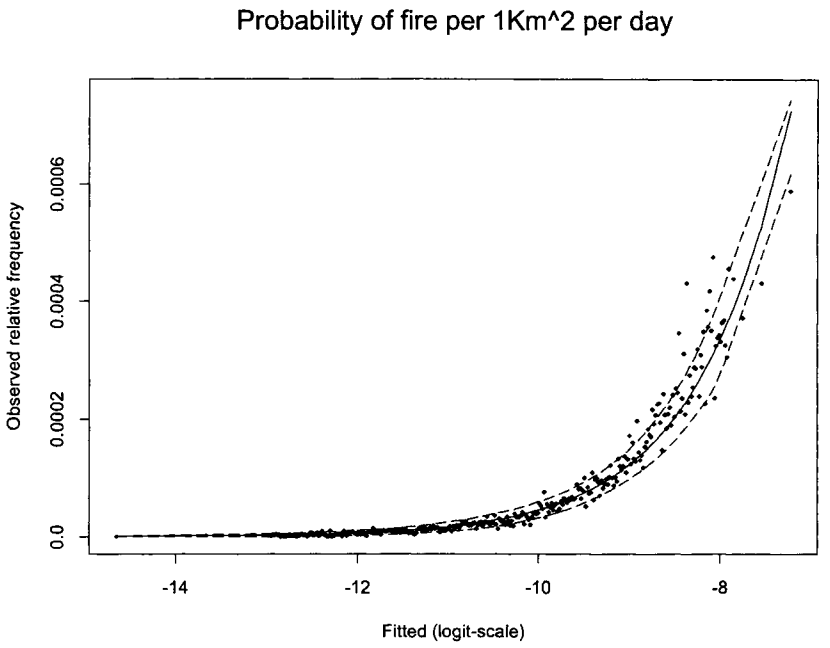


Figure 6: Observed relative frequencies of fire, after grouping the data into classes based on the fitted linear predictor, $\hat{\eta}$. The solid curve is the fitted logistic curve. The dashed lines are smoothed approximate 95% limits obtained via a binomial approximation.

6.2 An example

Forest managers are interested in estimates of the number of fires likely to occur in a given region during a given fire period. Amongst other things these estimates will help them allocate resources. As an example, using the output above, estimated fire probability values were produced for the shaded region within the box of Figure 2. This is the Heppner Ranger District of the Umatilla Forest in Oregon. It will be referred to as Region B. It is taken as representative of the sort of region of interest in wildfire risk estimates for the U.S. generally.

Figure 7 and Table 1 present some of the results. The solid line in Figure 7 gives the estimate of the monthly rates obtained by fitting the model with location, elevation, and day of year as explanatories. The shaped region gives approximate 95% marginal confidence limits for the estimate. These are obtained in the same jackknife computations as produced the errors bounds of Figure 5. In this case the linear predictions are perturbed by $\pm 2SE$ and converted to probabilities by the transform (6b). The resulting fire rates are seen to peak in the summer season with an estimated value of 6.74 fires for the month of August.

The dots in Figure 7 refer to the naive estimate of the rates of fires obtained by dividing the total number of fires for a month by the number of years of observation. The vertical lines give approximate 95% limits employing a Poisson approximation. One notes differences, but it may be remarked that the only physical characteristics of Region B being used in the present estimation are its elevations.

To be specific, managers are interested in things like

$$Prob\{i \text{ or more fires in July}\},\ i = 1, 2, ,...$$

Table 1 provides estimates of such together with approximate 95% confidence limits for the estimates. The count is based on the sum of Bernoulli variables. When the probabilities of the Bernoullis differ the distribution is called the Poisson-Binomial, [7]. These authors show that this distribution is well approximated by a Poisson in the case that the largest of the individual pixel probabilities is small. This seems to be the case in the present situation where the risk of fire is low, particularly since the pixels are small. For example the estimated probability of 7 or more fires in July for Region B is .324 with an approximate confidence interval of (.176,.520).

Here the only serious explanatory employed beyond location is elevation. The goal of future work is prediction using leading explanatories. It will be interesting to see the extent to which the confidence interval shrinks as these are brought into the model.

7 Discussion

The model, (6a,b), for the breaking out of wildfires has been set down and fit. The model was motivated by threshold considerations and is of nonparametric character.

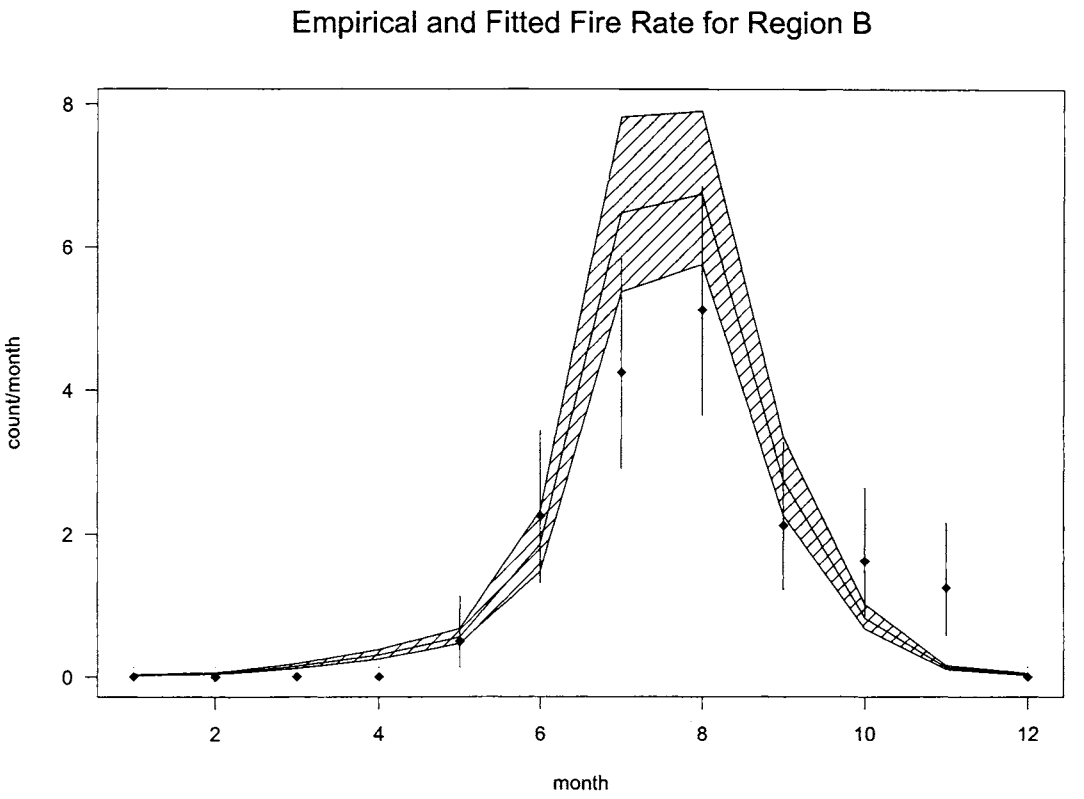


Figure 7: The solid central line gives the fitted rate of fires by month. The shaded region gives ± 2 SE limits. The points are the monthly empirical rates of fires. Vertical lines are ± 2 SE limits for the points.

i	*probability*	*confidence interval*
1	.989	(.967,.996)
2	.956	(.895,.983)
3	.887	(.771,.948)
4	.774	(.610,.883)
5	.628	(.441,.784)
6	.470	(.291,.658)
7	.324	(.176,.520)
8	.206	(.097,.386)
9	.121	(.049,.268)
10	.066	(.023,.174)
11	.033	(.010,.106)
12	.016	(.004,.060)
13	.007	(.001,.032)
14	.003	(.001,.016)
15	.003	(.000,.007)
16	.001	(.000,.003)
17	.000	(.000,.001)
18	.000	(.000,.001)
19	.000	(.000,.000)
20	.000	(.000,.000)

Table 1: Estimated probability of i or more fires and approximate 95% confidence limits for the month of July and Region B.

The assumptions include smooth dependence of the risk probability on location, elevation, day of the year. A logistic link function is employed.

There are many sources of variability present in addition to the random nature of wildfires. The simplest is that arising from using a sample of the no-fire voxels. The sampling was two-stage without replacement. Finite population correction factors were ignored in computing uncertainty estimates, but they are negligible. The sampling fraction, π, will be increased in future studies.

The principal source of the variability that shows itself in Figure 7 and Table 1 is possibly that arising from missing explanatories. To the extent possible this will be dealt with in future studies. Additional data sources to be included then are weather data from the Weather Information Management System [14], fire danger indices from NIFMID [14] and more topographic data, *e.g.*, aspect and distance to nearest population center.

This has been a pilot study with the purpose of obtaining baseline values. These will be used as standards for later forecasting models.

One difficulty needs to be mentioned. In using *gam*() it was found that the results of prediction runs depended on which other predictions were carried out at the same time. This may be due to the accumulation of roundoff error. Studies are continuing. The results presented in this paper are for runs of all the predictions being carried out at the same time. The reason for this choice is that the computing time required was less than, say, running one prediction at a time.

Acknowledgements

We thank Bob Burgan, Francis Fujioka, Rich Kimberlin, David Martell and the Referee for their helpful comments. Bob Burgan and Carolyn Chase supplied the data. Trevor Hastie and Phil Spector helped out when trouble arose employing *gam*(). Alan Aager chased down some geography.

D. R. Brillinger, Department of Statistics, University of California, Berkeley,
brill@stat.berkeley.edu

H. K. Preisler, USDA, Forest Service, PO Box 245, Berkeley, CA 94701,
hpreisler@fs.fed.us

J. Benoit, USDA, Forest Service, 4955 Canyon Crest Dr., Riverside, CA 92507,
jbenoit@fs.fed.us

References

[1] D. R. Brillinger and H. K. Preisler. Two examples of quantal data analysis: a) multivariate point process, b) pure death process in an experimental design. In *Proc. XIII International Biometric Conference, Seattle*, pages 94–113, 1986.

[2] Y. H. Chou, R. A. Minnich, and R. A. Chase. Mapping probability of fire occurrence in San Jacinto Mountains, California, USA. *Environmental Management*, 17:129–140, 1993.

[3] P. W. A. Dayananda. Stochastic models for forest fires. *Ecological Modelling*, 3:309–313, 1977.

[4] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, New York, 1993.

[5] P. M. Fishman and D. L. Snyder. The statistical analysis of space-time point processes. *IEEE Transactions on Information Theory*, IT-22:257–274, 1976.

[6] T. Hastie. Generalized additive models. In J. M. Chambers and T. J. Hastie, editors, *Statistical Models in S*, pages 249–307. Wadsworth, Pacific Grove, 1992.

[7] J. L. Hodges and L. M. Le Cam. The Poisson approximation to the Poisson Binomial distribution. *Annals of Mathematical Statistics*, 31:737–740, 1960.

[8] D. Mandallaz and R. Ye. Statistical model for the prediction of forest fires. Report Project Minerve II, ETH Zurich, 1996.

[9] D. Mandallaz and R. Ye. Prediction of forest fires with Poisson models. *Canadian Journal of Forest Research*, 27:1685–1694, 1997.

[10] D. L. Martell. A Markov chain model of daily changes in the Canadian forest fire weather index. *International Journal of Wildland Fire*, 9:265–274, 1999.

[11] D. L. Martell, S. Otukol, and B. J. Stocks. A logistic model for predicting daily people-caused forest fire occurrence in Ontario. *Canadian Journal of Forest Research*, 17:394–401, 1987.

[12] D. L. McKenzie, D. L. Peterson, E. Alvarado, J. K. Agee, and R. A. Norheim. Spatial models of fire frequency in the Columbia River Basin. `http://silvae.cfr.washington.edu/people/dmck/crb.html`, 1998.

[13] H. A. Nielsen. Summary: periodic spline basis. *s-news*, June 29, 1999.

[14] NIFMD. `http://www.fs.fed.us/fire/planning/nist/kcfast.html`.

[15] Y. Ogata. The asymptotic behaviour of maximum likelihood estimators for stationary point processes. *Annals of the Institute of Statistical Mathematics*, 30:243–261, 1978.

[16] R. Peng and F. Schoenberg. Estimation of wildfire hazard using spatial-temporal fire history data. Technical report, Statistics Department, UCLA, 2001.

[17] M. Poulin-Costello. People-caused forest fire prediction using poisson and logistic regression. Master's thesis, Dept. of Math. and Stat., University of Victoria, Victoria, Canada, 1993.

[18] H.K. Preisler and D. Weise. Forest fire models. In A. H. El-Shaarawi and W. W. Piegorsch, editors, *Encyclopedia of Environmetrics*, pages 808–810. Wiley, Chichester, 2002.

[19] S. L. Rathbun. Asymptotic properties of the maximum likelihood estimator for spatio-temporal point processes. *Journal of Statistical Planning and Inference*, 51:55–74, 1996.

[20] J. O. Roads, F. M. Fujioka, and R. E. Burgan. Development of a seasonal fire weather forecast for the contiguous United States. In *Third Symposium of Fire and Forest Meteorology*, pages 99–102. American Meteorological Society, 2000.

[21] B. Sagalovsky. *Maximum likelihood and related estimation methods in point processes and point process systems*. PhD thesis, University of California, Berkeley, 1983.

[22] F. Schoenberg, R. Peng, Z. Huang, and P. Rundel. Exploratory analysis of Los Angeles County wildfire data. Technical report, Statistics Dept. UCLA, 2000.

[23] F. P. Schoenberg. *Assessment of point process models*. PhD thesis, University of California, Berkeley, 1997.

[24] T. P. Speed. Negligible probabilities and nuclear reactor safety: another misuse of probability. Technical report, Mathematics Dept., University of Western Australia, 1977.

[25] T. P. Speed. The possibility of future damage to the Leslie Salt Company concentrator and diversion levees caused by adverse weather conditions: an evaluation of probabilities. Technical report, Mathematics Dept., University of Western Australia, 1980.

[26] T. P. Speed. Tow route risk asessment. Technical report, University of Western Australia, 1980.

[27] T. P. Speed. Probabilistic risk assessment in the nuclear industry: Wash 1400 and beyond. In L. M. LeCam and R. Olshen, editors, *Neyman-Kiefer Conference*, Pacific Grove, 1983. Wadsworth.

[28] USDA. National Fire Occurrence Data Base. http://www.fs.fed.us/fire/fuelman.

Appendix

The use of the jackknife for the model (6a,b) may justified by assumptions that K is large, that the values

$$(N_k, x_k, y_k, d_k, e_k),\ k = 1, ..., K$$

may be treated as an i.i.d. sample from some distribution and that the statistics computed are approximately additive in some i.i.d. variates. The variances estimated are overall as opposed to conditional. One reference to the jackknife is Chapter 11 in Efron and Tibshirani [4] where it is discussed as an approximation to the bootstrap.

In the implementation employed, the standard errors of the estimates of $g_1(x,y)$, $g_2(d)$, and $g_3(e)$ are obtained by splitting the 7 years of data into 7 random segments of 365 days. The values obtained do not differ greatly from those produced by *gam*.

On the Likelihood of Improving the Accuracy of the Census Through Statistical Adjustment

David A. Freedman and Kenneth W. Wachter

Abstract

In this article, we sketch procedures for taking the census, making adjustments, and evaluating the results. Despite what you read in the newspapers, the census is remarkably accurate. Statistical adjustment is unlikely to improve on the census, because adjustment can easily put in more error than it takes out. Indeed, error rates in the adjustment turn out to be comparable to—if not larger than—errors in the census. The data suggest a strong geographical pattern to these errors even after controlling for demography, which contradicts a basic premise of adjustment. Complex demographic controls built into the adjustment mechanism turn out to be counter-productive.

Proponents of adjustment have cited "loss function analysis" to compare the accuracy of the census and adjustment, generally to the advantage of the latter. However, these analyses make assumptions that are highly stylized and quite favorable to adjustment. With more realistic assumptions, loss function analysis is neutral or favors the census. At the heart of the adjustment mechanism, there is a large sample survey—the post enumeration survey. The size of the survey cannot be justified. The adjustment process now consumes too large a share of the Census Bureau's scarce resources, which should be reallocated to other Bureau programs.

Keywords: Census; adjustment; heterogeneity; correlation bias; demographic analysis; dual-system estimation; non-sampling error; loss function analysis

1 Introduction

The census has been taken every ten years since 1790. Counts are used to apportion Congress and redistrict states. Furthermore, census data are the basis for allocating federal tax money to cities and other local governments. For such purposes, the geographical distribution of the population matters rather than counts for the nation as a whole. Data from 1990 and previous censuses suggested there would be a net undercount in 2000; the undercount would depend on age, race, ethnicity, gender, and—most importantly—geography. This differential undercount, with its implications for sharing power and money, attracted considerable attention in the media and the court-house.

There were proposals to adjust the census by statistical methods, but this is advisable only if the adjustment gives a truer picture of the population and its geographical

distribution. The census turns out to be remarkably good, despite the generally bad press reviews. Statistical adjustment is unlikely to improve the accuracy, because adjustment can easily put in more error than it takes out.

In this article, which is an expanded version of Freedman and Wachter [15], we sketch procedures for taking the census, making adjustments, and evaluating results. (A sketch is what you want: detailed descriptions cover thousands of pages.) We have new data on errors in the adjustment, and on geographical variation in error rates. We discuss alternative adjustments, and point out critical flaws in oft-cited methods for comparing the accuracy of the census and adjustment. We close with pointers to the literature, including citations to the main arguments for and against adjustment, and a summary of the policy recommendations that follow from our analysis.

2 The Census

The census is a sophisticated enterprise whose scale is remarkable. In round numbers, there are 10,000 permanent staff at the Bureau of the Census. Between October 1999 and September 2000, the staff opened 500 field offices, where they hired and trained 500,000 temporary employees. In spring 2000, a media campaign encouraged people to cooperate with the census, and community outreach efforts were targeted at hard-to-count groups.

The population of the United States is about 280 million persons in 120 million housing units, distributed across 7 million "blocks," the smallest pieces of census geography. (In Boston or San Francisco, a block is usually a block; in rural Wyoming, a "block" may cover a lot of pastureland.) Statistics for larger areas like cities, counties, or states are obtained by adding up data for component blocks.

From the perspective of a census-taker, there are three types of areas to consider. In "city delivery areas" (high-density urban housing with good addresses), the Bureau develops a Master Address File. Questionnaires are mailed to each address in the file. About 70 percent of these questionnaires are filled out and returned by the respondents. Then "Non-Response Followup" procedures go into effect: for instance, census enumerators go out several times and attempt to contact non-responding households, by knocking on doors and working the telephone. City delivery areas include roughly 100 million housing units.

"Update/leave" areas, comprising less than 20 million households, are mainly suburban and have lower population densities; address lists are more difficult to construct. In such areas, the Bureau leaves the census questionnaire with the household while updating the Master Address File. Beyond that, procedures are similar to those in the city delivery areas.

In "update/enumerate" areas, the Bureau tries to enumerate respondents—by interviewing them—as it updates the Master Address File. These areas are mainly rural, and post-office addresses are poorly defined, so address lists are problematic. (A typical address might be something like Smith, Rural Route #1, south of Willacoochee, GA.)

Perhaps a million housing units fall into such areas. There are also special populations that need to be enumerated—institutional (prisons and the military), as well as non-institutional "group quarters." (For instance, 12 nuns sharing a house in New Orleans are living in group quarters.) About 8 million persons fall into these two categories.

3 Demographic Analysis

Demographic analysis estimates the population using birth certificates, death certificates, and other administrative record systems. The estimates are made for national demographic groups defined by age, gender, and race (Black and non-Black). Estimates for sub-national geographic areas like states are currently not available. According to demographic analysis, the undercount in 1970 was about 3 percent nationally. In 1980, it was 1 to 2 percent, and the result for 1990 was similar. The undercount for Blacks was estimated at about 5 percentage points above non-Blacks, in all three censuses.

Demographic analysis starts from an accounting identity:

$$\text{Population} = \text{Births} - \text{Deaths} + \text{Immigration} - \text{Emigration}.$$

However, data on emigration are incomplete. And there is substantial illegal immigration, which cannot be measured directly. Thus, estimates need to be made for illegals, but these are (necessarily) somewhat speculative.

Evidence on differential undercounts depends on racial classifications, which may be problematic. Procedures vary widely from one data collection system to another. For the census, race of all household members is reported by the person who fills out the form. In Census 2000, respondents were allowed for the first time to classify themselves into multiple racial categories: this is a good idea from many perspectives, but creates a discontinuity with past data. On death certificates, race of decedent is often determined by the undertaker. Birth certificates show the race of the mother and (usually) the race of father; procedures for ascertaining race differ from hospital to hospital. A computer algorithm is used to determine race of infant from race of parents.

Prior to 1935, many states did not have birth certificate data at all; and the further back in time, the less complete is the system. This makes it harder to estimate the population aged 65 and over. In 2000, demographic analysis estimates the number of such persons starting from Medicare records. Despite its flaws, demographic analysis has generally been considered to be the best yardstick for measuring census undercounts. Recently, however, proponents of adjustment have favored another procedure, the DSE ("Dual System Estimator").

4 DSE—Dual System Estimator

The DSE is based on a special sample survey done after the census—a PES ("Post Enumeration Survey"). The PES of 2000 came to be called ACE ("Accuracy and Coverage Evaluation Survey"): acronyms seem to be unstable linguistic compounds. The

ACE sample covers 25,000 blocks, containing 300,000 housing units and 700,000 people. An independent listing is made of the housing units in the sample blocks, and persons in these units are interviewed after the census is complete. This process yields the "P-sample."

The "E-sample" comprises the census records in the same blocks, and the two samples are then matched up against each other. In most cases, a match validates both the census record and the PES record. A P-sample record that does not match to the census may be a "gross omission," that is, a person who should have been counted in the census but was missed. Conversely, a census record that does not match to the P-sample may be an "erroneous enumeration," in other words, a person who got into the census by mistake. For instance, a person can be counted twice in the census—because he sent in two forms. Another person can be counted correctly but assigned to the wrong unit of geography: she is a gross omission in one place and an erroneous enumeration in the other.

Of course, an unmatched P-sample record may just reflect an error in ACE; likewise, an unmatched census record could just mean that the corresponding person was found by the census and missed by ACE. Fieldwork is done to "resolve" the status of some unmatched cases—deciding whether the error should be charged against the census or ACE. Other cases are resolved using computer algorithms. However, even after fieldwork is complete and the computer shuts down, some cases remain unresolved. Such cases are handled by statistical models that fill in the missing data. The number of unresolved cases is relatively small, but it is large enough to have an appreciable influence on the final results (Section 9).

Movers—people who change address between census day and ACE interview—represent another complication. Unless persons can be correctly identified as movers or non-movers, they cannot be correctly matched. Identification depends on getting accurate information from respondents as to where they were living at the time of the census. Again, the number of movers is relatively small, but they are a large factor in the adjustment equation (Section 9). More generally, matching records between the ACE and the census becomes problematic if respondents give inaccurate information to the ACE, or the census, or both. Thus, even cases that are resolved though ACE fieldwork and computer operations may be resolved incorrectly. We refer to such errors as "processing error."

The statistical power of the DSE comes from matching, not from counting better. In fact, the E-sample counts came out a bit higher than the P-sample counts, in 1990 and in 2000: the census found more people than the post enumeration survey.[1] As the discussion of processing error shows, however, matching (like so many other things) is easier said than done.

Some persons are missed both by the census and by ACE. Their number is estimated using a statistical model, assuming that ACE is as likely to find people missed by the census as people counted in the census—"the independence assumption." Following this assumption, a gross omission rate estimated from the people found by ACE is

extrapolated to the sort of people who are unlikely to be found, although the gross omission rate for the latter group may well be different. Failures in the independence assumption lead to "correlation bias." Data on processing error and correlation bias will be presented later.

5 Small-Area Estimation

The Bureau divides the population into "post strata" defined by demographic and geographic characteristics. For Census 2000, there were 448 post strata. One post stratum, for example, consisted of Asian male renters age 30–49, living anywhere in the United States. Another post stratum consisted of Blacks age 0–17 (male or female) living in owner-occupied housing in big or medium-size cities with high mail return rates, across the whole country. Persons in the P-sample are assigned to post strata on the basis of information collected during the ACE interview. (For the E-sample, assignment is based on the census return.)

Each sample person gets a "weight." If the Bureau sampled 1 person in 500, each sample person would stand for 500 in the population and be given a weight of 500. The actual sampling plan for ACE is more complex, so different people get different weights, ranging from 10 to 6000.[2] To estimate the total number of gross omissions in a post stratum, the Bureau simply adds the weights of all ACE respondents who were identified as (i) gross omissions and (ii) being in the relevant post stratum.

To a first approximation, the estimated undercount in a post stratum is the difference between the estimated numbers of gross omissions and erroneous enumerations.[3] The Bureau computes an "adjustment factor"; when multiplied by this factor, the census count for a post stratum equals the estimated true count from the DSE. About two-thirds of the adjustment factors exceed 1: these post strata are estimated to have undercounts. The remaining post strata are estimated to have been overcounted by the census; their adjustment factors are less than 1.[4]

How does the Bureau adjust small areas like blocks, cities, or states? Take any particular area. Each post stratum has some number of persons counted by the census in that area. (The number may be zero.) This census number is multiplied by the adjustment factor for the post stratum. The process is repeated for all post strata, and the adjusted count is obtained by adding the products; complications due to rounding are ignored for now. The adjustment process makes the "homogeneity assumption," that undercount rates are constant within each post stratum across all geographical units. This is not plausible, and was strongly contradicted by census data on variables related to the undercount. Failures in the homogeneity assumption are termed "heterogeneity." Ordinarily, samples are used to extrapolate upwards, from the part to the whole. In census adjustment, samples are used to extrapolate sideways, from 25,000 sample blocks to each and every one of the 7 million blocks in the United States. That is where the homogeneity assumption comes into play.

The political debate over adjustment is often framed in terms of sampling: "sam-

pling is scientific." However, from a technical perspective, sampling is not the issue. The crucial questions are about the size of processing errors, and the validity of statistical models for missing data, correlation bias, and homogeneity—in a context where the margin of allowable error is relatively small.

6 State Shares

All states would gain population from adjustment. Some, however, gain more than others. In terms of population share, gains and losses must balance—a subtle point often overlooked in the political debate. In 2000, even more than 1990, share changes were tiny. According to Census 2000, for example, Texas had 7.4094 percent of the population. Adjustment would have given it 7.4524 percent, an increase of $7.4524 - 7.4094 = .0430$ percent, or 430 parts per million. The next biggest winner was California, at 409 parts per million; third was Georgia, at 88 parts per million.

Ohio would have been the biggest loser, at 241 parts per million; then Michigan, at 162 parts per million. Minnesota came third in this sorry competition, at 152 parts per million. The median change (up or down) is about 28 parts per million. These changes are tiny, and most are easily explained as the result of sampling error in ACE. "Sampling error" means random error introduced by the luck of the draw in choosing blocks for the ACE sample; you get a few too many blocks of one kind or not quite enough of another: the contrast is with "systematic" or "non-sampling" error like processing error.

The map (Figure 1) shows share changes that exceed 50 parts per million. Share increases are marked "+"; share decreases, "−". The size of the mark corresponds to the size of the change. As the map indicates, adjustment would have moved population share from the Northeast and Midwest to the South and West. This is paradoxical, given the heavy concentrations of minorities in the big cities of the Northeast and Midwest—and political rhetoric contending that the census shortchanges such areas ("statistical grand larceny," according to New York's ex-Mayor Dinkins). One explanation for the paradox is correlation bias. The older urban centers of the Northeast and Midwest may be harder to reach, both for census and for ACE.

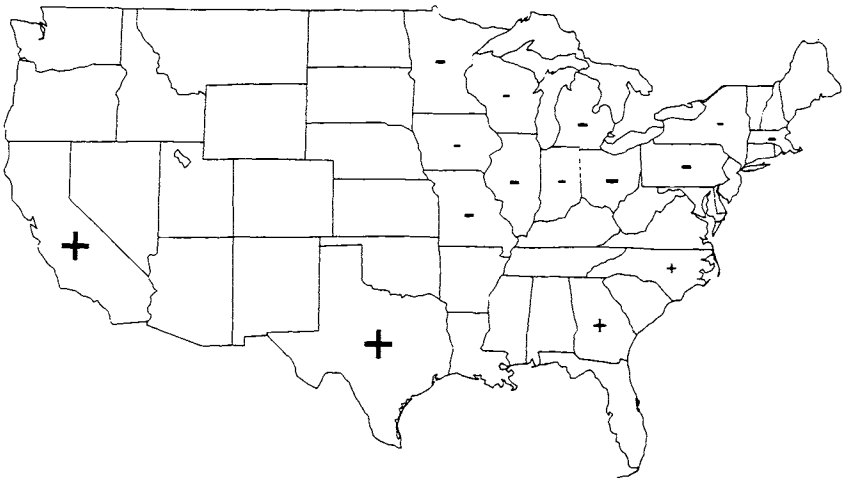


Figure 1: ACE adjustment: State share changes exceeding 50 parts per million[5]

7 The 1990 Adjustment Decision

Table 1: Errors in the adjustment of 1990

The adjustment	+5.3	
Processing error	−3.6	
Corrected adjustment		+1.7
Correlation bias		+3.0
Demographic analysis		+4.7

A brief look at the 1990 adjustment decision provides some context for discussions of Census 2000. In July 1991, the Secretary of Commerce declined to adjust Census 1990. At the time, the undercount was estimated as 5.3 million persons (Table 1). Of this, 1.7 million persons were thought by the Bureau to reflect processing errors in the post enumeration survey, rather than census errors. Later research has shown the 1.7 million to be a serious underestimate. Current estimates range from 3.0 million to 4.2 million, with a central value of 3.6 million. (These figures are all nation-wide, and net.) Thus, the bulk of the 1990 adjustment resulted from errors not in the census but in the PES. Processing errors generally inflate estimated undercounts, and subtracting them leaves a corrected adjustment of 1.7 million. (There is an irritating numerical coincidence here, as 1.7 million enters the discussion with two different roles.) Correlation

bias, estimated at 3.0 million, works in the opposite direction, and brings the undercount estimate up to the demographic analysis figure of 4.7 million.[6] The message is simple: on the scale of interest, most of the estimated undercount is noise.

8 Evaluating Census 2000

We see widespread—although by no means universal—agreement on two chief points. First, Census 2000 succeeded in reducing differential undercounts from their 1990 levels. Second, there are serious questions about the accuracy of proposed statistical adjustments. Mistakes in statistical adjustments are nothing new. Studies of the 1980 and 1990 data have quantified, at least to some degree, the three main kinds of error: processing error, correlation bias, and heterogeneity. In the face of these errors, it is hard for adjustment to improve on the accuracy of census numbers for states, counties, legislative districts, and smaller areas. Statistical adjustment can easily put in more error than it takes out, because the census is already very accurate.

In 1990, there were many studies on the quality of the adjustment. For 2000, evaluation data are only beginning to be available. However, the Bureau's preliminary estimates, based largely on the experience of 1990, suggested that processing error in ACE contributes about 2 million to the estimated undercount of 3.3 million.[7] (Errors in ACE will be discussed in more detail, below.) Errors in the ACE statistical operations may from some perspectives have been under better control than they were in 1990. But error rates may have been worse in other respects. There is continuing research, both inside the Bureau and outside, on the nature of the difficulties. The Bureau investigated a form of error called "balancing error"—essentially, a mismatch between the levels of effort in detecting gross omissions or erroneous enumerations. We think that troubles also occurred with a new treatment of movers (discussed in the next section) and duplicates. Some 25 million duplicate persons were detected in various stages of the census process, and removed.[8] But how many slipped through?

Besides processing error, correlation bias is an endemic problem that make it extremely difficult for adjustment to improve on the census. Correlation bias is the tendency for people missed in the census to be missed by ACE as well. Correlation bias in 2000 probably amounted, as it did in 1990, to millions of persons. These people cannot be evenly distributed across the country. If their distribution is uneven, the DSE creates a distorted picture of census undercounts. Heterogeneity is also endemic: undercount rates differ from place to place within population groups treated as homogeneous by adjustment. Heterogeneity puts limits on the accuracy of adjustments for areas like states, counties, or legislative districts. Studies of the 1990 data, along with more recent work discussed below, show that heterogeneity remains a serious concern.[9]

9 The Adjustment Decision: March 2001

In March 2001, the Secretary of Commerce—on the advice of the Census Bureau—decided to certify the census counts rather than the adjusted counts for use in redistricting (drawing congressional districts within state).[10] The principal reason was that, according to demographic analysis, the census had overcounted the population by perhaps 2 million people. Proposed adjustments would have added another 3 million people, making the overcounts even worse. Thus, demographic analysis and ACE pointed in opposite directions. The three population totals are shown in Table 2.[11]

Table 2: The population of the United States

Demographic analysis	279.6 million
Census 2000	281.4 million
ACE	284.7 million

If demographic analysis is right, there is a census overcount of .7 percent. If ACE is right, there is a census undercount of 1.2 percent. Demographic analysis is a particularly valuable benchmark, because it is independent (at least in principle) of both the census and the post enumeration survey that underlies proposed adjustments. While demographic analysis is hardly perfect, it was a stretch to blame demographic analysis for the whole of the discrepancy with ACE. Instead, the discrepancy pointed to undiscovered error in ACE. Evaluations of the ACE data are ongoing, so conclusions must be tentative. However, there was some information on missing data and on the influence of movers available in March 2001, summarized in Table 3.[12]

Table 3: Missing data in ACE, and impact of movers

Non-interviews	
P-sample	3 million
E-sample	6 million
Imputed match status	
P-sample	3 million
E-sample	7 million
Inmovers and outmovers	
Imputed residence status	6 million
Outmovers	9 million
Inmovers	13 million
Mover gross omissions	3 million

These figures are weighted to national totals, and should be compared to (i) a total census population around 280 million, and (ii) errors in the census that may amount to a

few million persons. For some 3 million P-sample persons, a usable interview could not be completed; for 6 million, a household roster as of census day could not be obtained (lines 1 and 2 in the table). Another 3 million persons in the P-sample and 7 million in the E-sample had unresolved match status after fieldwork: were they gross omissions, erroneous enumerations, or what? For 6 million, residence status was indeterminate—where *were* they living on census day? (National totals are obtained by adding up the weights for the corresponding sample people; non-interviews are weighted out of the sample and ignored in the DSE, but we use average weights.) If the idea is to correct an undercount of a few million in the census, these are serious gaps. Much of the statistical adjustment therefore depends on models used to fill in missing data. Efforts to validate such models remain unconvincing, despite some over-enthusiastic claims in the administrative and technical literature.[13]

The 2000 adjustment tried to identify both inmovers and outmovers, a departure from past practice. Gross omission rates were computed for the outmovers and applied to the inmovers, although it is not clear why rates are equal within local areas. For outmovers, information must have been obtained largely from neighbors. Such "proxy responses" are usually thought to be of poor quality, inevitably creating false non-matches and inflating the estimated undercount. As the table shows, movers contribute about 3 million gross omissions (a significant number on the scale of interest) and ACE failed to detect a significant number of outmovers. That is why the number of outmovers is much less than the number of inmovers. Again, the amount of missing data is small relative to the total population, but large relative to errors that need fixing. The conflict between these two sorts of comparisons is the central difficulty of census adjustment. ACE may have been a great success by the ordinary standards of survey research, but not good enough for adjusting the census.

10 Gross or Net?

Errors can reported either "gross" or "net," and there are many possible ways to refine the distinction. Given the uncertainties, we find that error rates in the adjustment are comparable to—if not larger than—error rates in the census, whether gross or net. For context, proponents of adjustment have lately favored measuring errors in the census on a gross basis rather than net, citing concerns about geographical imbalances. Some places may have an excess number of census omissions while other places will have an excess number of erroneous inclusions. Such imbalances could indeed be masked by net error rates. However, adjustment is hardly a panacea for geographical imbalance. The adjustment mechanism allows cancellation of errors within post strata—the homogeneity assumption at work. Much of the gross error is netted out, post stratum by post stratum; the rest is spread uniformly across geography within post strata. Adjustment fixes geographical imbalances in the census only if you buy the homogeneity assumption: location is accident, demography is destiny.

Proponents of adjustment have also objected to a comparison between undercount

estimates and estimated processing error (as in Section 7), on the grounds that net errors can after all be negative. We are unsympathetic to this complaint. For most areas with substantial populations—all states and 433/435 congressional districts—the adjustment is positive. Furthermore, estimated processing error is positive for all states and all congressional districts.[14] For such areas, adjustment adds to the population totals, and these increments mainly result from errors in the adjustment rather than errors in the census. All that said, a comparison of gross error rates will be instructive: see Table 4, where rows are numbered to match the paragraphs below.

Table 4: Errors in the census and ACE. Millions of persons. March figures.

	Positive Error	Negative Error	Gross Error	Paragraph
Census	1.00	4.26	5.26	(i)
ACE	1.75	.90	2.65	(ii)
Census	.10	2.51	2.61	(iii)
Census	3.1	6.4	9.5	(iv)
ACE			12.8	(v)

Sign convention. A "positive" error makes a population estimate too high, while a negative error makes the estimate too low. Thus, correlation bias counts as a negative error for ACE. Generally, processing error in ACE is positive.

(i) ACE would have added 4.26 million persons nationwide in certain post strata, and subtracted 1.00 million in other post strata. The net change is $4.26 - 1.00 = 3.26$ million but the gross is $4.26 + 1.00 = 5.26$ million. In effect, these are net and gross errors in the census, as estimated by ACE.[15]

(ii) For comparison, the Bureau's estimated biases in ACE (as of March 2001) add 1.75 million persons to the adjustment in certain post strata and subtract .90 million in other post strata, for a net error of $1.75 - .90 = .85$ million and a gross error of $1.75 + .90 = 2.65$ million.[16] The latter is about half the proposed gross adjustment. Thus, proponents of adjustment must think there are 5.26 million gross errors in the census that are detected by ACE and can be fixed by adjustment. But half of these represent errors in the adjustment mechanism itself, rather than errors in the census—even on the March figures for biases in ACE, which were largely based on extrapolation from 1990.

(iii) This comparison, however, is far too generous to ACE, because biases in ACE are counted against the census. Instead, we can estimate errors in the census from a bias-corrected ACE, sticking with the March figures for bias. On this basis, gross error rates in the census are virtually the same as those in ACE. In line (iii), for instance, the negative error for the census can be computed from the data in lines (i) and (ii), as

$4.26 - 1.75 = 2.51$ million, and similarly for the positive error.[17] The gross error in the census is 2.61 million; in ACE, 2.65 million.

(iv) Some number of persons were left out of Census 2000 and some were counted in error. Even if ACE had been done with surgical precision, there is no easy way to estimate the size of these two errors separately. Many people were counted a few blocks away from where they should have been counted: they are both gross omissions and erroneous enumerations. Many other people were classified as erroneous enumerations because they were counted with insufficient information for matching; they should also come back as gross omissions in the ACE fieldwork. With some rough-and-ready allowances for this sort of double-counting, the Bureau estimated that 6–8 million people were left out of the census while 3–4 million were wrongly included, for a gross error in the census of 9–12 million; the Bureau's preferred values are 6.4 and 3.1, for a gross of 9.5 million.[18] Much of this nets out within post strata: see line (i).

(v) For comparison, gross errors in ACE amount to 11.7 million after weighting to national totals, with an additional 1.1 million for correlation bias: here, cancellation is not allowed within post strata.[19] Doubtless, the 11.7 million double-counts some errors; and in any event, much of the error will net out within post strata. Still, on this basis, gross error rates in ACE are substantially larger than those in the census.

It is puzzling to see proponents of adjustment reciting gross error rates for the census, like those in line (iv) of the table, as if such data justified their position.[20] Errors that cancel within post strata cannot matter to the adjusters, or at least to those who care about logical consistency, because such errors—according to their theories—affect the accuracy neither of the census nor of the adjustment. Moreover, gross error rates in ACE are comparable to, if not larger than, gross error rates in the census.

11 Error Rates in ACE: October 2001

In October 2001, the Bureau decided not to adjust the census as a base for postcensal population estimates. This sounds even drier than redistricting, but $200 billion a year of tax money are allocated using such estimates. The decision was made after further analysis of the data, carried out between March and October. The Bureau added 2.2 million to the demographic analysis; and processing error in ACE went from 2 million to 5–6 million. Moreover, the Bureau confirmed that gross errors in ACE were well above 10 million, with another 15 million cases whose status remains to be resolved.[21] Any way you slice it, a large part of the adjustment comes about because of errors in the adjustment process rather than the census.

Before the October decision, we tried to reconcile the figures in Table 2 for the population of the United States—279.6 million from demographic analysis, 281.4 million from the census, and 284.7 million from ACE. There are good (albeit post hoc) arguments for increasing the demographic analysis figure, perhaps by 2 million; the census seemed about right to us, or even a little high; and in our view, the net processing error in ACE was probably 5–6 million, partially offset by correlation bias amounting to

2–3 million. In short, the Bureau's preliminary estimates for processing error in ACE needed to be doubled or tripled, and so did Bureau estimates for correlation bias. In the main, these forecasts are confirmed by the October decision document: U. S. Census Bureau [25]. Bureau estimates for correlation bias, however, are still based on a fiction—that is there is no correlation bias for women.[22]

12 Heterogeneity in 2000

Table 5: Measuring heterogeneity. In the first column, post stratification is either (i) by the Bureau's 448; or (ii) by the 64 post-stratum groups, that is, collapsing age and sex; or (iii) by the 16 evaluation post strata. "II" means whole-person substitutions, and "LA" is late census adds. In the last two columns, "P-S" stands for post strata; these are of three different kinds, according to rows of the table.

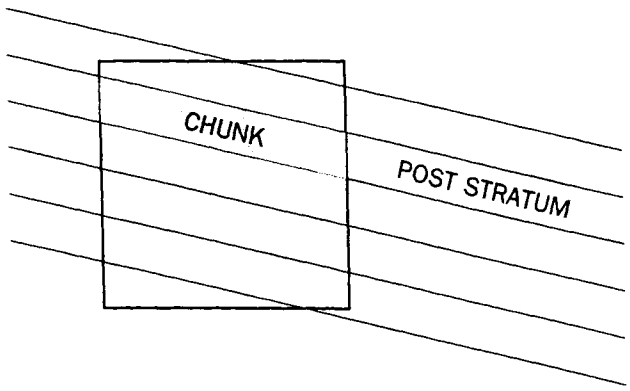
		Standard Deviation		
Proxy & Post Stratification	Level	Across states	Across P-S	Within P-S across states
II 448	.0208	.0069	.0134	.0201
II 64	.0208	.0069	.0131	.0128
II 16	.0208	.0069	.0133	.0089
LA 448	.0085	.0036	.0070	.0118
LA 64	.0085	.0036	.0069	.0074
LA 16	.0085	.0036	.0056	.0046

Note. The level does not depend on the post stratification, and neither does the SD across states. These two statistics do depend on the proxy.

In this section, we show that substantial heterogeneity remains in the data, despite the Bureau's elaborate post stratification; in fact, the post stratification seems on the whole to be counter-productive. We measure heterogeneity as in Freedman and Wachter [13], with "whole-person substitutions" and "late census adds" as "proxies" (surrogates) for undercount.[23] For example, 2.08% of the census count came from whole-person substitutions ("II" in the first line of the table, for obscure historical reasons). We compute these substitution rates, not only for the whole country, but for each state and DC: the standard deviation (SD) of the 51 rates is .69 of 1%. We also compute the rate for each post stratum: across the 448 post strata, the SD of the substitution rates is 1.34%: the post strata do show considerably more variation than the states.

On the other hand, we can think of each state as being divided into "chunks" by the post strata, as in the sketch below. (Alabama, for instance, is divided into 251

chunks by the post strata; post stratum #1 is divided into 7 chunks by states.) We compute the substitution rate for each chunk with a non-zero census count, then take the SD across chunks within post stratum, and finally the root-mean-square over post strata. We get 2.01%. If rates were constant across states within post strata, as the homogeneity assumption requires, this SD should be 0. Instead, it is larger than the SD across post strata, and almost as large as the overall imputation level.



We made similar calculations for two coarser post stratifications. (i) The Bureau considers its 448 post strata as coming from 64 PSGs, a PSG being a "post-stratum group." (Each PSG divides into 7 age-sex groups, giving back $64 \times 7 = 448$ post strata.) We use the 64 PSGs as post strata in the second line of Table 5. (ii) The Bureau groups PSGs into 16 EPS, or "evaluation post strata." We use these as post strata in the third line of Table 5. Remarkably, there is less rather than more variability within post-stratum group than within post stratum—and even less within evaluation post stratum; the air of paradox may be dispelled by Freedman and Wachter [13, p. 482].[24] Results for late census adds (LA) are similar, in lines 4–6 of the table. If the proxies are good, refining the post stratification is counter-productive: with more post strata, there is more heterogeneity rather than less.

13 Alternative Post Stratifications

The Bureau computed "direct DSEs" for the 16 evaluation post strata, by pooling the data in each: we constructed an adjustment factor, as the direct DSE divided by the census count.[25] We adjusted the United States using these 16 factors rather than the Bureau's 448. For states and congressional districts, there is hardly any difference: the scatter diagram in Figure 2 shows results for congressional districts. There are 435 dots, one for each congressional district. The horizontal axis shows the change in population count that would have resulted from adjustment with 448 post strata; the vertical, from adjustment with 16 post strata. There is little to choose between the two. (For some geographical areas with populations below 100,000, however, the two adjustments are

likely to have different consequences.)

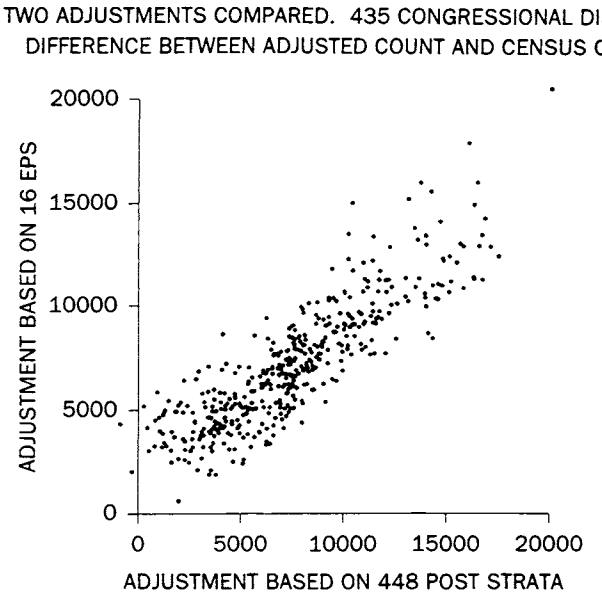


Figure 2: Changes to congressional district populations. The production adjustment, with 448 post strata, is plotted on the horizontal. An alternative, based only on the 16 evaluation post strata (EPS), is plotted on the vertical.

For example, take CD 1 in Alabama, with a 2000 census population of 646,181. Adjustment with 448 post strata would have increased this figure by 7630; with 16 post strata, the increase would have been 7486. The corresponding point is plotted at (7630, 7486). The correlation between the 435 pairs of changes is .87, as shown in the third line of Table 6.[26] For two out of the 435 districts, adjustment by 448 post strata would have reduced the population count: their points are plotted just outside the axes, at the lower left.

Within a state, districts are—by court order—almost exactly equal in size when redistricting is done shortly after census counts are released. Over the decade, of course, people move from one district to another. Variation in population sizes at the end of the decade is therefore of considerable policy interest. In California, for one example, 52 districts were drawn to have equal populations according to Census 1990. According to Census 2000, the range in their populations is 583,000 to 773,000.

Table 6: Comparing the production adjustment based on 448 post strata to one based on 16 evaluation post strata. Correlation coefficients for changes due to adjustment.

Changes in state population counts	.99
Changes in state population shares	.90
Changes in congressional district counts	.87
Changes in congressional district shares	.85

Table 6 and Figure 2 show that the Bureau's elaborate post stratification does not remove much heterogeneity. Whatever there was with 448 remains with 16, and that is a lot (Table 5). Experience from 1990 and 2000 teaches a sad lesson. Heterogeneity is not to be removed by the sort of post stratification that can be constructed by the Bureau. The impact of heterogeneity on errors in adjustment is discussed by Freedman and Wachter [13, pp. 479–81]: heterogeneity is likely to be much more of a problem than sampling error.

14 Loss Function Analysis

Proponents of adjustment often rely on a statistical technique called "loss function analysis." In effect, this technique attempts to make summary estimates of the error levels in the census and the adjustment, generally to the advantage of the latter. However, the apparent gains in accuracy—like the gains from adjustment—tend to be concentrated in a few geographical areas, and heavily influenced by the vagaries of chance. At a deeper level, loss function analysis turns out to depend more on wishful assumptions than on data.

For example, adjustment makes the homogeneity assumption: census errors occur at a uniform rate within post strata across wide stretches of geography. Loss function analysis assumes that and more: error rates in the census are uniform, and so are error rates in ACE. A second example: loss function analysis depends on models for correlation bias, and the Bureau's model assumes there is no correlation bias for women. The idea that only men are hard to reach—for the census and the post enumeration survey—is unlikely on its face. It is also at loggerheads with the data from 1990. See Wachter and Freedman [29]. For such reasons, we cannot believe that loss function analysis will clarify remaining issues about use of adjusted census data.

The discussion now becomes progressively—but unavoidably—more technical. (Readers can skip to Section 15 or 16, without losing the thread of the argument.) Loss function analysis tries to compare accuracy of census and adjusted figures for defined geographical areas. By way of example, take counts for states. The main ingredients for the comparison are the following.

(a) Census counts.

(b) Production adjustment counts.

(c) Variances in production adjustment counts.

(d) Biases in production adjustment counts.

However, the quantities in (d) must themselves be estimated, and estimates have variances. Thus, we amend (d) and add (e).

(d) Estimated biases in production adjustment counts.

(e) Variances in estimated biases.

In 1990, estimates for the variances in (c) were questionable. That issue may not arise for 2000, but estimates for the variances in (e) remain problematic. Here is why. As noted above, much of the evaluation data used by the Bureau in March 2001 comes from the 1990 Evaluation Followup, a sample survey done several months after the post enumeration survey was completed. This survey was based on 919 block clusters; the 1990 PES, on 5290; the 2000 ACE, on 11,303. On this basis, variances should be $11{,}303/919 = 12$ times bigger than the ACE variances. Instead, they are about 4 times smaller. The variances for estimated biases are, by such a reckoning, too small by a factor of $4 \times 12 = 48$. Other calculations give much larger factors, but 48 is surely enough to make the point.[27]

Where did the missing variance go? Processing error can be estimated from Evaluation Followup in fine-grain geographical detail. However, the sample is small, so variances for direct estimates would be huge. Instead, errors are aggregated to broad population groups (16 evaluation post strata in 2000) and then shared back down to constituent post strata, using proportionality assumptions. Finally, errors are spread across state or substate areas assuming constant error rates within post strata across geography, for correlation bias as well as processing error. Thus, variance in estimated errors is converted to bias by the sharing and spreading—but that particular bias is ignored in loss function analysis.

The statistical theory of loss function analysis. If we use survey data to estimate a parameter, loss can be defined as squared error. Risk is expected squared error, that is, averaged over hypothetical replications of the survey. Loss function analysis tries to make unbiased estimates of risk, as the variance of the estimator plus the square of its bias. Bias has to be estimated, and the variance of the bias estimator has to be accounted for. Unbiased estimators of bias, and unbiased estimators of their variances, are needed to make the calculation work.

For the intended application, consider two competing "estimators" of the population of California at census day in the year 2000: the census itself, and the adjustment based on ACE. The Bureau estimated a risk from the census: variance is nil, and bias is estimated primarily from ACE with some refinement from evaluation studies in 1990. We replicated the Bureau's calculation, and found the estimated risk to be 167 billion; units are "squared people." Likewise, there is an estimated risk for adjustment, which is 12.2 billion. Adjustment seems to make the smaller error. This process can be repeated for each state and DC. Summing the results gives a total estimated risk of 362 billion for the census, compared to 46 billion for adjustment. See line 1 of Table 7: a billion is 10^9, and $362 - 46 = 316$.

Table 7: Loss function analysis for counts. States. Weights inverse to census counts.

Wtd	Cov	Diff	SE	Scale
No	1	316	139	10^9
No	25	316	432	10^9
Yes	1	193	95	10^2
Yes	25	193	303	10^2

The Bureau preferred to divide the estimated risk for each state by its population, at least in its March report. We measure population by the census count: for California, this is 33.1 million.[28] The census risk for California is 167 billion/33.1 million $\doteq$ 5050 while the adjustment risk is 12.2 billion/33.1 million $\doteq$ 369. (Here, $\doteq$ means "approximately equal.") In total over all 51 states (and DC), we get 24,558 for the census risk and 5208 for the risk from adjustment.[29] The difference is 24,558 – 5208 = 19,350, which is rounded to 193×10^2 in line 3 of Table 7; "wtd" indicates division by population counts.

In Table 7, biases are estimated from the Bureau's "preferred" model.[30] If "Cov" is 25, the Bureau's covariance matrix for estimated biases is multiplied by 25, which brings the variances closer to what might be anticipated on the basis of sample size, as discussed above. Let "Cen" be estimated census loss and "Adj" be estimated adjustment loss. Then "Diff" in the table is Cen – Adj, which is the estimated gain in accuracy from adjusting. Diff is estimated from sample data, ACE and Evaluation Followup, and is therefore subject to sampling error. The "SE" column gives the standard error for Diff, and gauges the likely magnitude of sampling error in this estimate.

The last column indicates the units. Thus, in the first line of Table 7, the estimated gain in accuracy from adjustment is

$$(316 \pm 139) \times 10^9.$$

Diff is over twice its SE, and such a large value for Diff is hard to explain as the result of sampling error alone. (Diff is "statistically significant.") However, the calculation rides on Bureau estimates for the sampling variability in the biases—which are too low. Correcting these, as in the second line of Table 7, makes Diff noticeably smaller than its SE, and readily explained as the result of chance. Correcting the covariance matrix for the biases does not change Diff itself, but has a pronounced effect on its estimated SE.

Table 8: Loss function analysis for shares. States. Weights inverse to census shares.

Wtd	Cov	Diff	SE	Scale
No	1	306	297	10^{-9}
No	25	306	778	10^{-9}
Yes	1	57	63	10^{-7}
Yes	25	57	153	10^{-7}

Table 8 turns to state shares; weights are inversely proportional to census population shares. Diff is at the chance level. For congressional districts, the Bureau's loss function uses shares within state, but weights states by the square of the census count. This seems both cumbersome and unnatural—at least to us. We replicated the Bureau's analysis, but also examined numerical accuracy with the squared error loss function and no weights (Table 9).

Table 9 treats congressional districts as 435 areas across the country, with populations ranging from 500,000 to 1,000,000. As before, the estimated gain in accuracy from adjustment is significant if we use Bureau variances for the bias estimates, but insignificant when we correct for under-estimation. The District of Columbia does not come into Table 9, and state boundaries play no special role.

Table 9: Loss function analysis for counts. Congressional districts, unweighted.

Wtd	Cov	Diff	SE	Scale
No	1	137	65	10^8
No	25	137	202	10^8

Tables 7 and 9 show that the statistical significance of loss function analysis for counts is strongly dependent on the modeling—among other things, on the homogeneity assumption for biases. Table 8 shows that, for shares, Diff is at the chance level. Still, Diff is positive in all the summary tables. Perhaps that means adjustment is better than the census? We think not. The Bureau's March estimates for processing error and correlation bias were on the low side. Table 10 doubles the Bureau's allowance for processing error, post stratum by post stratum; it doubles the Bureau's allowance for correlation bias in states likely to have had unusually high levels of correlation bias in the 1990 adjustment (Wachter and Freedman [29] Table 5). This brings processing error to 4 million and correlation bias to 1.5 million; it allows for some geographical variation in rates of correlation bias, a possibility which is excluded by the Bureau's model. The corrected loss function analysis favors the census.

Table 10: Loss function analysis for counts. States. Partial correction for underestimated processing error and correlation bias. Some differentials in correlation bias. Weights inverse to census counts.

Wtd	Cov	Diff	SE	Scale
No	1	−185	99	10^8
No	25	−185	421	10^8
Yes	1	−214	65	10^2
Yes	25	−214	295	10^2

Table 11 gives detail for "Diff" in lines 3 and 4 of Table 8. Five states (CA, IA, MN, MO, TX) account for over half the estimated loss from the census. For the third line of Table 7, four states (CA, FL, GA, TX) account for over half the estimated loss from the census. Unweighted results (line 1 in Tables 7 and 8) are dominated by two states—CA and TX. In short, estimated gains from adjustment are concentrated in a few states, and subject to large uncertainties. Unbiased estimates of risk can be negative; that happens in Tables 7–10, and is explicit in Table 11.

Table 11: Estimated losses in accuracy from the census and from adjustment. State shares. Weights inverse to census shares. Parts per 10 million. Detail for "Diff" in lines 3 and 4 of Table 8. Alabama—Minnesota.

	Cen	Adj
Alabama	−0.6	0.6
Alaska	1.0	0.8
Arizona	−1.3	1.7
Arkansas	−0.4	1.0
California	12.6	2.5
Colorado	−1.1	1.1
Connecticut	−0.2	0.4
Delaware	0.7	0.4
DC	7.7	2.5
Florida	2.8	3.2
Georgia	7.3	2.1
Hawaii	1.1	2.4
Idaho	−0.6	1.0
Illinois	3.2	0.8
Indiana	4.9	0.7
Iowa	10.1	1.5
Kansas	5.2	0.7
Kentucky	−0.2	1.1
Louisiana	1.5	1.0
Maine	−1.8	2.9
Maryland	8.0	7.1
Massachusetts	3.7	0.8
Michigan	3.1	1.1
Minnesota	18.2	0.8

Table 11, Continued: Estimated losses in accuracy from the census and from adjustment. State shares. Weights inverse to census shares. Parts per 10 million. Detail for "Diff" in lines 3 and 4 of Table 8. Mississippi—Wyoming.

	Cen	Adj
Mississippi	−0.9	0.7
Missouri	12.2	0.5
Montana	−1.4	3.2
Nebraska	4.1	0.5
Nevada	0.3	0.8
New Hampshire	−0.3	0.8
New Jersey	−1.5	1.3
New Mexico	−1.0	3.1
New York	−0.9	2.2
North Carolina	0.2	0.8
North Dakota	3.2	0.7
Ohio	9.1	1.1
Oklahoma	−0.6	1.4
Oregon	−0.7	0.7
Pennsylvania	2.9	2.9
Rhode Island	0.3	0.5
South Carolina	1.0	1.9
South Dakota	3.4	0.9
Tennessee	−0.1	0.6
Texas	15.4	2.5
Utah	−1.0	1.1
Vermont	−1.0	1.4
Virginia	2.2	1.7
Washington	−0.9	3.1
West Virginia	1.0	2.9
Wisconsin	5.0	0.6
Wyoming	−0.5	0.9
Total	134.3	76.9

Given the levels of ACE processing error reported in U. S. Census Bureau [25], loss function analysis is an academic exercise. However, this sort of analysis seems to have played a salient role in Bureau deliberations over the 1990 adjustment, and was even a factor in the decision for Census 2000: see U. S. Census Bureau [24].[31] We think it is time to stop using loss function analysis. The assumptions are too fanciful.

15 Artificial Population Analysis

In essence, loss function analysis justifies the homogeneity assumption by making an even stronger assumption: not only are error rates in the census constant within post strata across geography, so are error rates in ACE. Proponents of adjustment may cite report B14: using proxy variables for overcounts and undercounts, that report creates artificial populations where truth is known. Bias in loss function analysis that

results from failures in the homogeneity assumption can then be measured—for the artificial populations. However, some proxies favor adjustment and some do not: Fay and Thompson [10, p. 82], Freedman and Wachter [13, pp. 484–5] .

Detailed results in B14 are rather mixed. Moreover, the B14 artificial populations are, well, artificial. Overcounts and undercounts are measures of difficulty in data collection. Intuition and data analysis suggest the following two criteria for proxies.

(i) The proxies for overcount and undercount should be positively correlated, but not perfectly correlated.

(ii) Proxies should be correlated with other indicators of poor data quality.

Bureau report B14 used four artificial populations, and Table 1 in that report lists the proxies. Populations #2 and #4 violate condition (i), using the same proxy for overcounts as for undercounts:

non-substituted persons in #2, non-mailbacks in #4.

With artificial population #3, there must be an inverse rather than a direct relationship between the two proxies, also violating condition (i):

persons with 2 or more allocations,

persons with non-allocated age and birth-date.

With artificial population #1, the proxies violate condition (ii), because they relate to goodness rather than badness of data:

non-substituted persons,

persons with non-allocated age and birth-date.

In consequence, the impact of heterogeneity on loss function analysis remains to be determined.[32]

16 Pointers to the Literature

Reviews and discussions of the 1980 and 1990 adjustments can be found in *Survey Methodology* 18 (1992) 1–74 and *Statistical Science* 9 (1994) 458–537. *Journal of the American Statistical Association* 88 (1993) 1044–1166 has a lot of useful descriptive material. Although tilted toward adjustment, the collection does include an insightful paper on heterogeneity—Hengartner and Speed [18]—and a comment on the imputation model by Wachter [27]. Other exchanges worth noting include *Jurimetrics* 34 (1993) 59–115 and *Society* 39 (2001) 3–53: these are easy to read, and informative. Pro-adjustment arguments are made by Anderson and Fienberg [1, 2], but see Stark [23] and Ylvisaker [30]. Prewitt [21] may be a better source, although he too must be taken with several grains of salt. Proponents of adjustment often cite Zaslavsky [31] to demonstrate the comparative advantages of adjustment; however, that paper makes all the mistakes discussed in Section 14 above, and others too. Cohen, White, and Rust [7] try to answer arguments on the 1990 adjustment, but miss many points.[33] Skerry [22] has an accessible summary of the arguments, leaning against adjustment. Darga [8, 9] is the sternest of critics. Freedman, Stark and Wachter [12]

have a probability model for census adjustment, which may help to clarify some of the issues.

17 Policy Implications

Census adjustment has become an expensive program for the Bureau, especially in terms of senior management time. The cost of ACE is driven in part by its complexity, and in part by the sample size. The large sample is needed because there are so many post strata, so the sample is spread very thin. However, given the results in Tables 5–6, the number of post strata would be very hard to justify. Moreover, the large sample size, while helpful on the sampling error side, must contribute to non-sampling error. Bigger samples are harder to manage, and non-sampling error is the critical issue (Sections 4–12).

The sample size for the post enumeration survey should therefore be scaled back dramatically. If 448 post strata are cut back to 16, as suggested by Sections 12–13, the size of the sample can be reduced by a factor of $448/16 = 28$ while maintaining the average number of households per post stratum. The sample can therefore be reduced from 300,000 households to $300{,}000/28 \doteq 10{,}000$, although that seems too optimistic. If the program is to be continued, the focus should be research and evaluation not adjustment, and a sample in the range 10,000–25,000 households should be adequate.[34]

If the adjustment program is scaled back, the savings could well be used elsewhere:

Demographic Analysis (DA)
American Community Survey (ACS)
Maintaining the Master Address File (MAF)
Community outreach between census years
Research into counting methods
Non-response followup.

Putting a few million dollars into demographic analysis now would make a big difference in 2010. Despite its flaws, DA allows reasonably accurate estimates for the sizes of major population groups, and these estimates could readily be improved—if resources were made available. DA is a more promising tool for census evaluation than ACE.

Decennial census long form data (income, education, occupation, country of origin, and so forth) are collected on a sample of about 1/6 of the respondents. The ACS will collect such data with a rolling sample survey, by interviewing 3% of the nation's households each year. The data will be available continuously, rather than every 10 years. From many perspectives, answers to ACS questionnaires are likely to be of better quality. Furthermore, the burden on the census will be markedly reduced. (The coverage of the ACS, however, is not likely to be as good as the census.)

The accuracy of a mail-out-mail-back census depends on having a list of addresses that is nearly complete, with few duplicates. Building such a list every 10 years is a

huge undertaking. Building it once and then maintaining it might be more productive. With ACS in place, maintaining the MAF seems like a promising activity.

Given the decades of effort spent in developing post enumeration surveys for census adjustment, the decision not to adjust must have been a wrenching one for the Bureau. We are confident they made the right decision. Statistical adjustments were considered in 1980, twice in 1990, and twice again in 2000. These adjustments could not improve the accuracy of the census. The adjustment technology does not work well enough to use. It is time to move on.

Endnotes

1. The Census Bureau provided detailed summary data on the census and the adjustment, by memorandum of understanding with the National Academy of Science and congressional oversight bodies. We were given access to these data; many of our results—like coverage comparisons—are computed from these data. The sum of P-sample non-movers and inmovers is about 1.6 million persons less than the number of E-sample persons (upweighted to national totals).

2. B2 p. 30 and App. 5. Also see R32, Attachment 2, Table 2. "B2" and "R32" are Bureau reports. For bibliographic details, see U. S. Census Bureau [24].

3. The procedure for estimating gross omissions is called "capture-recapture" in the statistical literature: capture is in the census, recapture is in the post enumeration survey. Erroneous enumerations are one major complication, and there are others. For a discussion of the 1990 procedures, see Hogan [19]. As yet, there is nothing comparable for 2000; report Q37 may be the best source.

4. Computed from data described in note 1.

5. Computed from data described in note 1.

6. See Wachter and Freedman [29], with cites to the literature; Breiman [4] is a primary source for alternative error estimates.

7. See B13 and B14 for data sources, and note 14 for methods. The positive component of processing error totals 2.14 million across all post strata, and the negative component .15 million: $2.14 - .15 \doteq 2$ million. Line (ii) of Table 4 combines processing error with correlation bias, giving a smaller net error.

8. B3 Table 8 shows 10.7 million housing units with two forms, and .5 million with three or more. Another 2.4 million forms were flagged as duplicates very late in the process (B1 p. 2). We are reckoning 2 persons per form, although this is rather rough.

9. On heterogeneity in 1990, see Freedman and Wachter [13], Wachter and Freedman [28]. Data for 2000 are presented here, in Table 5. On correlation bias in 1990, see Wachter and Freedman [29].

10. The U. S. Supreme Court had already precluded the use of adjustment for reapportionment, that is, allocating congressional seats among the states; previously, it had upheld Secretary Mosbacher's decision not to adjust Census 1990. See 517 U. S. 1 (1996), 525 U. S. 316 (1999), available on-line at `http://supct.law.cornell.edu/supct/`. For discussion, see Brown *et al.* [5]. The Bureau's recommendation is explained in U. S. Census Bureau [24].

Other litigation. Efforts by Los Angeles and the Bronx among others to compel adjustment have been rejected by the courts (City of Los Angeles *et al.* v. Evans *et al.*, Central District, California); appeals are pending in the Ninth Circuit. Utah has sued to preclude the use of imputations but their suit was denied by the Supreme Court (Utah *et al.* v. Evans *et al.*, `http://supct.law.cornell.edu/supct/`). Members of the House have also sued to compel release of block-level adjusted counts; this case is pending (Waxman et al. v. Evans *et al.*, Central District, California), along with a similar case in the Southern District of Texas (Cameron County *et al.* v. Evans *et al.*).

11. B4, Appendix Table 1, Cols. 1–4. We can replicate the census and ACE figures from the data described in note 1.

12. On missing data, see B1 p. 4 and B6 p. 29; report B7 is useful too. Results for movers were computed from data described in note 1.

13. B1 p. 39. Also see Belin and Rolph [3], with a response by Freedman and Wachter [13]. In brief, the "validation" is a coincidence of two rates. One rate is computed from all PES data that required imputation. The other is a benchmark computed from all cases re-interviewed and resolved in Evaluation Followup, a smaller survey done many months after the Post Enumeration Survey, designed to check the quality of the PES. Only stronger cases were sent to Evaluation Followup, and of these, only the strongest could be found and resolved. Thus, the benchmark rate is computed from only 25% of the data, and by no means a random 25% either.

14. The Bureau has provided two sets of "targets" for each of the post strata, computed (i) with correlation bias, and (ii) without correlation bias. Each set comprises 1,000 replicates. The average target in set (i) represents their idea of the adjustment factors, corrected for processing error. The difference between the mean target in set (i) and the corresponding adjustment factor represents the estimated net effect of processing error, post stratum by post stratum. Similarly, the difference between the means of the two sets represents the allowance for correlation bias, post stratum by post stratum.

These target adjustment factors were developed for 416 "collapsed" post strata, e.g., post strata 06-2, 06-4, and 06-6 are pooled due to small sample size; we take the target for a pooled post stratum and treat it as the (common) target for the components.

The targets summarize the Bureau view on errors in ACE, as of March 2001. Error estimates were largely derived from 1990 data. See B13 and B19. In turn, the 1990 data mainly derive from the Evaluation Followup (note 13).

The biases for states and congressional districts were computed from the targets; adjustments were computed from other data described in note 1. By way of example, take

CD 1 in Alabama, whose population according to Census 2000 was 646,181. Adjustment would have added 7630 to this figure, of which 3983 is due to processing error. (This CD is the first in our data file.)

15. If the adjustment factor for a post stratum is greater than 1.00, adjustment adds to the population; in total, such post strata add 4.26 million to the census count: apparently, the census is below the true population for such post strata, and has made a negative error. If the adjustment factor is less than 1.00, adjustment subtracts from the population; in total, such post strata would subtract 1.00 million from the census: apparently, the census is above the true population for such post strata, and has made a positive error.

16. See note 14 on the sources of the data, and the targets. We take the mean target in set (i), with correlation bias, for each post stratum; subtract this from the corresponding adjustment factor; and multiply by the census count for the post stratum. The sum of the positive numbers is 1.75 million; the sum of the negatives, .90 million.

17. The "negative error" in the census comes from omissions, estimated by ACE as 4.26 million; but 1.75 million of this figure reflects errors in ACE, rather than census errors. The positive error in the census can be computed from lines (i) and (ii) as $1.00 - .90 = .10$.

18. Briefing by Census Bureau staff to congressional oversight committees, 21 March 2001.

19. Gross errors in ACE were discussed in Bureau report B19, and were also provided to us in computer-readable format (note 1); disentangling the sign conventions in the computer file seemed more trouble than it was worth, so we report gross error only. Gross errors reported by Freedman and Wachter [15, p. 31] were derived from B19 p. 80, and do not net out processing error within post strata.

20. Anderson and Fienberg [1, 2]; Fienberg [11].

21. The Bureau's decision is explained in U. S. Census Bureau [25]; funding allocation is mentioned on p. 3. As Figure 1 suggests, however, the impact of adjustment on the distribution of funds would have been minor, at least in relative terms. Also see U. S. Commerce Department [26, Appendix 15]. See U. S. Census Bureau [25, p. 6] on DA; p. 10 reports that there were 3–4 million undetected duplications, which must be added to the previous figure of 2 million for processing error, giving the range 5–6 million; p. 11 discusses the 15 million cases that remain in doubt; and p. 12 mentions 10 million gross errors in mover status, with many other large gross errors mentioned elsewhere. The report notes on p. 13 that missing data create "considerable" uncertainty; the quantification at 500,000 is optimistic, as discussed on p. 14. The numbers may change when further analysis is done. Balancing error is no longer considered a serious problem by the Bureau.

22. Arguments for increasing DA estimates were made by Jeff Passel at the Urban Institute and Bureau demographers, although the two groups reach somewhat different conclusions. Our own estimates were presented at a conference on census adjustment

in Berkeley, on 24 September 2001. For details on correlation bias and data from 1990, see Wachter and Freedman [29]. The Bureau has so far not recognized the problem (B1 p. 45). While B12 p. 16 grants our premises, it denies the arithmetic. We return to this point when discussing loss function analysis.

23. Some persons are counted in the census with no personal information, in which case all their personal characteristics are imputed. "Late census adds" are persons who file a census form too late to be run through the ACE process. These may or may not be "data-defined," i.e., have enough characteristics for matching. Our late adds include the non-data-defined late adds, whereas our IIs exclude those records.

As a further complication (note 8), 2.4 million forms were found to be duplicates late in the process; of these, about 1.4 million were taken completely off the table, but 1.0 million were put into the ACE process as late adds. These forms correspond to 2.3 million people, and in 2000, the late adds were basically just these people. (Neither imputes or late adds are eligible for matching in ACE; such records are subtracted from the census, on the theory that the corresponding persons—if they really exist—will come back as gross omissions.) For details and an attempted explanation of the logic, see Q43.

For consistency, Table 5 only covers the "ACE target population," i.e., persons living in group quarters or institutions are excluded, as is remote rural Alaska.

24. Freedman and Wachter [13, p. 482] make the connection with analysis of variance, identifying the inequalities that must hold—and those that can be violated. As indicated there, results may depend on weights. Table 5 is unweighted: all states are treated as equals, likewise for post strata and "chunks." Undoubtedly, some of the effect in Table 5 is due to random variation in the smallest chunks—by the time you spread the population over 448 post strata, 50 states (and D. C.), some of the cells have to be tiny. However, the variance correction in equation (4) of Wachter and Freedman [28] does not affect the pattern in the table, although estimated heterogeneity is reduced. Similarly, we can restrict attention to chunks with a census population in excess of 100, for instance. Or, we can restrict attention, say, to PSGs #1–60 or even to #1–48, #53–60: this markedly reduces the variation due to small chunks without changing the pattern in Table 5. (The excluded PSGs have small populations, and cut across 50 states as well as D.C.) The table does not seem to be an artifact of small-sample variation.

Congressional districts. We were able to compute the analog of Table 5 for congressional districts. The SD across districts is about 1.5 times the SD across states. The SD within post strata across districts is about 3.5 times the SD within post strata across states, for the 448 post strata. Going to 64 or 16 post strata does not change the SD within post strata across districts. These results apply to both proxies.

25. See B19. We used the census count for the ACE target population, as defined at the end of note 23.

26. Some additional summary statistics may be of interest, for instance, for adjustments to congressional districts in Table 6 and Figure 2 (counts). The average district had a

census population of 646,000 in 2000; adjustment by 448 post strata would have added 7500 persons on average, compared to 7200 from the 16 evaluation post strata; the SDs are 3800 and 3000. From the perspective of "loss function analysis," the coarser post stratification is to be preferred (last paragraph of note 30 below).

27. A "block cluster" contains one block in densely populated areas, and many contiguous blocks in the hinterland. On average, there seem to be about two blocks per cluster. See, e.g., pp. 1048, 1088–89 in Hogan [19] for 1990 sample sizes and B11 p. 3 for 2000 data. There are published discussions of problems with Bureau estimates of variance for bias estimators: see, e.g., Freedman *et al.* [17, pp. 268–69], Freedman and Wachter [13, pp. 532–33]. These papers explain loss function analysis in some detail; also see Mulry and Spencer [20], Brown *et al.* [5]. For 2000, see B13 and B19.

The factor of 4 comes from comparing the traces of the covariance matrices for the adjustment factors and the targets (note 14). For a discussion of the process for generating the targets, see B13 and B19; Fay and Thompson [10, p. 77] may be clearer, or Mulry and Spencer [20, p. 1089], although these refer to the 1990 adjustment. On the 1990 variances, see Freedman *et al.* [16].

28. We are using the ACE target population as defined at the end of note 23. (This is due to a quirk in our computer code.) The full census population of California (including group quarters and institutions) is 33.9 million.

29. The ratio is $24{,}558/5{,}208 \doteq 4.715$, compared to 4.895 in B13, Table 1A, line 1. We also replicated the other results in line 1, to the same precision. We have not replicated the targets themselves; preliminary calculations suggest some incongruities (Don Ylvisaker, personal communication), but to resolve these, additional data would be needed.

30. The Bureau's "preferred model" for correlation bias corresponds to line 1 of Table 1A in B13; this version of correlation bias is built into the set of targets and hence the estimated biases (note 14) that underlie Tables 7–11.

Weights. See B13 p. 5, where the loss function for state counts is weighted inversely to population; for shares, weights are inverse to population share. We follow the Bureau, but weight by the census rather than the adjusted census or bias-corrected adjusted census, to avoid random weights. Due to the aforementioned quirk in our code (notes 28), we weight by census counts for the ACE target population in Tables 7 and 10, rather than the full census count. (Our code for shares uses the full census count.) For reasons that are at best obscure, the Bureau's loss function for congressional districts uses shares within-state, but weights the states proportional to the square of their census count. We followed suit in replication, but not in Table 9.

Terminology. The ACE target population (defined at the end of note 23) has nothing to do with the targets in note 14.

The model behind loss function analysis. The model, and the procedure for estimating the SE of Diff, are discussed in Freedman *et al.* [17]; also see Freedman and Wachter [14, pp. 368–70] or Brown *et al.* [5, pp. 372–75]. For state counts, say, let C be the

51×448 matrix whose (i,j)th entry is the census count for state i intersected with post stratum j. Let $\hat{\phi}$ be the 448×1 vector of adjustment factors, and $\hat{\psi}$ the 448×1 vector of estimated biases in $\hat{\phi}$. Write $\mathrm{E}(\hat{\phi}) = \phi$ and $\mathrm{E}(\hat{\psi}) = \psi$. The little model behind loss function analysis assumes that $\hat{\psi}$ is unbiased, so that ψ is the "true" bias in the adjustment factors, and true population counts for the states are given by $C(\phi - \psi)$; estimators are jointly normal, with $\hat{\phi}$ independent of $\hat{\psi}$; finally $\mathrm{cov}(\hat{\phi})$ and $\mathrm{cov}(\hat{\psi})$ are taken as given.

The adjusted state counts (ACE target population) are given by $C\hat{\phi}$; likewise, the biases in these counts are estimated by $C\hat{\psi}$. The covariances are

$$\mathrm{cov}(C\hat{\phi}) = C\,\mathrm{cov}(\hat{\phi})\,C' \quad \text{and} \quad \mathrm{cov}(C\hat{\psi}) = C\,\mathrm{cov}(\hat{\psi})\,C'. \tag{1}$$

Of course, when (1) is applied to data, the covariance matrices $\mathrm{cov}(\hat{\phi})$ and $\mathrm{cov}(\hat{\psi})$ would be replaced by sample-based estimates. The estimated census risk is

$$\mathrm{Cen} = \|C(\hat{\phi} - \hat{\psi} - 1)\|^2 - \mathrm{trace}\,\mathrm{cov}(C\hat{\phi}) - \mathrm{trace}\,\mathrm{cov}(C\hat{\psi}). \tag{2}$$

The estimated risk from adjustment is

$$\mathrm{Adj} = \|C\hat{\psi}\|^2 + \mathrm{trace}\,\mathrm{cov}(C\hat{\phi}) - \mathrm{trace}\,\mathrm{cov}(C\hat{\psi}). \tag{3}$$

Notice that $\mathrm{trace}\,\mathrm{cov}(C\hat{\psi})$ cancels when computing $\mathrm{Diff} = \mathrm{Cen} - \mathrm{Adj}$.

The census population, adjustment factors, and their covariances were supplied by the Bureau (note 1). Biases are computed from the targets (note 14), and their covariance is the empirical covariance of the 1000 sets of targets provided by the Bureau. In Tables 7–11, we are using the targets that correspond to the "preferred model" for correlation bias (see the beginning of this note; the "preferred model" is for correlation bias, not loss function analysis more generally).

The SE of Diff. Index the the 50 states and DC by $k = 1, \ldots, 51$. Let μ_k be the error in the census population count for area k. Let X_k be the production dual system estimate for μ_k; thus, $X = C(\hat{\phi} - 1)$. The bias in X_k is denoted β_k; this is estimated as $\hat{\beta} = C\hat{\psi}$. Let $G = \mathrm{cov}(\mathrm{X})$ and $H = \mathrm{cov}(\hat{\beta})$, which are assumed known. The estimated risk from the census for area k is $(X_k - \hat{\beta}_k)^2 - G_{kk} - H_{kk}$, while the estimated risk from adjustment is $\hat{\beta}_k^2 + G_{kk} - H_{kk}$. The estimated risk difference is

$$\hat{R}_k = (X_k - \hat{\beta}_k)^2 - \hat{\beta}_k^2 - 2G_{kk}. \tag{4}$$

By a tedious but routine calculation,

$$\mathrm{cov}(\hat{R}_i, \hat{R}_j) = 4\mu_i\mu_j G_{ij} + 2G_{ij}^2 + 4E(X_iX_j)H_{ij}. \tag{5}$$

The displayed covariance can be estimated from sample data as

$$\hat{K}_{ij} = 4(X_i - \hat{\beta}_i)(X_j - \hat{\beta}_j)\hat{G}_{ij} - 2G_{ij}^2 - 4G_{ij}H_{ij} + 4X_iX_jH_{ij}, \tag{6}$$

and var(Diff) can be estimated as $\sum_{ij} \hat{K}_{ij}$. When (4) and (6) are applied to data, the covariance matrices G and H would be replaced by sample-based estimates, whose variability is ignored in (5).

Issues. Of course, G and H are derived from $\text{cov}(\hat{\phi})$ and $\text{cov}(\hat{\psi})$ respectively—equation (1). The calculations take these matrices as known, or at least estimated in a reasonable way. In fact, our calculations suggest that trace $\widehat{\text{cov}}(\hat{\psi})$ is too small by a factor of 50 or more, $\widehat{\text{cov}}(\hat{\psi})$ being the Bureau's estimate for $\text{cov}(\hat{\psi})$. But scaling $\widehat{\text{cov}}(\hat{\psi})$ by a constant factor may not be a reliable correction. A more sensible thing to do is to make direct estimates of bias, and compute $\text{cov}(\hat{\psi})$ by jackknifing. (This would require much more data than we have.) The calculations also assume that estimates for the biases in the DSE are unbiased, which is rather questionable.

At some level, the Bureau must have been aware of the problems with $\text{cov}(\hat{\psi})$ when it considered adjusting the post-censals in 1992. If it believed in its own estimates for $\text{cov}(\hat{\psi})$, it would have adjusted not by $\hat{\phi}$ but by $\hat{\phi} - \hat{\psi}$. See Fay and Thompson [10, p. 74].

Footnote to a footnote. Loss function analysis "shows" that adjustment by the 16 EPS (Section 13) is comparable to or even better than the production adjustment (448 post strata). We adjust starting from the direct DSEs for the EPS (defined in Section 12), and evaluate using the Bureau's preferred targets. An exact comparison waits on the 16×16 covariance matrix for the direct DSEs, which we do not have; but a factor of 1.5 on estimated MSEs for state counts, unweighted, is plausible—within the peculiar conventions of loss function analysis.

31. On 1990, see Mulry and Spencer [20] or Fay and Thompson [10]. The Undercount Steering Committee cited loss function analysis throughout its report on the 1990 adjustment (U. S. Commerce Department [26] Appendix 4).

32. Also see U. S. Census Bureau [25, p. 17].

33. Some examples give the flavor.

Processing error. Cohen, White, and Rust [7, Chap. 4] defend the adjustment of 1990 without mentioning Breiman's [4] estimates of processing error, although they do take up one of his minor points (pp. 73–74).

Heterogeneity. According to Cohen *et al.* (p. 78), Bureau staff were in the early 1990s "well aware" of problems with the synthetic assumption. This may be so, after the adjustment decisions were made: Fay and Thompson [10]. While those decisions were being made, however, the facts were less evident. For instance, the Bureau's Undercount Steering Committee was "convinced from the data, that, in general, block parts are homogeneous within post-strata"; they saw "no evidence to indicate there are any serious flaws" in the homogeneity assumption at the state level. U. S. Commerce Department [26, Appendix 4 p. 7].

The imputation model. Cohen *et al.* (p. 72) defend the putative validation in 1990 (note 13) as "strong evidence." They ignore data showing the flimsiness of this evidence (Freedman and Wachter [13] pp. 535–56).

Correlation bias. Cohen *et al.* (p. 82) argue that relatively few people are added to the counts by statistical modeling. The point is that *more* people need to be added—but not in the places expected by the modelers. On the same page, Cohen et al. respond to a minor point in Darga [8], relating to correlation bias—but fail to address the major finding: adjustment would have had a perverse effect on many sex ratios.

Smoothing and loss function analysis. Variances for adjustment factors in 1990 are discussed in Freedman et al. [16]. The variances were obtained from a "smoothing model." Cohen *et al.* (pp. 58–62) defend the model, although they do not state the crucial assumptions, nor do they respond to our points. Similarly, Cohen *et al.* (pp. 68–85) defend loss function analysis, without responding to the points raised in Section 14—and in our previous publications (Freedman *et al.* [17]; Freedman and Wachter [13]). Also see Brown *et al.* [5, pp. 364–65, pp. 371–75].

Citro, Cork, and Norwood [6, p. 33] find that the Bureau's decision not to adjust was "justifiable," but the "fact that the Bureau did not recommend adjusting the census counts to be provided for redistricting does not carry any implications for the usefulness of statistical adjustment methods based on dual-system estimation."

34. The 16 evaluation post strata could probably be reconfigured with advantage. Among other things, one of the major variables used to define the post strata—the mail back rate—turned out to have paradoxical features: many post strata with high mail back rates had high rather than low measured undercounts.

Sampling error in the PES is relatively easy to quantify, and is not the major problem, at least at the national level. For example, the standard error for the national undercount estimate of 3.3 million is around 400,000. See B19 Table 20. Reducing the sample size by a factor of 30 would increase the SE to roughly $\sqrt{30} \times .4 \doteq 2$ million. (Of course, our calculations only give rough guidelines.) On this basis, sampling error would still be dominated by uncertainties due to non-sampling error (see Tables 1–4 and Section 11). The extent to which non-sampling error could be reduced with a smaller PES remains to be seen.

In 2000, ACE matched outmovers rather than inmovers. This treatment of movers came about by historical accident (Brown et al. [5] p. 367). The Bureau had planned to do sample-based non-response followup (SNRFU). With SNRFU, inmovers cannot be matched back to the census blocks they came from: the records may not be in the census because the people were not selected into the SNRFU sample. The Supreme Court over-ruled the Bureau on SNRFU, as part of the decision cited in note 10, but outmovers stayed in the plan, probably due to inertia. If there is a PES in 2010, let it match inmovers not outmovers—like the PES of 1990.

Authors' footnote. We thank Liza Levina, Chad Schafer and Don Ylvisaker for many useful comments. Both of us testified for the United States against adjustment, in the Census cases of 1980 and 1990. We have also testified in Congressional hearings, and consulted for the Department of Commerce on the adjustment decision.

Finally, a word about Terry Speed. Terry is a wonderful friend and colleague, who

for years has helped us by answering all kinds of questions, not least about census adjustment – although that is only one of his many interests.

David A. Freedman, Department of Statistics, University of California, Berkeley,
freedman@stat.berkeley.edu

Kenneth W. Wachter, Department of Demography, University of California, Berkeley,
wachter@demog.berkeley.edu

References

[1] M. Anderson and S. E. Fienberg. *Who Counts? The Politics of Census-Taking in Contemporary America.* Russell Sage Foundation, New York, 1999.

[2] M. Anderson and S. E. Fienberg. Census 2000 and the politics of census taking. *Society*, 39:17–25, 2001.

[3] T. R. Belin and J. E. Rolph. Can we reach consensus on census adjustment? *Statistical Science*, 9:458–537, 1994 (with discussion).

[4] L. Breiman. The 1991 census adjustment: Undercount or bad data? *Statistical Science*, 9:458–537, 1994 (with discussion).

[5] L. D. Brown, M. L. Eaton, D. A. Freedman, S. P. Klein, R. A. Olshen, K. W. Wachter, M. T. Wells, and D. Ylvisaker. Statistical controversies in Census 2000. *Jurimetrics*, 9:347–375, 1999.

[6] C. F. Citro, D. L. Cork, and J. L. Norwood, editors. *The 2000 Census: Interim Assessment.* National Academy Press, Washington, D. C., 2001.

[7] M. L. Cohen, A. A. White, and K. F. Rust, editors. *Measuring a Changing Nation: Modern Methods for the 2000 Census.* National Academy Press, Washington, D. C., 1999.

[8] K. Darga. *Sampling and the Census.* The AEI Press, Washington, D. C., 1999.

[9] K. Darga. *Fixing the Census until It Breaks.* Michigan Information Center, Lansing, 2000.

[10] R. E. Fay and J. H. Thompson. The 1990 post enumeration survey: Statistical lessons, in hindsight. In *Proceedings, Bureau of the Census Annual Research Conference*, Washington, D. C., 1993. Bureau of the Census.

[11] S. E. Fienberg. The New York City census adjustment trial: Witness for the plaintiffs. *Jurimetrics*, 34:65–83, 1993.

[12] D. A. Freedman, P. B. Stark, and K. W. Wachter. A probability model for census adjustment. *Mathematical Population Studies*, 9:165–180, 2001.

[13] D. A. Freedman and K. W. Wachter. Heterogeneity and census adjustment for the intercensal base. *Statistical Science*, 9:458–537, 1994 (with discussion).

[14] D. A. Freedman and K. W. Wachter. Planning for the census in the year 2000. *Evaluation Review*, 20:355–377, 1996.

[15] D. A. Freedman and K. W. Wachter. Census adjustment: Statistical promise or illusion? *Society*, 39:26–33, 2001. Sections 1–9 are adapted from this source.

[16] D. A. Freedman, K. W. Wachter, D. Coster, R. Cutler, and S. Klein. Adjusting the census of 1990: The smoothing model. *Evaluation Review*, 17:371–443, 1993.

[17] D. A. Freedman, K. W. Wachter, R. Cutler, and S. Klein. Adjusting the census of 1990: Loss functions. *Evaluation Review*, 18:243–280, 1994.

[18] N. Hengartner and T. P. Speed. Assessing between-block heterogeneity within the poststrata of the 1990 post-enumeration survey. *Journal of the American Statistical Association*, 88:1119–1125, 1993.

[19] H. Hogan. The 1990 post-enumeration survey: Operations and results. *Journal of the American Statistical Association*, 88:1047–1060, 1993.

[20] M. H. Mulry and B. D. Spencer. Accuracy of 1990 census and undercount adjustments. *Journal of the American Statistical Association*, 88:1080–1091, 1993.

[21] K. Prewitt. Accuracy and coverage evaluation: Statement on the feasibility of using statistical methods to improve the accuracy of Census 2000. *Federal Register*, 65:38373–38398, 2000.

[22] P. Skerry. *Counting on the Census? Race, Group Identity, and the Evasion of Politics*. Brookings, Washington, D. C., 2000.

[23] P. B. Stark. Review of *Who Counts? Journal of Economic Literature*, 39:592–595, 2001.

[24] U. S. Census Bureau. *Report of the Executive Steering Committee for Accuracy and Coverage Evaluation Policy.* With supporting documentation, Reports B1–B24. Washington, D. C., 2001. `http://www.census.gov/dmd/www/EscapRep.html`.

[25] U. S. Census Bureau. *Report of the Executive Steering Committee for Accuracy and Coverage Evaluation Policy on Adjustment for Non-Redistricting Uses.* With supporting documentation, Reports 1–24. Washington, D. C., 2001. `http://www.census.gov/dmd/www/EscapRep2.html`.

[26] U. S. Commerce Department. Office of the Secretary. *Decision on Whether or Not a Statistical Adjustment of the 1990 Decennial Census of Population Should Be Made for Coverage Deficiencies Resulting in an Overcount or Undercount of the Population, Explanation.* Three volumes. Reprinted in part in *Federal Register* 56: 33582–33642 (July 22). Washington, D. C., 1991.

[27] K. W. Wachter. Ignoring nonignorable effects. *Journal of the American Statistical Association*, 88:1161–1163, 1993.

[28] K. W. Wachter and D. A. Freedman. Measuring local heterogeneity with 1990 U. S. census data. *Demographic Research,* an on-line journal of the Max Planck Institute. Volume 3, Article 10, 2000. `http://www.demographic-research.org/Volumes/Vol3/10/` .

[29] K. W. Wachter and D. A. Freedman. The fifth cell. *Evaluation Review*, 24:191–211, 2000.

[30] D. Ylvisaker. Review of *Who Counts? Journal of the American Statistical Association*, 96:340–341, 2001.

[31] A. M. Zaslavsky. Combining census, dual system, and evaluation study data to estimate population shares. *Journal of the American Statistical Association*, 88:1092–1105, 1993.

A Brief Introduction to Genetics

Darlene R. Goldstein

Abstract

This very brief introduction to genetics is included to provide greater accessibility to the papers in this volume. More extensive details are available in genetics textbooks and the literature.

Keywords: DNA sequencing; genetic map; genome; microarray; molecular biology; physical map

1 Introduction

What follows is a very brief introduction to genetics concepts to provide greater accessibility to the papers in this volume. More extensive details are available in genetics and molecular biology textbooks, *e.g.* [2, 3, 4, 6, 8].

2 Genomes

The *genome* of an organism consists of the biological information content of a cell. This information is necessary for all cellular processes required by the organism. With the exception of some viruses, genomes are comprised of *deoxyribonucleic acid*, or *DNA*. DNA is a double-stranded, linear polymer consisting of a sugar-phosphate backbone attached to subunits called *nucleotides*. There are four nucleotides: the *purines*, *adenine (A)* and *thymine (T)*, and the *pyrimidines*, *cytosine (C)* and *guanine (G)*. Although DNA can form other tertiary structures, the best known is that of the double helix. The two strands of the double helix are held together by weak hydrogen bonds between complementary bases on the strands. Base pairing occurs as follows: A pairs with its complementary base T, and G pairs with C. The sequence complementarity provides a mechanism for DNA replication: each strand may serve as a template for sythesis of a new DNA molecule. *Ribonucleic acid*, or *RNA*, is similar to DNA but (i) contains the sugar ribose rather than deoxyribose, (ii) uses the base *uracil* (U) instead of thymine (T), and (iii) is usually single-stranded rather than double-stranded.

Genomic DNA is distributed along *chromosomes* in the cell nucleus. A *gene* is a segment of DNA that codes for a *protein*. Proteins perform a large number of diverse functions, serving as enzymes or antibodies, providing storage or transportation for other molecules, and providing structure (*e.g. collagen*). Proteins are made up of

subunits called *amino acids*; there are 20 amino acids. The set of rules associating the DNA sequence of a gene with the amino acid sequence of a protein is called the *genetic code*.

Genes occur at particular sites, or *loci* along a chromosome. A gene may exist in multiple versions, called *alleles*. The two alleles at a genetic locus comprise the *genotype* at that locus. If both alleles are the same, the genotype is *homozygous*; otherwise, the genotype is *heterozygous*. A locus may refer more generally to genetic units other than genes; for example, to sequences of DNA smaller than genes. Genetic entities such as these following normal hereditary laws are referred to as *markers*. Loci represented by more than one allelic variant in a population are said to be *polymorphic*. Examples of types of polymorphic marker systems include *restriction fragment length polymorphisms*, or *RFLP*s, and *single nucleotide polymorphisms*, or *SNP*s.

Gene expression is the process whereby the genetic information in a gene is made available to the cell. When a gene is expressed, it is said to be "turned on." Gene expression occurs in two major steps: *transcription* of the gene DNA sequence into *messenger RNA* (*mRNA*), followed by *translation* of the *mRNA* into protein. A number of intermediate steps also occur during expression, the details are omitted here.

Gene expression depends on not only allele status (genotype), but also chromosomal structure, DNA modifications, and gene-gene interactions (*epistasis*). An example of these other effects is *imprinting*, the phenomenon that genes are differently expressed depending on whether they came from the mother or father.

3 Molecular Laboratory Techniques

Advances in genetic knowledge have been made possible by innovative techniques in the molecular biology laboratory. Important techniques involve manipulations of DNA: hybridization, copying, cutting or binding, labeling and visualization.

Hybridization refers to the annealing of complementary strands of DNA. The two strands of DNA can be *denatured* (separated) by heating; upon cooling, the strands bind, restoring the original molecule.

This hybridization property of DNA can be exploited to *amplify* (copy) sequences of DNA. The process for amplifying DNA is the *polymerase chain reaction*, or *PCR*. PCR is used to amplify specific DNA sequences when the ends of the sequence are known. In PCR, the source DNA is denatured into single strands, short sequences complementary to one end of each strand are added in great excess, then the temperature is lowered so that the short sequences hybridize with their complementary sequences. The genomic DNA remains denatured, because the complementary strands are at too low a concentration to encounter each other during the period of incubation. The hybridized ends then serve as *primers* for DNA synthesis, which begins upon addition of a supply of nucleotides and a temperature resistant polymerase, an enzyme for synthesizing DNA. When the synthesis cycle is complete, there are approximately twice as many DNA molecules as there were at the start. Repeated cycles (25 – 30) of denaturing and

synthesis quickly provide many copies of the original DNA.

When a bacterium is invaded by a DNA-containing organism (*e.g.* a virus), it can defend itself with *restriction enzymes*, also called *restriction endonucleases*. Restriction enzymes recognize a specific short sequence of DNA and cut both strands at that sequence. They are used in the laboratory as "molecular scissors" for cutting large DNA molecules into smaller fragments. Restriction fragments may also be joined with the enzyme *DNA ligase*. This ability to join fragments is an important step in the creation of artificial *cloning vectors*, DNA molecules able to replicate inside a host. *Bacterial artificial chromosomes* (*BAC*s) and *yeast artificial chromosomes* (*YAC*s) are high capacity cloning vectors capable of cloning large fragments of DNA. These are used in large scale DNA sequencing projects.

Southern [9] showed that is possible to detect a specific DNA fragment within a mixed pool of fragments. Crucial to this process is DNA labeling, which enables the location and visualization of a particular DNA molecule. DNA may be labeled radioactively, then visualized by X-ray (autoradiography). Labeling for many procedures is also done with nonradioactive alternatives, most commonly with fluorescent markers.

4 Linkage and Genetic Maps

Most cells of *diploid* individuals, that is individuals whose cell nuclei contain two of each chromosome, contain a homologous pair of each *autosome*, or non-sex chromosome, and two sex chromosomes. However, gametic cells (sperm or egg) are an exception to this general rule, as they contain only one chromosome of each autosomal pair and one sex chromosome (*haploid*). Gametes are produced via a reduction divison process called *meiosis*. During meiosis, a diploid gametic precursor cell replicates DNA once and divides twice, producing four gametes.

It is also during meiosis that *crossing over* occurs. For each chromosome, after DNA replication, the two sets of chromosomal pairs (the four *chromatids*) become aligned, at which time pairs of nonidentical homologous chromosomes form regions of contact (*chiasmata*). Because physical exchange of chromosomal DNA occurs in these regions, the gametic chromosomes of an individual are generally not exact copies of the originals, but rather are combinations of the original pair.

It is one of these new combinations which is then passed on to offspring. The combination of alleles (at different loci) an offspring receives from one parent is called a *haplotype*. A *recombination* between two loci has occurred when the exchange of DNA is such that the resulting haplotype passed to an individual contains alleles at the two loci contributed by different grandparents. It is on the basis of haplotypes passed from parent to offspring that recombinations can be recognized. However, recombinations can only be distinguished from nonrecombinations when at least one parent is heterozygous at each locus.

Pairs of (gene or marker) loci on different chromosomes, or so far apart on the same chromosome that there is the same chance of recombination as nonrecombination, are

said to be *unlinked*. Two loci are *linked* when they are *not* passed on independently. When loci are in *linkage equilibrium*, the haplotype frequency is the product of the individual allele frequencies in the population; when this rule does not hold, the loci are in *linkage disequilibrium*.

The probability of recombination, or *recombination fraction*, measures the extent of linkage between loci, thereby providing a means for creating a *genetic map*. A measure of genetic distance is given by the expected number of crossovers on a single strand between two loci; the unit for this distance is the *Morgan* (*M*); distances are more commonly specified as centimorgans (*cM*; 100 *cM* = 1 *M*).

5 Physical Maps and Genome Sequencing

Genetic maps show the position of genes and other types of genetic loci in terms of genetic distance. These maps are constructed using techniques such as cross-breeding experiments and analysis of *pedigrees* (families). Prior to large scale, whole genome level sequencing, a genetic map should be supplemented by a *physical map*, constructed by direct examination of DNA molecules.

Techniques for physical mapping include: *restriction mapping*, which locates relative positions of cut sites for restriction enzymes on a DNA molecule; *fluorescent in situ hybridization* (FISH), whereby marker locations are mapped by hybridizing the marker to intact chromosomes; and *sequence tagged site* (STS) mapping, where positions of short sequences are mapped by PCR and hybridization analysis of genome fragments. Since STS is quick and not too technically demanding, it has been used for creating detailed maps of large genomes.

A single experiment is capable of directly sequencing DNA molecules with lengths of up to around 750 bp (base pairs, or nucleotides). Therefore, the sequence of an entire chromosome, which has length measured in mega-base pairs (Mb), must be constructed from smaller sequences. *Shotgun sequencing* is the standard approach used for smaller genomes. With this method, long DNA molecules are broken into fragments of sizes that can be sequenced directly. The fragments are individually sequenced, and the entire original sequence is reconstructed using computational algorithms to search for overlaps between contiguous DNA sequences (*contigs*). This approach does have some problems, though. When a genetic map is available, sequencing can proceed using variations of shotgun method: the *clone-contig* approach [7] or *directed shotgun* [10].

6 Microarray Technologies

Measuring the amounts of *mRNA* can provide information on which genes are being expressed, or used by, a cell. Microarrays provide a means to measure gene expression. Common areas currently under study with microarray experiments include: differential gene expression, that is, which genes are expressed differently between two

(or more) sample types; similar gene expression patterns (profiles) across treatments; tumor sub-class identification using gene expression profiles; classification of malignancies into known classes; and identification of genes associated with clinical outcomes, such as response to treatment or survival time. There are several microarray technologies in current use, but the two most widely used are *cDNA (complementary DNA) microarrays* and *high-density (short) oligonucleotide gene chips* produced by the company Affymetrix.

cDNA microarrays consist of thousands of individual cDNA *probe* sequences printed in a high-density array on a glass microscope slide. The relative abundance of these spotted DNA sequences in two DNA or RNA samples may be assessed by monitoring the differential hybridization of the two samples to the sequences on the array. For mRNA samples, the two samples (*targets*) are reverse-transcribed into cDNA, labeled using different fluorophores ("dyes"), usually Cyanine 5 (Cy5), which fluoresces at red wavelengths, and Cyanine 3 (Cy3), which fluoresces at green wavelengths. The labeled samples are then mixed in equal proportions and hybridized with the arrayed DNA sequences. After this competitive hybridization, the slides are scanned and fluorescence measurements are made separately for each dye at each spot on the array. The ratio of red to green fluorescence intensity for each spot is indicative of the relative abundance of the corresponding DNA probe in the two nucleic acid target samples. See [1] for a more detailed introduction to the biology and technology of cDNA microarrays and oligonucleotide chips.

Affymetrix gene chip arrays use a photolithography approach to synthesize probes directly onto a silicon chip. In addition to a number of short sequences used to probe each gene, the *perfect match* (*PM*) probes, there is an equal number of negative controls, the *mismatch* (*MM*) probes. A single labeled sample is hybridized to the array, so that absolute rather than relative measures of gene expression are obtained. Further details are available in [1, 5].

Darlene R. Goldstein, Bioinformatics Core Facility, Institut Suisse de Recherche Expérimentale sur le Cancer, 1066 Epalinges, Switzerland; and Institut de mathematiques, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland,
darlene.goldstein@isrec.unil.ch

References

[1] *The Chipping Forecast*, volume 21, January 1999. Supplement to Nature Genetics.

[2] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J. D. Watson. *Molecular Biology of the Cell.* Garland Science, New York, 4th edition, 2002.

[3] T. A. Brown. *Genomes.* BIOS Scientific, Oxford, 1999.

[4] A. J. F. Griffiths, J. H. Miller, D. T. Suzuki, R. C. Lewontin, and W. M. Gelbart. *An Introduction to Genetic Analysis*. W. H. Freeman and Company, New York, 6th edition, 1996.

[5] D. J. Lockhart, H. L. Dong, M. C. Byrne, M. T. Follet tie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, and H. Horton. Expression monitoring by hybridization to high-density oligonu cleotide arrays. *Nature Biotechnology*, 14:1675–1680, 1996.

[6] H. Lodish, A. Berk, L. Zipursky, P. Matsudaira, D. Baltimore, and J. Darnell. *Molecular Cell Biology*. W. H. Freeman, New York, 4th edition, 2000.

[7] S. G. Oliver, Q. J. van der Aart, M. L. Agostini-Carbone, M. Aigle, L. Alberghina, D. Alexandraki, G. Antoine, R. Anwar, J. P. Ballesta, P. Benit, and (128 others). The complete DNA sequence of yeast chromosome III. *Nature*, 357:38–46, 1992.

[8] J. Ott. *Analysis of Human Genetic Linkage*. Johns Hopkins University Press, Baltimore, 3rd edition, 1999.

[9] E. M. Southern. Detection of specific sequences among DNA fragments separated by gel electrophoresis. *Journal of Molecular Biology*, 98:503–517, 1975.

[10] J. C. Venter, M. D. Adams, G. G. Sutton, A. R. Kerlavage, H. O. Smith, and M. Hunkapiller. Shotgun sequencing of the human genome. *Science*, 280:1540–1542, 1998.

Common Long Human Inversion Polymorphism on Chromosome 8*p*

Karl W. Broman, Naomichi Matsumoto, Sabrina Giglio, Christa Lese Martin, Jessica A. Roseberry, Orsetta Zuffardi, David H. Ledbetter and James L. Weber

Abstract

In an analysis of human crossover interference, we identified apparent triple recombination events, in a short region on chromosome 8*p*, on the maternally-derived chromosomes in four individuals (two from each of two families). While this may have indicated an error in marker order, the inverted order was inconsistent with recombination events in other individuals. We were thus led to the hypothesis of an inversion polymorphism in the region, which was subsequently confirmed by fluorescent *in situ* hybridization (FISH). The inversion spans approximately 12 cM on the female genetic map and 2.5 – 5.3 Mb on the physical map. The allele frequency of the inverted order (D8S1130 telomeric; D8S351 centromeric) in 50 individuals of European ancestry was 21%. This is only the second known common, long inversion polymorphism in the human genome.

Keywords: CEPH; FISH; inversion; polymorphism

1 Introduction

Inversions in gene order along chromosomes have frequently been observed by comparing related species [14, 24, 25], including great apes [16, 21, 22, 29]. Human inversion mutations occur at a low, but detectable frequency. Paracentric (not involving the centromere) inversions that are large enough to be detectable by standard cytogenetic analysis occur at a frequency of 1 – 5 per 10,000 individuals [23]. The frequency of human submicroscopic inversions is unknown, although inversions have been identified as the cause of specific heritable disorders (see, for example, [1, 8, 15, 19, 20]). Chromosomal inversions are of particular clinical interest because recombination within the inverted region in heterozygotes can lead to segmental aneusomies and concomitant abnormalities.

The only well characterized common human inversion polymorphism is the 48 kb inversion of the Emery-Dreifuss muscular dystrophy and filamin genes on the X chromosome [26]. This inversion is present in populations of European descent at a frequency of about 18%. Page and colleagues also recently made a preliminary report of a potentially common 3 Mb inversion polymorphism on chromosome Y*p* flanked

by inverted 300 kb repeats [27]. Here we describe a common, paracentric inversion polymorphism spanning > 2.5 Mb in chromosome band $8p23.1 - 8p22$.

2 Materials and Methods

We considered high-density genotype data on eight of the CEPH reference families [6]. These families, which were recruited in order to form the first human genetic maps, are largely three-generation families, with 10–15 siblings each. They have been genotyped at > 8,000 short tandem repeat polymorphisms (STRPs, also known as microsatellites). The genotype data are publicly available (see the Marshfield web site, `http://research.marshfieldclinic.org/genetics`).

Initial marker order was taken from [3]. Haplotypes were constructed with use of the *chrompic* option of the CRI-MAP program [12]. The physical length of the inverted region was estimated based on the December 22, 2001, version of the University of California, Santa Cruz, draft human sequence (see `http://genome.ucsc.edu`).

Fluorescent *in situ* hybridization (FISH) was carried out as previously described [5]. A minimum of five spreads were examined for each individual. BAC clones were obtained from Genome Systems (St. Louis, Missouri, USA).

3 Results and Discussion

In an examination of the sites of meiotic recombination in eight of the CEPH reference families, as part of an analysis of human crossover interference [4], we observed that the maternally inherited chromosomes in two offspring from each of CEPH families 1362 and 1413 show similar and highly unlikely crossover patterns (Figure 1). Even in the absence of crossover interference, the probability of four triple crossovers within 12 cM is vanishingly small. All four of these chromosomes revert to single crossovers when the region between and including D8S351 and D8S1130 is inverted. However, the inverted order of markers is inconsistent with recombination events in other CEPH families (see Figure 1).

Fluorescent *in situ* hybridization (FISH) was used to confirm and extend the initial evidence for inversion. BAC clones encompassing D8S351 and D8S1130 near the ends of the inverted segment (see Figure 1) were used as probes. We arbitrarily defined the normal allele as having the marker order in Figure 1, and the inverted allele as having the markers between D8S351 and D8S1130 inverted. Shown in Figure 2 are representative metaphase results from two individuals with each of the three possible genotypes, including the mother (1362-02) (panel c) of CEPH family members of 1362-10 and 1362-11, who is homozygous for the inverted order (relative to the order shown in Figure 1). CEPH individuals 1362-10, 1413-03, and 1413-02 (mother of 1413-03 and 1413-09) were also found to be homozygous for the inverted order (data not shown).

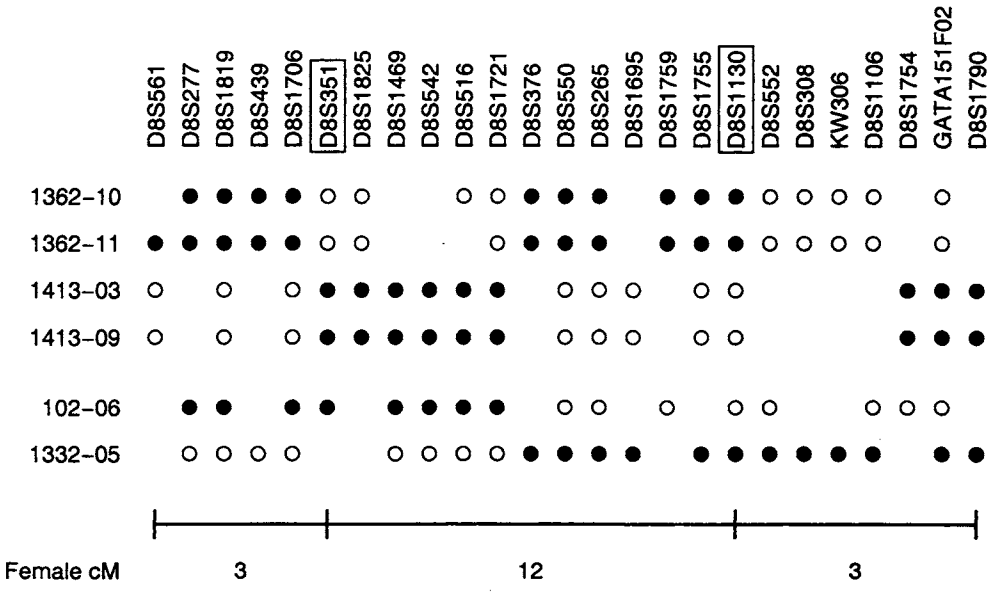


Figure 1: Maternal haplotypes for a small portion of chromosome 8*p* for six CEPH family children (identified by family – individual). Filled symbols indicate alleles from the maternal grandfather, open symbols alleles from the maternal grandmother, and blank spaces indicate missing data (due mostly to homozygous markers in the mother). The order of markers is telomeric (left) to centromeric (right). BACs encompassing the two markers shown in boxes were used in the FISH experiments.

Metaphase FISH carried out on 50 unrelated individuals of European ancestry revealed 33 homozygotes with the order shown in Figure 1, 13 heterozygotes, and 4 homozygotes for the inverted order (inversion frequency 21%; 95% confidence interval, assuming Hardy-Weinberg equilibrium, 13 – 30%). The genotype frequencies showed no significant deviation from Hardy-Weinberg equilibrium.

The inversion polymorphism appears to be either extremely old or the result of recurrent mutations. With only a single, relatively recent, inversion mutation event, and assuming no recombination in heterozygotes, there should only be one common "inverted" haplotype. With at least two inversion events occurring upon different haplotype backgrounds, recombination events in parents homozygous for the inversion could produce many different haplotypes. Although we don't know which orientation is ancestral, construction of haplotypes in the CEPH families using available genotyping data revealed several different haplotypes for each orientation. All three haplotypes for the order shown in Figure 1 were quite different with multiple (up to 17) repeat differences between alleles (data not shown). Similarly, all six haplotypes for the inverted order were very different. Since short tandem repeats (microsatellites) nearly always mutate by gain or loss of one or two repeat units [2, 28], it is unlikely that a single in-

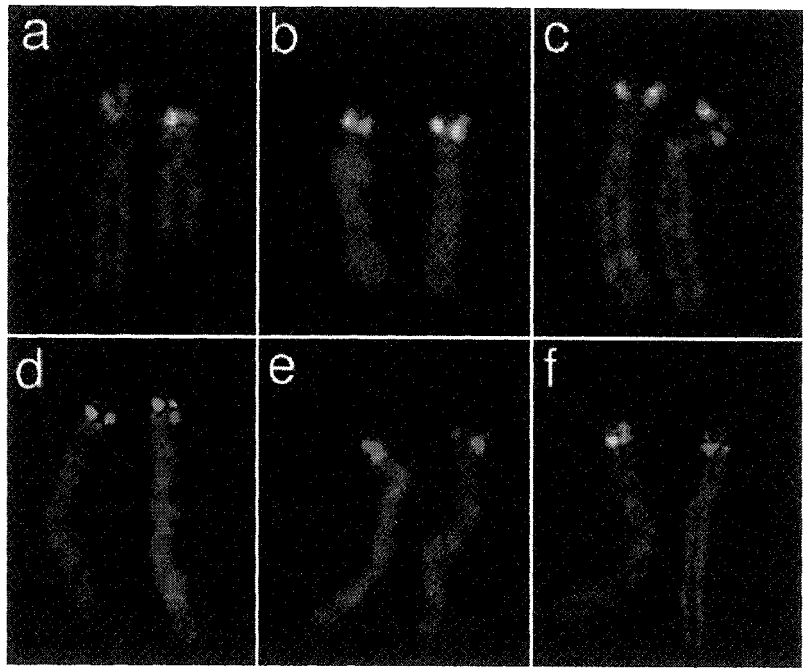


Figure 2: Metaphase FISH results from CEPH family individuals (lymphoblastoid cell lines) and other individuals (peripheral blood lymphocytes). Probes were DNA from BAC 173O4 (encompassing D8S351) labeled with Spectrum Green and BAC 257O3 (encompassing D8S1130) labeled with Spectrum Orange. a) CEPH individual 102-01, homozygous for the order shown in Figure 1. b) CEPH individual 1331-02, heterozygous for the inversion. c) CEPH individual 1362-02, homozygous for the inverted order. d) e) f) Individuals homozygous for the Figure 1 order, heterozygous, and homozygous for the inverted order, respectively.

version mutation event occurred within the last few hundred thousand years. Although we cannot rule out a single event that occurred longer ago, we favor the alternative that at least two inversion mutation events occurred relatively recently.

We further characterized the inversion through examination of relevant genome maps. The genetic length of the inverted region is approximately 12 and 2 cM on the female and male genetic maps, respectively [3]. Using the December 22, 2001, version of the University of California-Santa Cruz draft human sequence, the length of the inverted region was estimated to be at least 2.5 Mb and possibly as long as 5.3 Mb. The sequence assembly in this region of 8*p* is still crude with many gaps, both large and small, and other uncertainties. Sites of the inversion breakpoints are not yet precisely known. From both the CEPH and FISH results, the inversion breakpoints appear to be at similar locations in all individuals; however, the precision of these approaches is limited.

The inversion is likely mediated by two clusters of olfactory receptor genes that flank the inverted segment at both ends [9]. Olfactory receptor genes are found on nearly every human chromosome [11]. The flanking repeated sequences are apparently in inverted orientation (Matsumoto *et al.*, in preparation). The 48 kb emerin/filamin inversion on the X chromosome is also flanked by 11 kb inverted repeat sequences [26]. Intrachromatid recombination between inverted non-adjacent repeat sequences results in the inversion of the intervening segment. As the human genomic sequence becomes finished, it may be possible to identify additional inversion polymorphisms through searches for intrachromosomal inverted repeats with high sequence similarity.

The 8*p* inversion may have substantial clinical impact. For example, Giglio *et al.* [9] studied eight mothers of children with the inverted duplication 8it p rearrangement, and found that all were heterozygous for the inversion described herein. Inv dup (8*p*) is a well-known chromosomal abnormality of maternal origin that causes multiple abnormalities including mental retardation [7, 13, 17]. The frequency of inv dup (8*p*) has been estimated at 1/15,000 [9]. It may be that women heterozygous for the chromosome 8*p* inversion that we identified are more likely to bear children with the inv dup (8*p*) rearrangement. Giglio *et al.* [10] recently identified another inversion polymorphism, on chromosome 4p16, which is also flanked by clusters of olfactory receptor genes. These two chromosomal inversions, on chromosomes 4 and 8, appear to be involved in the recurrent t(4;8)(p16;p23) translocation.

Heterozygotes for the chromosome 8*p* inversion may also have slightly reduced fertility compared to homozygotes of either genotype due to unbalanced gametes produced through recombination within the inverted region. The rearrangement may also affect the expression of genes near the inversion breakpoint. Such effects are well known for translocations [18]. Genes within or adjacent to the inverted segment include several defensins, GATA-binding protein 4 (GATA4), cathepsin B (CTSB), tankyrase (TNKS), and methionine sulfoxide reductase A (MSRA).

Submicroscopic inversions are difficult to identify. Use of improbable meiotic products as an inversion signature (see Figure 1) becomes much more difficult as the size of the inversion decreases. For inversion of only two or three adjacent markers, the phase patterns will masquerade as genotyping errors or mutations. Also, a recombination event is required within the inverted region for detection, and it may be necessary for the parent to be homozygous for the inversion for recombination to occur. A better approach to detect inversion polymorphisms is likely to be comparison of various genome maps, especially including sequence assemblies, which are prepared using DNA from different donors. Our results clearly demonstrate that differences in marker order between various genome maps should not automatically be dismissed as errors.

Acknowledgments

This work was supported by NIH grants N01–HV–48141 to JLW and R01–HD–36111 to DHL.

Karl W. Broman, Center for Medical Genetics, Marshfield Medical Research Foundation (current address: Department of Biostatistics, Johns Hopkins University), kbroman@jhsph.edu

Naomichi Matsumoto, Department of Human Genetics, University of Chicago (current address: Department of Human Genetics, Nagasaki University School of Medicine)

Sabrina Giglio, Department of Human Genetics, University of Chicago (current address: Diagnostica e Ricerca, San Raffaele, Milan)

Christa Lese Martin, Department of Human Genetics, University of Chicago

Jessica A. Roseberry, Department of Human Genetics, University of Chicago

Orsetta Zuffardi, Biologia Generale e Genetica Medica, Università di Pavia

David H. Ledbetter, Department of Human Genetics, University of Chicago

James L. Weber, Center for Medical Genetics, Marshfield Medical Research Foundation

References

[1] M. L. Bondeson, N. Dahl, H. Malmgren, W. J. Kleijer, T. Tonnesen, B. M. Carlberg, and U. Pettersson. Inversion of the IDS gene resulting from recombination with IDS-related sequences is a common cause of the Hunter syndrome. *Human Molecular Genetics*, 4:615–621, 1995.

[2] B. Brinkmann, M. Klintschar, F. Neuhuber, J. Huhne, and B. Rolf. Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. *American Journal of Human Genetics*, 62:1408–1415, 1998.

[3] K. W. Broman, J. C. Murray, V. C. Sheffield, R. L. White, and J. L. Weber. Comprehensive human genetic maps: individual and sex-specific variation in recombination. *American Journal of Human Genetics*, 63:861–869, 1998.

[4] K. W. Broman and J. L. Weber. Characterization of human crossover interference. *American Journal of Human Genetics*, 66:1911–1926, 2000.

[5] S. S. Chong, S. D. Pack, A. V. Roshke, A. Tanigami, R. Carrozzo, A. C. M. Smith, W. B. Dobyns, and D. H. Ledbetter. A revision of the lissencephaly and Miller-Dieker syndrome critical regions in chromosome 17p13.3. *Human Molecular Genetics*, 6:147–155, 1997.

[6] J. Dausset, H. Cann, D. Cohen, M. Lathrop, J. M. Lalouel, and R. White. Centre d'Etude du Polymorphisme Humain (CEPH): collaborative genetic mapping of the human genome. *Genomics*, 6:575–577, 1990.

[7] C. E. M. de Die-Smulders, J. J. Engelen, C. T. Schrander-Stumpel, L. C. Govaerts, B. de Vries, J. S. Vles, A. Wagemans, S. Schijns-Fleuren, G. Gillessen-Kaesbach, and J.-P. Fryns. Inversion duplication of the short arm of chromosome 8: clinical data on seven patients and review of the literature. *American Journal of Medical Genetics*, 59:369–374, 1995.

[8] Y. J. M. de Kok, G. F. M. Merkx S. M. van der Maarel, I. Huber, S. Malcolm, H. H. Ropers, and F. P. M. Cremers. A duplication/paracentric inversion associated with familial X-linked deafness (DFN3) suggests the presence of a regulatory element more than 400 kb upstream of the POU3F4 gene. *Human Molecular Genetics*, 4:2145–2150, 1995.

[9] S. Giglio, K. W. Broman, N. Matsumoto, V. Calvari, G. Gimelli, T. Neumann, H. Ohashi, L. Voullaire, D. Larizza, R. Giorda, J. L. Weber, D. H. Ledbetter, and O. Zuffardi. Olfactory receptor gene clusters, genomic-inversion polymorphisms, and common chromosome rearrangements. *American Journal of Human Genetics*, 68:874–883, 2001.

[10] S. Giglio, V. Calvari, G. Gregato, G. Gimelli, S. Camanini, R. Giorda, A. Ragusa, S. Guerneri, A. Selicorni, M. Stumm, H. Tonnies, M. Ventura, M. Zollino, G. Neri, J. Barber, D. Wieczorek, M. Rocchi, and O. Zuffardi. Heterozygous submicroscopic inversions involving olfactory recepter-gene clusters mediate the recurrent t(4;8)(p16;p23) translocation. *American Journal of Human Genetics*, 71:276–285, 2002.

[11] G. Glusman, I. Yanai, I. Rubin, and D. Lancet. The complete human olfactory subgenome. *Genome Research*, 11:685–702, 2001.

[12] P. Green, K. Falls, and S. Crooks. *Documentation for CRI-MAP, version 2.4*, 1990.

[13] W. J. Guo, F. Callif-Daley, M. C. Zapata, and M. E. Miller. Clinical and cytogenetic findings in seven cases of inverted duplication of 8p with evidence of a telomeric deletion using fluorescence *in situ* hybridization. *American Journal of Medical Genetics*, 58:230–236, 1995.

[14] M. A. Huynen, B. Snel, and P. Bork. Inversions and the dynamics of eukaryotic gene order. *Trends in Genetics*, 17:304–306, 2001.

[15] M. A. Jobling, G. Williams, K. Schiebel, A. Pandya, K. McElreavey, L. Salas, G. A. Rappold, N. A. Affara, and C. Tyler-Smith. A selective difference between human Y-chromosomal DNA haplotypes. *Current Biology*, 8:1391–1394, 1998.

[16] H. Kehrer-Sawatzki, B. Schreiner, S. Tanzer, M. Platzer, S. Muller, and H. Hameister. Molecular characterization of the pericentric inversion that causes differences between chimpanzee chromosome 19 and human chromosome 17. *American Journal of Human Genetics*, 71:375–388, 2002.

[17] A. Kleczkowska, J.-P. Fryns, F. D'Hondt, J. Jaeken, and H. van den Berghe. Partial duplication 8p due to interstitial duplication: inv dup (8)(p21.1 – p22). *Annals of Genetics*, 30:47–51, 1987.

[18] D.-J. Kleinjan and V. van Heyningen. Position effect in human genetic disease. *Human Molecular Genetics*, 7:1611–1618, 1998.

[19] K. Lagerstedt, S. L. Karsten, B. M. Carlberg, W. J. Kleijer, T. Tonnesen, U. Pettersson, and M. L. Bondeson. Double-strand breaks may initiate the inversion mutation causing the Hunter syndrome. *Human Molecular Genetics*, 6:627–633, 1997.

[20] D. Lakich, H. H. Kazazian, S. E. Antonarakis, and J. Gitschier. Inversions disrupting the factor VIII gene are a common cause of severe haemophilia A. *Nature Genetics*, 5:236–241, 1993.

[21] S. Muller, R. Stanyon, P. Finelli, N. Archidiacono, and J. Wienberg. Molecular cytogenetic dissection of human chromosomes 3 and 21 evolution. *Proceedings of the National Academy of Sciences USA*, 97:206–211, 2000.

[22] E. Nickerson and D. L. Nelson. Molecular definition of pericentric inversion breakpoints occurring during the evolution of humans and chimpanzees. *Genomics*, 50:368–372, 1998.

[23] M. J. Pettenati, P. N. Rao, M. C. Phelan, F. Grass, K. W. Rao, P. Cosper, A. J. Carroll, F. Elder, J. L. Smith, M. D. Higgins, J. T. Lanman, R. R. Higgins, M. G. Butler, F. Luthardt, E. Keitges, C. Jackson-Cook, J. Brown, S. Schwartz, D. L. Van Dyke, and C. G. Palmer. Paracentric inversions in humans: a review of 446 paracentric inversions with presentations of 120 new cases. *American Journal of Medical Genetics*, 55:171–187, 1995.

[24] M. T. Pletcher, T. Wiltshire, D. E. Cabin, M. Villanueva, and R. H. Reeves. Use of comparative physical and sequence mapping to annotate mouse chromosome 16 and human chromosome 21. *Genomics*, 74:45–54, 2001.

[25] J. M. Ranz, F. Casals, and A. Ruiz. How malleable is the eukaryotic genome? Extreme rate of chromosomal rearrangement in the genus Drosophila. *Genome Research*, 11:230–239, 2001.

[26] K. Small, J. Iber, and S. T. Warren. Emerin deletion reveals a common X-chromosome inversion mediated by inverted repeats. *Nature Genetics*, 16:96–99, 1997.

[27] C. A. Tilford, T. Kuroda-Kawaguchi, H. Skaletsky, S. Rozen, L. G. Brown, M. Rosenberg, J. D. McPherson, K. Wylie, M. Sekhon, T. A. Kucaba, R. H. Waterston, and D. C. Page. A physical map of the human Y chromosome. *Nature*, 409:943–945, 2001.

[28] X. Xu, M. Peng, Z. Fang, and X. Xu. The direction of microsatellite mutations is dependent upon allele length. *Nature Genetics*, 24:396–399, 2000.

[29] J. J. Yunis and O. Prakash. The origin of man: a chromosomal pictorial legacy. *Science*, 215:1525–1530, 1982.

The Roles of Mutation Rate and Selective Pressure on Observed Levels of the Human Mitochondrial DNA Deletion mtDNA^{4977}

William C. Navidi, Simon Tavaré and Norman Arnheim

Abstract

The mitochondrial deletion mtDNA^{4977} has been found at high levels in individuals with certain neuromuscular and neurological diseases, and at lower levels in older normal individuals. We use experimental estimates of the mutation rate of mtDNA^{4977} and of the half-life of mitochondrial genomes to construct a model of mitochondrial replication and mutation that is consistent with observed levels of the deletion. We conclude that deleted genomes have a slight selective advantage, at least in some tissues. Our results suggest that for an individual to attain a clinically significant level of the deletion, between 0.2% to 0.5% of the mitochondrial genomes in the original oöcyte must have been deleted.

Keywords: branching process; Kearns-Sayre syndrome; mitochondria; selection

1 Introduction

The human mitochondrial mutation mtDNA^{4977} is a 4977 base pair deletion originating between two 13 bp direct repeats in normal mtDNA. This deletion is associated with the neuromuscular and neurological diseases progressive external ophthalmoplegia (PEO), Kearns-Sayre syndrome (KSS) and Pearson's marrow/pancreas syndrome. Symptoms of these sporadic diseases range from mild to severe, depending on the level to which the deleted molecules have accumulated. For a review of diseases associated with mutations in mitochondrial DNA, see DiMauro and Wallace [10], Wallace [26], DiMauro [9] and Bianchi *et al.* [3]. MITOMAP (`http://www.mitomap.org`) is a very useful resource for human mitochondrial data and references.

The mtDNA^{4977} deletion has also been found at low levels in normal adults and appears to accumulate with age, primarily in non-mitotic tissues (Cortopassi *et al.* [8], Arnheim and Cortopassi [1], Corral-Debrinski *et al.* [7], Hattori *et al.* [14], Yen *et al.* [27], Zhang *et al.* [28]). The level of accumulation is found to vary among different tissues and even within tissues. For example, studies on the brains of old normal individuals has shown that the substantia nigra, caudate and putamen can have hundreds of fold higher levels of mtDNA^{4977} than the cerebellum (Corral-Debrinski *et al.* [7], Soong *et al.* [25]).

The degree of accumulation of mtDNA[4977], whether at the high levels found in patients with disease, or at the low levels found in normal adults, is determined by the mutation rate, by selective factors that may favor deleted molecules, and by the initial level of deletions present at conception. We describe models that enable us to discuss quantitatively the roles played by each of these factors in the accumulation of deletions in both growing and stable populations of cells.

2 A Stochastic Model for Deletions

The mitochondrial genomes in a human cell are distributed among many mitochondria, with an average of between four and ten genomes per mitochondrion. Mitochondria can turn over by being engulfed by lysosomes. In what follows we refer to this turnover as mitochondrial death.

Let ω_1 and ω_2 represent the probabilities that a nonmutant and a mutant, respectively, die before replication. Let λ represent the mutation rate, defined as the probability that replication of a nonmutant sequence produces a mutant. We assume that a mutant always gives rise to mutants upon replication.

Define one generation to be the length of time between replications of a nonmutant molecule. Let $2r$ be the mean number of descendants of a single mutant molecule after one generation, conditional on survival to replication of the original molecule. Then the unconditional expectation of the one generation clone size of a single mutant is $2r(1-\omega_2)$, and that of a single nonmutant is $2(1-\omega_1)$.

We define the selective advantage ν of mutants over nonmutants in terms of the ratio of the expected one-generation clone sizes: $\nu = r(1-\omega_2)/(1-\omega_1)$. Notice that any value of ν can be obtained by setting $r = 1$, and choosing ω_1 and ω_2 appropriately. For the sake of simplicity, we assume henceforth that $r = 1$, and that any selective advantage is due to differences in the death rates ω_1 and ω_2.

We model the evolution of the population of mitochondria as a two-type Galton-Watson process, the two types being nonmutant (type 1) and mutant (type 2). We calculate the probability $p_i(x,y)$ that a parent of type i ($i = 1,2$) produces x nonmutants and y mutants in a single mitochondrial genome replication. We have

$$p_1(0,0) = \omega_1, \qquad p_1(2,0) = (1-\omega_1)(1-\lambda), \qquad p_1(1,1) = (1-\omega_1)\lambda$$

$$p_2(0,0) = \omega_2, \qquad p_2(0,2) = 1-\omega_2, \tag{1}$$

and $p_i(x,y) = 0$ for other values of x and y.

Let M denote the 2×2 matrix whose ij^{th} element m_{ij} is the mean number of offspring of type j produced by a parent of type i in a single replication. From (1) we have

$$M = \begin{pmatrix} (1-\omega_1)(2-\lambda) & (1-\omega_1)\lambda \\ 0 & 2(1-\omega_2) \end{pmatrix}. \tag{2}$$

Given values m_{g-1}, n_{g-1} for the expected number of mutant and nonmutant genomes, respectively, after $g-1$ generations, the values m_g, n_g can be computed as

$$(n_g, m_g) = (n_{g-1}, m_{g-1})M. \tag{3}$$

By repeatedly applying (3) starting from $g = 1$ and values for n_0, m_0, we can compute m_g and n_g for any value of g. We approximate the proportion of mutants in the population after g generations by $m_g/(m_g + n_g)$. For theoretical results in this spirit, see Olofsson and Shaw [19].

3 Constant-size Populations

We can use our model to evaluate the possible roles that both mutation and selection could play in the accumulation of deletions over time. We first investigate the relationships between mutation rate, selective advantage, and the proportion of mutants expected in regions of the brain. For example, the level of mtDNA^{4977} deletions in the substantia nigra has been estimated to be as high as 0.5% in an 80-year-old normal individual (Soong *et al.* [25]).

As before, let ω_1 and ω_2 represent the probabilities that a nonmutant and a mutant, respectively, die before replication, and let λ represent the probability that a replication of a nonmutant genome produces a mutant genome. Let m_0, n_0 be the initial number of mutants and nonmutants, respectively, present at birth, and let m_g, n_g be the expected numbers of mutants and nonmutants after g mitochondrial generations.

From (2) and (3), the quantities m_g and n_g satisfy the recursive equations

$$m_g = 2(1-\omega_2)m_{g-1} + \lambda(1-\omega_1)n_{g-1} \tag{4}$$

$$n_g = (2-\lambda)(1-\omega_1)n_{g-1}. \tag{5}$$

We use the ratio

$$\nu = (1-\omega_2)/(1-\omega_1) \tag{6}$$

to express the selective advantage of mutants over nonmutants. Values of ν greater than 1 correspond to an advantage for mutants, and values less than 1 correspond to an advantage for nonmutants.

Since the brain is primarily a non-dividing tissue, the number of genomes $G = m_g + n_g$ is assumed to be constant across generations. It follows that in each generation g, the death rates ω_1 and ω_2 satisfy the following equations:

$$1-\omega_1 = G/2(\nu m_{g-1} + n_{g-1}) \tag{7}$$

$$1-\omega_2 = \nu(1-\omega_1). \tag{8}$$

Given values of ν, m_0, and n_0, we can calculate values of m_g and n_g for any value of g by repeatedly calculating ω_1 and ω_2 from equations (7) and (8) and substituting into equations (4) and (5).

In our calculations, we took the mitochondrial generation time to be 45 days, based on experimental estimates of turnover in the rat brain (Gross *et al.* 1969). Table 3 gives values for the expected proportion m_{650}/G of mutants after 650 generations (about 80 years) as a function of the selective advantage ν for three values of λ, assuming no mutants were present at birth. The values of λ we chose bracket the estimate of Shenkar *et al.* [23]), who estimated the mutation rate of mtDNA^{4977} in cultured human cells to be $5.95 \times 10^{-8} \pm 2.28 \times 10^{-8}$.

It is clear that only a small selective advantage is needed to produce a large proportion of mutants, and that the value of the mutation rate has less impact. The second column in Table 3 shows that this is consistent with a selective advantage in the range of 1.012. Estimates of mtDNA^{4977} levels in putamen of old individuals may be as high as 12% (Corral-Debrinski *et al.* [7]) which would be consistent with a selective advantage in the range of 1.018.

The model yields a different conclusion regarding the cerebellar grey matter. Of thirteen regions of the brain studied by Soong *et al.* [25], this had the smallest fraction of deleted genomes, with the proportion being only 0.0013% in an 82 year old individual. Table 3 shows that this would be consistent with the absence of a selective advantage in the cerebellar grey matter, if we were to assume that the mutation rates in this region is the same as in the substantia nigra.

Of course, differences in mutation rates in different tissues can also explain differences in accumulation. Table 3 gives values for the expected proportion of mutants after 650 generations as a function of mutation rate, assuming no selective advantage. Comparing Tables 3 and 3 shows that an increase of about two orders of magnitude in the mutation rate (from 6×10^{-8} to 5×10^{-6}) is needed to produce the same result as a selective advantage of 1%, i.e. $\nu = 1.01$.

4 Growing Populations

The mtDNA^{4977} deletion accumulates to at least a level of 40% of the mitochondrial genomes in muscle cells of the vast majority of children with KSS (Shanske *et al.* [22]). We develop a model to investigate the roles of selective pressure and mutation rate on the accumulation of mtDNA^{4977} in these children. In order to accomplish this we need to take into consideration three phases of the child's development.

The first phase begins with a single fertilized oöcyte (zygote) containing approximately 150,000 mitochondrial genomes (Chen *et al.* [6]). This cell divides until there are about 125 descendants, 40 of which comprise the inner cell mass and are destined to become the embryo (Hardy *et al.* [13]). We assume that a cell needs a minimum of about 7500 mitochondrial genomes to survive. It follows that the number of mitochondria in the 125 cells must be about 7500×125. Therefore the number of mitochondria in the original oöcyte (150,000) must have multiplied by a factor of 6.25 ($= 7500 \times 125 / 150{,}000$).

In the second phase the 40 embryo cells replicate to form a fetus which we estimate contains approximately 2.5×10^{12} cells at birth. During these two mitotic phases the

Table 1: Expected proportions of mutants after 650 mitochondrial generations for various values of the mutation rate and selective advantage.

Selective advantage (ν)	$\lambda = 1.0 \times 10^{-8}$	$\lambda = 5.95 \times 10^{-8}$	$\lambda = 1.0 \times 10^{-7}$
1.022	0.24010	0.65278	0.75960
1.020	0.08866	0.36663	0.49313
1.018	0.02930	0.15224	0.23185
1.016	0.00936	0.05327	0.08640
1.014	0.00299	0.01755	0.02914
1.012	0.00097	0.00574	0.00961
1.010	0.00032	0.00191	0.00320
1.008	0.00011	0.00066	0.00110
1.006	0.00004	0.00024	0.00040
1.004	0.00002	0.00009	0.00015
1.002	0.00001	0.00004	0.00007
1.000	0.00000	0.00002	0.00003

rate of mitochondrial turnover is likely to be insignificant. The third phase covers the period after birth where we only consider mitochondrial turnover and not cell replication. Our goal is to express the proportion of genomes that have the deletion in the muscle cells of a child 10 years of age as a function of the number of deleted genomes in the oöcyte and the selective advantage ν of deleted over non-deleted genomes, assuming a constant mutation rate.

Let N be the number of deleted genomes in the oöcyte. In the first phase, these N molecules multiply to a level of approximately $6.25N$ deleted molecules that are distributed among 125 cells. We assume that the distribution is random, so that the number of deleted molecules in a single one of the cells has a binomial distribution with parameters $6.25N$ and $1/125$. We approximate this with a Poisson distribution with mean $\mu = 6.25N/125 = N/20$.

To model the second and third phases, we focus on a single embryo cell and use the stochastic model described in equations (4) and (5). To apply the model, we specified values for the mutation rate λ, the death rates ω_1 and ω_2, the initial number m_0 of mutants and n_0 of non-mutants, and the number of generations g. We used the value 5.95×10^{-8} for λ, as estimated by Shenkar *et al.* [23]). For the second (mitotic) phase, we assumed there is no mitochondrial turnover, so that $\omega_1 = \omega_2 = 0$. Note that this implies an absence of a selective advantage in the mitotic phase. As explained below, the degree of selective advantage in the second mitotic phase has little impact on the final level of deletions, so this assumption is not crucial. We assumed that the cell contained a total of $m_0 + n_0 = 7500$ mitochondrial genomes. As described above, the quantity m_0 has a Poisson distribution with mean $\mu = N/20$. We assumed that

Table 2: Expected proportions of mutants after 650 generations for various values of the mutation rate assuming no selective advantage.

Mutation rate	Expected proportion of mutants
1.0×10^{-8}	0.00000
5.0×10^{-8}	0.00002
1.0×10^{-7}	0.00003
5.0×10^{-7}	0.00016
1.0×10^{-6}	0.00032
5.0×10^{-6}	0.00162
1.0×10^{-5}	0.00324
5.0×10^{-5}	0.01612
1.0×10^{-4}	0.03198
5.0×10^{-4}	0.15000
1.0×10^{-3}	0.27753
5.0×10^{-3}	0.80349
1.0×10^{-2}	0.96154

the second phase begins with 40 cells and ends with 2.5×10^{12} cells. It follows that the number of mitochondrial generations g satisfies the equation $40 \times 2^g = 2.5 \times 10^{12}$. This provides the estimate $g = 36$, to the nearest integer.

For the third phase, we used the method (described earlier for the brain) for a constant size population, in which we consider mtDNA turnover. We specify a value for the selective advantage ν, then compute ω_1 and ω_2 from equations (7) and (8). To estimate the number of generations g', we used an experimental estimate of a half-life of one week for mitochondria in muscle cells (Gross *et al.* [12]), which corresponds to a mean turnover time of about 10 days. Therefore 10 years corresponds to $g' \approx 350$ mitochondrial generations. The initial numbers of mutants and nonmutants for the third phase were of course the final numbers for the second phase.

Given values for the initial number N of deleted mitochondria in the oöcyte and the selective advantage ν, we computed the fraction of deleted mitochondria deriving from a single inner mass cell for all feasible values of m_0, as described above. We then averaged this fraction over the Poisson distribution of m_0 to obtain expected fraction $d(N,\nu)$ of deleted mitochondria. For many different values of N, we computed the value of ν for which $d(N,\nu) = 0.5$. This fraction is chosen to reflect the observed levels of deletions seen in skeletal muscle of a child with Kearns-Sayre syndrome.

Figure 1 presents the results for values of N ranging from 0 to 75,000. The latter number corresponds to the situation in which one half of the original genomes have the deletion. Figure 2 presents the results for values of N ranging from 0 to 1000. In order to check the impact of our assumptions concerning the lack of turnover in the mitotic

phase, we redid the calculations setting ω_1 to 0.1, and assuming the selective advantage was the same in both the mitotic and constant-size phases. The results (not shown) were nearly identical to those presented.

It is clear from Figure 1 that unless the initial level of deletions in the oöcyte is nearly one-half, the deleted molecules must have a selective advantage in order to reach the levels of 50% observed in a child. Figure 2 shows that a selective advantage of 4% would result in the proportion of deletion reaching one-half when there were no initial oöcyte deletions, and deletions arose post-zygotically by mutations alone. Our earlier analysis of deletion levels in the brain estimates selective values in the range 1.2% to 1.8%, suggesting that a selective advantage of 4% is implausible. If we assume that the selective advantage is indeed in the range 1.2% to 1.8%, then the number of initial deletions leading to a level of 50% after ten years is of the order of 300 to 750, or between 0.2% and 0.5% of the oöcyte's mitochondrial genomes.

There is some empirical evidence about the frequency of deleted molecules in oöcytes (cf. Chen *et al.* [6], Brenner *et al.* [4], Barritt *et al.* [2]). For example, Chen *et al.* [6] observed that in a sample of 15 oöcytes, approximately one half had no detectable deletions, while the rest had an average of roughly 50 deletions each. The maximum number of deletions observed was in the range 100–200. Figure 2 shows that if the selective advantage were greater than about 2%, this initial level of deletions in an oöcyte would lead to a level of 50% deletions by age ten, and thus a disease prevalence of about 7%. However, the observed prevalence of the disease is between 1/100,000 and 1/500,000 (Larsson *et al.* [16]). It follows that the selective advantage of the deleted molecules must be less that 2%, and we conclude that KSS is due to the very rare (frequency of around 1/100,000) oöcytes that have 350-700 deletions, with selective advantage in the range 1.2% - 1.8%.

5 Discussion

We have used a model of mitochondrial replication and mutation to evaluate the possible roles that both mutation and selection could play in the accumulation of deletions over time, in both expanding and stable mitochondrial populations. We note however that the mechanism by which mutant mtDNA accumulate in patients with mitochondrial diseases is a matter for debate (cf. Marchington *et al.* [17], Reynier *et al.* [21], Jansen [15], Shoubridge [24], Brown *et al.* [5], Elson *et al.* [11], Qintana-Murci *et al.* [20]), and that more details of this process will be needed for a definitive quantitative analysis.

We begin with the stable population case. Using experimental estimates of the half-life of a mtDNA genome (Gross *et al.* [12]) and of the mtDNA^{4977} mutation rate (Shenkar *et al.* [23]), we have shown that starting from a collection of brain cells that do not divide, each of which begins with no mutant mitochondria, after 650 generations (80 human years) we would expect no more than 0.002% mutant genomes if the mutants had no selective advantage.

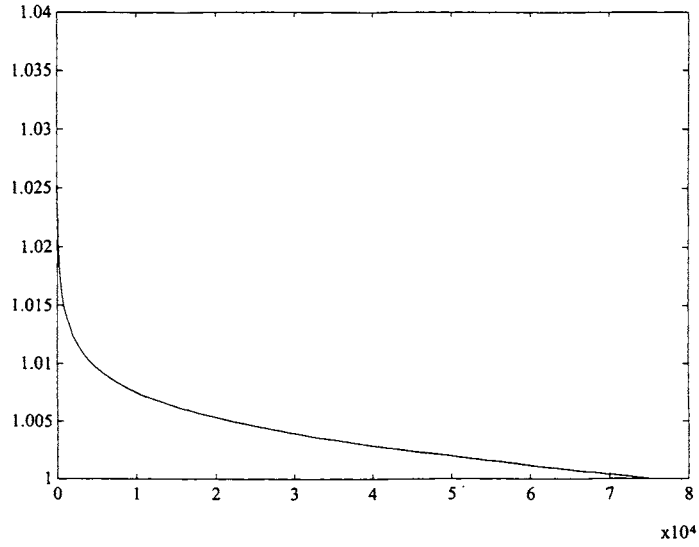


Figure 1: Value of selective advantage ν (y axis) required to give $d(N,\nu) = 0.5$ for different values of N (x axis).

That deleted genomes do have a replicative selective advantage is plausible *a priori*, since mtDNA^{4977} is two-thirds the size of normal mtDNA genomes. Using the estimates of the half life and the mutation rate as above, we calculate that even a selective advantage of as little as 1.2% could result in the levels observed in the substantia nigra of old individuals (Table 3). Experimental estimates of this selective advantage based on differences in the rates of completion of mtDNA circles during replication showed no evidence of any difference between deleted and undeleted genomes (Moraes and Schon [18]), but such studies could not have detected an advantage on the order of a few percent.

mtDNA^{4977} accumulation to the reported levels in some brain regions (0.5% to 12%) as a result of mutation alone would require a mutation rate about three orders of magnitude higher than the experimental estimate. It has been argued that the mutation rate may be high in the substantia nigra (as well as the caudate and putamen), due to the extra burden of oxidative damage that might result from the high levels of metabolism of the neurotransmitter dopamine by MAO-B (Soong *et al.* [25], Corral-Debrinski *et al.* [7]). This enzyme generates H_2O_2, which can react with iron deposits inside cells to produce hydroxyl radical leading to DNA damage. On the other hand, some brain regions show deletion levels hundreds of times smaller than in the substantia nigra and putamen. If mutation rates were the same, it is not clear how selective pressures could vary greatly among cells in different regions of the brain.

We applied our method to determining the extent to which mutations in oöcytes and selective pressure contribute to the deletion levels observed in the mitochondria of chil-

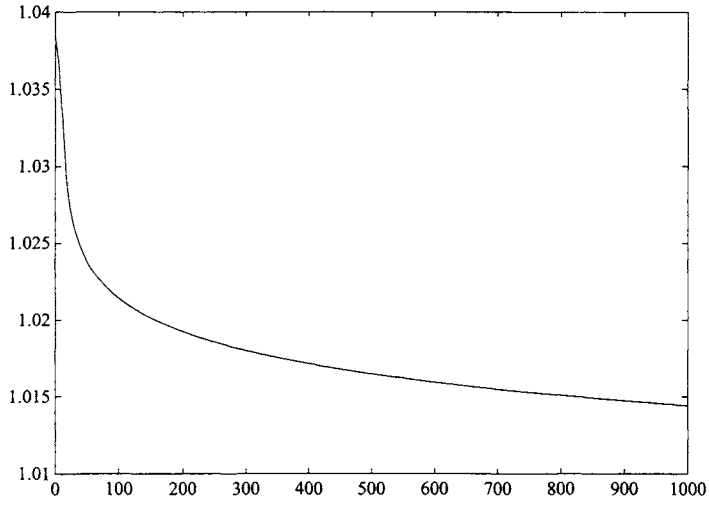


Figure 2: Value of selective advantage ν (y axis) required to give $d(N,\nu) = 0.5$ for different values of N (x axis).

dren with Kearns-Sayre syndrome. We described a hypothesis that quantifies the effects of the initial levels of deletions in oöcytes, the mutation rates, and the selective advantage of mutants. Although the number of mutant genomes in a person is random, we have based our analysis on expected values. We have estimated the selective advantage and the number of initial deletions necessary for the mean number of deletions after 10 years to amount to 50% of the total number of mitochondrial genomes. Of course, a few individuals will have deletion levels several standard errors above the mean. The standard error of the number of deletions produced in the second and third stages, however, is of the order of the square root of the mean, which is negligible in proportion to the mean. We conclude therefore that Kearns-Sayre syndrome cannot result solely from an unusually large number of mutations in the second and third stages.

This method can be generalized to situations involving several mutant types, with point mutations as well as deletions, and using time intervals other than the mitochondrial generation time. For example, the element in the second row and first column of the matrix (2) represents the rate of back point mutations multiplied by the mutant death rate.

The method outlined above is applicable in a wide variety of biological settings. Constant-size non-mitotic populations can be modelled by choosing values for ω_1 and ω_2 which indicate that on average half the genomes die each generation. Growing populations are modelled by choosing smaller values for these death rates. A selective advantage (for mutants, say) can be incorporated by choosing $\omega_2 < \omega_1$, indicating that mutants are more likely to replicate before dying than non-mutants are.

Dedication

It is a pleasure to dedicate this article to Terry Speed, friend, collaborator and teacher, on the occasion of his 60th birthday. The fields of statistics and genetics have been tightly bound together since their inception, and Terry's seminal contributions to the statistical analysis of molecular data stand in the forefront of this long tradition. Above all, Terry has shown us that one need not sacrifice mathematical rigor to obtain biological relevance. In this, he has set a standard toward which we continue to strive.

William C. Navidi, Department of Mathematics, Colorado School of Mines, `wnavidi@mines.edu`

Simon Tavaré, Program in Molecular and Computational Biology, University of Southern California, `stavare@usc.edu`

Norman Arnheim, Program in Molecular and Computational Biology, University of Southern California, `arnheim@molbio.usc.edu`

References

[1] N. Arnheim and G. Cortopassi. Deleterious mitochondrial DNA mutations accumulate in aging human tissues. *Mutation Research*, 275:157–167, 1992.

[2] J. A. Barritt, C. A. Brenner, S. Willadsen, and J. Cohen. Spontaneous and artificial changes in human ooplasmic DNA. *Human Reproduction*, 15(Suppl. 2):207–217, 2000.

[3] N. O. Bianchi, M. S. Bianchi, and S. M. Richard. Mitochondrial genome instability in human cancers. *Mutation Research*, 488:9–23, 2001.

[4] C. A. Brenner, Y. M. Wolny, J. A. Barritt, D. W. Matt, S. Munné, and J. Cohen. Mitochondrial DNA deletion in human oocytes and embryos. *Molecular Human Reproduction*, 4:887–892, 1998.

[5] D. T. Brown, D. C. Samuels, E. M. Michael, D. M. Turnbull, and P. F. Chinnery. Random genetic drift determines the level of mutant mtDNA in human primary oocytes. *American Journal of Human Genetics*, 68:533–536, 2001.

[6] X. Chen, R. Prosser, S. Simonetti, J. Sadlock, G. Jagiello, and E. A. Schon. Rearranged mitochondrial genomes are present in human oöcytes. *American Journal of Human Genetics*, 57:239–247, 1995.

[7] M. Corral-Debrinski, T. Horton, M. T. Lott, J. M. Shoffner, M. F. Beal, and D. C. Wallace. Mitochondrial DNA deletions in human brain: regional variability and increase with advanced age. *Nature Genetics*, 2:324–329, 1992.

[8] G. A. Cortopassi, D. Shibata, N-W. Soong, and N. Arnheim. A pattern of accumulation of a somatic deletion of mitochondrial DNA in aging human tissues. *Proceedings of the National Academy of Sciences USA*, 89:7370–7374, 1992.

[9] S. DiMauro. Lessons from mitochondrial DNA mutations. *Seminars in Cell and Developmental Biology*, 12:397–405, 2001.

[10] S. DiMauro and D. C. Wallace. *Mitochondrial DNA in human pathology*. Raven Press, New York, 1993.

[11] G. L. Elson, D. C. Samuels, D. M. Turnbull, and P. F. Chinnery. Random intracellular drift explains the clonal expansion of mitochondrial DNA mutations with age. *American Journal of Human Genetics*, 68:802–806, 2001.

[12] N. J. Gross, G. S. Getz, and M. Rabinowitz. Apparent turnover of mitochondrial deoxyribonucleic acid and mitochondrial phospholipids in the tissues of the rat. *Journal of Biological Chemistry*, 244:1552–1562, 1969.

[13] K. Hardy, A. H. Handyside, and R. M. Winston. The human blastocyte: cell number, death and allocation during late preimplantation development in vitro. *Development*, 107:597–604, 1989.

[14] K. Hattori, M. Tanaka, S. Sugiyama, T. Obayashi, T. Ito, T. Satake, Y. Hanaki, J. Asai, M. Nagano, and T. Ozawa. Age dependent increase in deleted mitochondrial DNA in the human heart: Possible contributory factor to presbycardia. *American Heart Journal*, 121:1735–1742, 1991.

[15] R. P. S. Jansen. Germline passage of mitochondria: quantitative considerations and possible embryological sequelae. *Human Reproduction*, 15(Suppl. 2):112–128, 2000.

[16] N-G. Larsson, E. Holme, B. Kristiansson, A. Oldfors, and M. Tulinius. Progressive increase of the mutated mitochondrial DNA fraction in Kearns-Sayre Syndrome. *Pediatric Research*, 28:131–136, 1990.

[17] D. R. Marchington, V. Macaulay, G. M. Hartshorne, D. Barlow, and J. Poulton. Evidence from human oocytes for a genetic bottleneck in an mtDNA disease. *American Journal of Human Genetics*, 63:769–775, 1998.

[18] C. T. Moraes and E. A. Schon. Replication of a heteroplasmic population of normal and partially-deleted human mitochondrial genomes. In F. Palmieri, S. Papa, C. Saccone, and M. N. Gadaleta, editors, *Progress in Cell Research: Symposium on "Thirty Years of Progress in Mitochondrial Bioenergetics and Molecular Biology"*, volume 5, pages 209–215. Elsevier, 1995.

[19] P. Olofsson and C. A. Shaw. Exact sampling formulas for multi-type Galton-Watson processes. *Journal of Mathematical Biology*, 45:279–293, 2002.

[20] L. Quintana-Murci, A. Rötig, A. Munnich, P. Rustin, and T. Bourgeron. Mitochondrial DNA inheritance in patients with deleted mtDNA. *Journal of Medical Genetics*, 38:e28, 2001.

[21] P. Reynier, M-F. Chrétien, F. Savagner, G. Larcher, V. Rohmer, P. Barriére, and Y. Mathiéry. Long PCR analysis of human gamete mtDNA suggests defective mitochondrial maintenance in spermatozoa and supports the bottleneck theory for oocytes. *Biochemical and Biophysical Research Communications*, 252:373–377, 1998.

[22] S. Shanske, C. T. Moraes, A. Lombes, A. F. Miranda, E. Bonilla, P. Lewis, M. A. Whelan, C. A. Ellsworth, and S. DiMauro. Widespread tissue distribution of mitochondrial DNA deletions in Kearns-Sayre syndrome. *Neurology*, 40:24–28, 1990.

[23] R. Shenkar, W. C. Navidi, S. Tavaré, M. H. Dang, A. Chomyn, G. Attardi, G. Cortopassi, and N. Arnheim. The mutation rate of the human mtDNA deletion $mtDNA^{4977}$. *American Journal of Human Genetics*, 59:772–780, 1996.

[24] E. A. Shoubridge. Mitochondrial DNA segregation in the developing embryo. *Human Reproduction*, 15(Suppl. 2):229–234, 2000.

[25] N-W. Soong, D. R. Hinton, G. Cortopassi, and N. Arnheim. Mosaicism for a specific somatic mitochondrial DNA mutation in adult human brain. *Nature Genetics*, 2:318–323, 1992.

[26] D. C. Wallace. Mitochondrial DNA sequence variation in human evolution and disease. *Proceedings of the National Academy of Sciences USA*, 91:8739–8746, 1994.

[27] T-C. Yen, J-H. Sue, K-L. King, and Y-H. Wei. Aging associated 5 kb deletion in human liver mitochondrial DNA. *Biochemical and Biophysical Research Communications*, 178:124–131, 1991.

[28] C. Zhang, A. Baumer, R. J. Maxwelland A. W. Linnane, and P. Nagley. Multiple mtDNA deletions in an elderly individual. *FEBS Letters*, 297:34–38, 1992.

DNA-Protein Binding and Gene Expression Patterns

Hongyu Zhao, Baolin Wu and Ning Sun

Abstract

Although many clustering methods have been applied to analyze gene expression data, genes in the same cluster may have neither common functions nor common regulation. As a result, computational approaches have been developed to identify motifs in the regulatory regions of a cluster of genes or of genes with similar gene expression levels that are responsible for DNA-protein binding and similar gene expression levels. However, these motifs are neither sufficient nor necessary for a transcription factor to bind to the promoter region of a gene with these motif patterns. More recently, molecular methods have been developed to directly measure DNA-protein binding at the genomic level. In this article, we first evaluate the predictive power of computational approaches to predict DNA-protein binding from a study involving nine transcription factors in the cell cycle. We then compare how much variation in gene expression levels can be explained either by the observed DNA-protein binding or by the binding predicted through computational approaches. We find that current computational approaches may be limited both in predicting DNA-protein binding as well as in predicting gene expression levels. We also observe indirectly that the correspondence between gene expression levels and protein levels may be rather poor, which suggests that there may be difficulty in modeling genetic networks purely through gene expression data. To better understand gene expression patterns, an integrated approach to incorporating different kinds of information should be developed.

Keywords: gene expression; DNA-protein binding; motif; microarray

1 Introduction

With the completion of the Human Genome Project, large-scale gene expression experiments have become common practice in the scientific community. Such experiments normally have different objectives: (1) to identify differentially expressed genes, (2) to identify genes expressed in a coordinated manner across a set of conditions, (3) to identify gene expression patterns that distinguish different samples (*e.g.* normal versus tumor tissues), and (4) to define global biological pathways. Genomics research is different from traditional molecular biology in that traditional approaches focus on the study of individual genes considered in isolation, whereas functional genomics allows researchers to determine the principles underlying complex biological processes (*e.g.*

development) by examining the expression patterns of large numbers of genes in parallel, taking into consideration temporal, as well as anatomical, patterns. Identification and characterization of regulatory *cis*-elements and *trans*-factors of a gene is essential for understanding the mechanisms of the control of gene expression, which can further shed light on gene function.

Currently, three types of statistical methods are under active development for gene expression data, including methods to identify differentially expressed genes (*e.g.* [8, 20] for cDNA arrays and [9, 22] for Affymetrix arrays), methods to identify clusters of genes with correlated expression patterns, *e.g.* [3, 10, 16, 21, 31], and methods to use gene expression patterns to distinguish phenotypes and predict clinical outcomes, *e.g.* [7, 13, 15, 32]. Although clustering methods have given some insight into gene function, similar gene expression patterns imply neither similar functions nor similar regulation for a group of genes. In addition, clustering results strongly depend on the set of experiments used to define similarities among genes, and results from different clustering algorithms may disagree with each other [12].

In contrast to standard statistical treatments of microarray data where data are mostly treated as a two-dimensional matrix, bioinformatics tools have been developed to use other information, mostly sequence information, to assist in the interpretation of gene expression patterns. For example, motif searches have been integrated in gene expression analysis in yeast studies, *e.g.* [4, 5, 24, 30]. The rationale is that genes having similar expression patterns are more likely to share common regulatory motifs in their promoter regions. These methods represent integration of expression data with sequence information. A more ambitious goal has been taken by some researchers to develop computational methods to reconstruct genetic networks, *e.g.* correlation metric construction [2], Boolean networks [1, 23, 28], and Bayesian networks [11, 14]. Unfortunately, most of these computational methods were not developed specifically for the analysis of gene expression data; therefore, it is difficult to incorporate biological information in these methods. They may generate results that are both hard to interpret and to verify, and they impose assumptions that are likely to be violated in real biological systems. This computational approach is in contrast to biologically driven approaches to dissecting pathways [18]. It has become clear that "the combination of predictive modeling with systematic experimental verification will be required to gain a deeper insight into living organisms, therapeutic targeting and bioengineering" [6].

Although many computational approaches have been proposed to identify DNA-protein binding motifs from gene expression patterns, such analyses may only provide indirect inference on binding. In addition, binding motifs are neither necessary nor sufficient for a given transcription factor to bind to the regulatory region of a gene [29]. Regulatory networks cannot be accurately deduced from expression profiles, partly because it is difficult to distinguish direct and indirect effects. Recently, experimental procedures have been developed to directly identify the *in vivo* genome binding sites for known transcription factors [19, 26]. Using this method, Simon *et al.* [29] studied genomic targets of nine known cell cycle transcription activators: Swi4, Swi6, Mbp1,

Fkh1, Fkh2, Mcm1, Ndd1, Ace2, and Swi5. MBF (Swi4 and Swi6) and SBF (Mbp1 and Swi6) control late G1 (cell cycle gap 1 phase) genes. Mcm1, together with Fkh1 or Fkh2, recruits the Ndd1 protein in late G2 (cell cycle gap 2) and controls G2/M (cell cycle gap 2 and mitosis phases) genes. Mcm1 is involved in M/G1 genes, whereas Swi5 and Ace2 control late M and early G1 genes [29]. Although Simon *et al.* [29] were able to infer binding motifs for each factor based on their data, they noted that the putative binding motifs are neither sufficient nor necessary to identify binding sites for a transcription factor.

In this article, using both gene expression data and binding data, we study how much DNA binding information explains gene expression levels through two approaches. In the first approach, we directly model expression levels as a function of the empirically measured binding of known transcription factors. In the second approach, we first infer putative motifs for each transcription factor based on the binding data, then predict binding based on these putative motifs, and finally model expression levels as a function of the predicted binding. Therefore, the second approach is an "indirect" computational method. We found that although the existing computational approaches yield significant associations between gene expression levels and predicted binding, the proportion of variation explained by these computational methods are much lower than those explained by empirically measured binding data. Our results suggest that better computational models and methods are needed to identify binding motifs and then to predict DNA-protein binding in the analysis and interpretation of gene expression data.

2 Methods

2.1 Gene expression data

We analyze cell cycle gene expression data reported in Spellman *et al.* [30], where yeast cell cultures were synchronized by three independent methods: α factor arrest, elutriation, and arrest of a cdc15 temperature-sensitive mutant. Approximately 800 genes, $>10\%$ of all yeast protein-coding genes, were identified as cell cycle regulated. In this article, we analyze the time course data from the α factor based synchronization experiment and gene expression levels of cell cycle regulated genes. The expression patterns of these genes were studied in detail by Spellman *et al.* [30] and a number of clusters of genes based on expression levels were investigated; this investigation included the identification of motifs for each gene cluster.

2.2 DNA binding data

The DNA binding data used in this article are those collected by Simon *et al.* [29]; the details of their experiments and statistical analysis of binding data can be found in Ren *et al.* [26]. Each experiment was done in triplicate. An estimate of the ratio of binding intensities of two fluorescents was calculated for each promoter region for a given transcription factor. This ratio, called the *binding ratio* here, is a measure of

the binding intensity of the given transcription factor. A statistical procedure was used by Simon *et al.* [29] to evaluate the statistical significance of the binding. In this article, we use their estimated p-values to assess statistical evidence for binding. These data revealed that genes encoding several of the cell cycle transcriptional regulators are themselves bound by other cell cycle regulators. Their data also suggested partial functional redundancy between homologous activators.

2.3 Motif searches

We use AlignACE [17, 27] to identify motifs that are over-represented in the upstream regulatory regions of a set of genes. In this article, we apply AlignACE to nine sets of genes, each of which were bound by the nine transcription factors, respectively. We then use CompareACE to identify those putative motifs that are similar to known motifs in yeast. Finally, ScanACE, a program that searches a genome for close matches to a motif found by AlignACE [17], is used to scan the upstream regions of the cell cycle regulated genes to identify those containing putative motifs. For each putative motif, each gene is defined as either having (coded "1") or not having (coded "0") this motif.

3 Results

3.1 Gene clusters based on binding data and gene clusters based on gene expression data

Transcription factors induce expression levels of cell cycle genes at different stages of the cell cycle. Simon *et al.* [29] observed consistency between DNA binding and gene expression levels. For example, SBF (Swi4 and Swi6) and MBF (Mbp1 and Swi6) are important activators of late G1 genes, and the expression levels for most of the genes bound by Swi4, Swi6, or Mbp1 are highest at the late G1 stage in the cell cycle [30]. When we cluster the nine transcription factors according to their binding ratios across the genome, transcription factors active at the same stage of the cell cycle are clustered together (Figure 1). However, when these factors are clustered according to their expression levels reported in Spellman *et al.* [30], there is no such ordering among them (Figure 2). This indicates that gene expression levels of the nine transcription levels are rather uninformative for correlating their functions in the cell cycle.

3.2 Binding motifs and binding ratios

To investigate how much computational methods can offer in predicting binding ratios, we apply AlignACE to genes bound by the same transcription factor to identify common motifs in the upstream promoter regions of these genes. We then select those putative motifs that are similar to known motifs, and run ScanACE on all cell cycle regulated genes to determine whether these motifs occur in the promoter regions of these genes. After this step, for each transcription factor, we fit a linear regression model

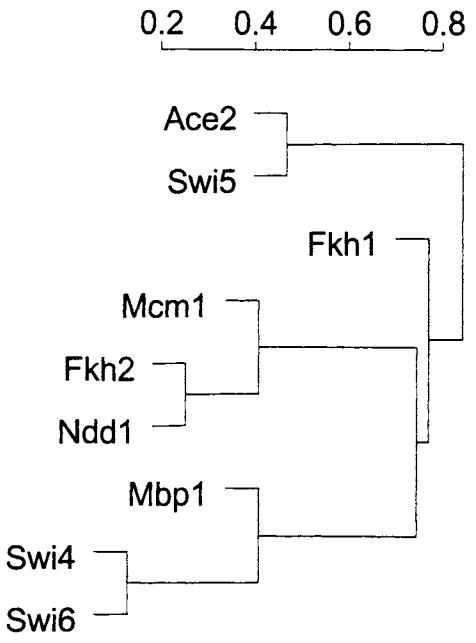


Figure 1: Clustering of nine transcription factors based on DNA binding data

with the observed binding ratios for cell cycle regulated genes as the response variable and the presence or absence of each putative motif in each gene as predictors, *i.e.*

$$y_i = \beta_1 M_{i1} + ... + \beta_k M_{ik} + e_i,$$

where y_i is the observed binding ratio for the ith gene, M_{ij} is a binary variable representing the presence ($M_{ij} = 1$) or absence ($M_{ij} = 0$) of the jth putative motif for this transcription factor in the i^{th} gene, and k is the number of putative motifs for this factor. In addition to this additive model, we also consider interactions among the M_{ij}, *i.e.*

$$y_i = \sum_{j=1}^{k} \beta_j M_{ij} + \sum_{j=1}^{k-1} \sum_{l=j+1}^{k} \gamma_{jl} M_{ij} M_{il} + e_i.$$

The results are summarized in Table 1, where all significant predictors for each transcription factor are listed, together with the proportion of variation in binding ratios explained by these predictors (R^2).

There are a few common features across all factors. First, SCB is the most commonly shared motif in these factors. Second, there are significant interaction terms for

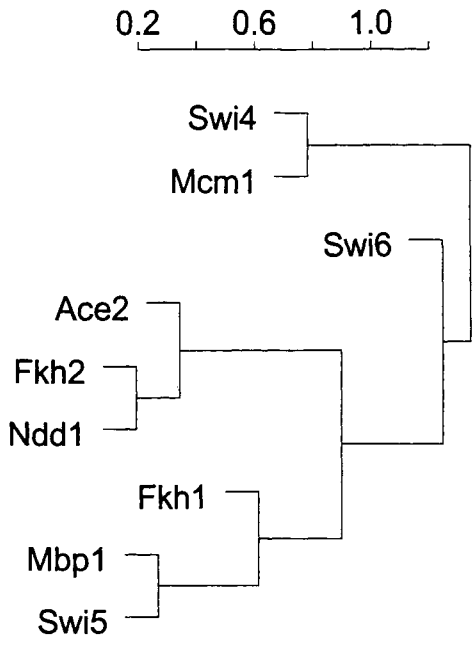


Figure 2: Clustering of nine transcription factors based on gene expression data

all the factors, which suggests that these putative motifs may interact with each other to recruit factors to the promoter regions. It is also clear from this table that the proportion of variation explained by the putative binding motifs varies across different transcription factors, with the variation in the binding ratios for Swi4, Swi6, and Ace explained most by the putative motifs. But overall, the R^2 is rather modest, which suggests that either there is substantial amount of measurement variation in binding ratios, or the motif search and binding prediction methods are far from satisfactory, or both.

3.3 Gene expression levels and empirically measured binding ratios

We consider how useful the binding ratios are to predict gene expression levels for cell cycle regulated genes. We analyze two sets of genes separately. The first set of genes includes all cell cycle regulated genes defined by Spellman *et al.* [30], whereas the second set of genes includes only those 298 genes that were found to be significantly bound (p-value < 0.001) by at least one of the nine transcription factors [29]. At each time point, for each set of genes, we first fit regression models with gene expression levels as the response variable and the observed binding ratios as predictors, *i.e.*

Table 1: Significant binding motifs as well as significant interactions among these motifs in the prediction of the binding ratios for each of the nine transcription factors studied. The last column is the proportion of variation explained by the joint effects of the binding motifs on the observed binding ratios

TF	Significant motifs and interactions	R^2
Swi4	SCB LEU MCB PDR SCB:MCB PDR:MCB	19%
Swi6	MCB STRE SCB:MCB MCB:STRE	14%
Mbp1	MCB RRPE MCB:RRPE	4%
Fkh1	RRPE STRE RRPE:STRE	1%
Fkh2	SCB RPN SCB:RPN	2%
Mcm1	SCB LYS SCB:LYS	4%
Ndd1	SCB REB SCB:REB	6%
Ace2	SCB LEU RAP STRE SCB:RAP SCB:STRE LEU:RAP	21%
Swi5	SCB STRE SCB:STRE	6%

$$y_i = \beta_1 R_{i1} + ... + \beta_9 R_{i9} + e_i,$$

where R_{ij} is the binding ratio between the ith gene and the jth factor, β_j is the regression parameter for the jth factor, and y_i is the observed expression level of the ith gene. The R^2 of the model for each of the 18 time points in the cell cycle are plotted in Figures 3 and 4 (solid lines).

It can be seen from these figures that the proportion of variation explained by the binding ratios is a function of time in the cell cycle, with the most variation explained at the S/G2 phase. The R^2 is increased if we focus on the subset of genes with each gene bound by at least one of the nine transcription factors. In the above analyses, we only consider the additive effects of different transcription factors. When interactions among factors are included in the model, we observe a significant increase in the R^2 for all time points. The comparisons between the additive models and the models with two-way interactions for the second set of genes are summarized in Figure 5, and the significant individual factors as well as significant interacting factors at each time point are summarized in Table 2. It can be seen that some interaction terms are significant, and that including interaction terms does improve the overall proportion of variation explained by the binding of these nine factors.

3.4 Gene expression levels and computationally predicted binding ratios

To evaluate the power of the predicted binding ratios in explaining gene expression levels, we fit regression models with the same response variable, *i.e.* gene expression

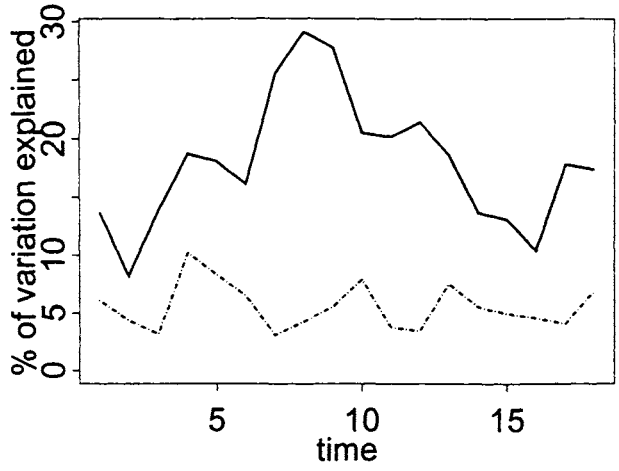


Figure 3: The proportions of variation explained by the observed binding data (solid line) and by the predicted binding (dotted line) for all cell cycle genes

levels, as above, but we use the predicted binding ratios through putative motifs as predictors this time. The R^2 of the model is plotted in Figures 3 and 4 (dashed line). It is clear from this figure that the observed binding data provides better information to explain expression levels. When interactions are included in the models, the overall R^2 is improved, but is still lower than that based on the empirically measured binding ratios. Therefore, although computational approaches are able to identify binding motifs that explain a statistically significant proportion of the variation in gene expression levels, their utility is limited compared to the directly observed binding data. Because we only consider nine transcription factors here, the unexplained proportion of the variation may be due to the effects from those transcription factors not included in the analysis, measurement errors in binding ratios, and sample variation in gene expression levels. Despite these other uncertainties, it is remarkable that these nine factors could explain up to 56% of the total variation at certain time points.

3.5 Estimation of transcription factor levels

These binding data also allow us to estimate relative protein expression levels for the transcription factors if we make the simple assumption that the effects of each transcription factor on inducing other genes' expression levels are proportional to the protein levels of the transcription factors in the cell. To estimate the protein levels of the nine transcription factors, we find the regression coefficients in the following regression model for each time point:

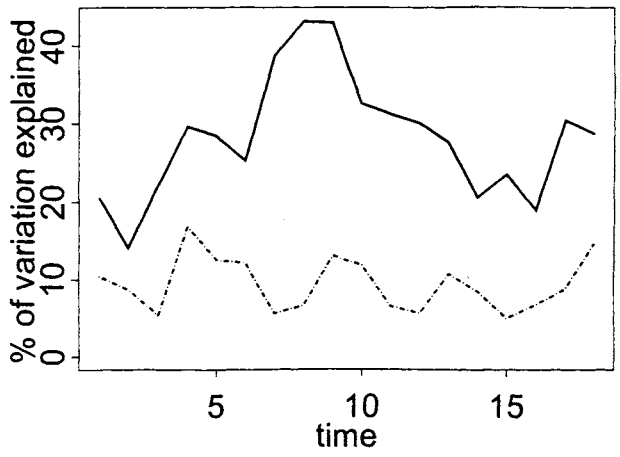


Figure 4: The proportions of variation explained by the observed binding data (solid line) and by the predicted binding (dotted line) for the 298 genes significantly bound by at least one of the nine transcription factors

$$y_i = B_{i1}L_1 + ... + B_{i9}L_9 + e_i,$$

where B_{ij} is the binding ratio between the ith gene and the jth transcription factor for the ith gene, and y_i is the gene expression level for the ith gene. Then the estimated L_j is the estimated protein expression level. Note that because the binding data only measure the relative levels, we should interpret the estimated L_j as equal to some constant times the protein level. Because we normalize the levels for the same protein across different time points in our summary (Figure 6), this is a reasonable approach to examine how the protein levels change across time. In Figure 6, we plot the observed gene expression levels and estimated protein expression levels at all 18 time points for each of the nine transcription factors. It can be seen from this figure that the correspondence between gene expression levels and protein levels is rather poor for some genes (*e.g.* Ace2), strong for some genes (*e.g.* Fkh1, Fkh2, and Ndd1), and a phase delay for other genes (*e.g.* Swi4 and Swi5).

4 Conclusions

We have first studied how well computational approaches can predict empirically observed DNA-protein interactions. Although we found that the computational approaches can yield results that are statistically significantly associated with the observed data, the

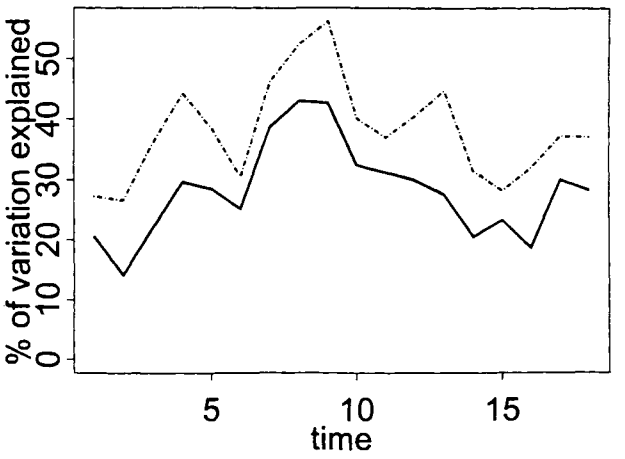


Figure 5: The proportions of variation explained by the observed binding data when only additive models are considered (solid line) and when interactions among factors are also considered (dotted line) for the 298 genes significantly bound by at least one of the nine transcription factors

correlation is rather modest. Current computational methods search for binding motifs separately; however, our results suggest the presence of interactions among putative binding motifs to jointly determine binding ratios. Similar observations were made by Pilpel *et al.* [25]. This suggests that interaction effects need to be taken into account in the search for binding motifs. Overall, even after interactions are taken into account, the proportion of variation in binding ratios explained by binding motifs through linear models is low. Therefore, there is ample room for methodology developments to predict DNA-protein binding.

We studied how well gene expression levels can be explained by DNA binding through two approaches. We found that a significant proportion of expression level variation across genes can be explained by the empirically measured DNA binding data. Similarly, computationally predicted binding also explain a significant proportion of the observed expression variation, but at much lower levels. We also investigated whether the predicted binding provide extra information to explain gene expression levels in addition to the observed binding by including both the observed binding and the predicted binding in the model. We found that the improvement of the model by the inclusion of the predicted binding was not significant (data not shown). Because it is well known that other transcription factors are involved in the cell cycle, we expect that the availability of binding data from other factors will further improve the prediction of the model. We also found that there is statistically significant evidence that

Table 2: Significant transcription factors and interacting terms in the prediction of gene expression levels at different time points

Time	Significant Terms
1	Ndd1, Ace2, Mbp1, Swi4, Mcm1:Swi4, Mcm1:Swi6, Mbp1:Swi6, Ndd1:Mcm1
2	Fkh1, Fkh2, Mbp1, Swi4, Mcm1:Swi4, Mcm1:Swi6, Mbp1:Swi6, Fkh1:Ndd1
3	Fkh2, Ndd1, Mbp1, Swi6, Ndd1:Swi5, Ace2:Swi5, Ndd1:Mbp1, Fkh1:Ndd1, Mcm1:Swi6
4	Fkh2, Ace2, Mbp1, Swi6, Fkh2:Ndd2, Ndd1:Swi6, Mcm1:Swi4
5	Ndd1, Mcm1, Ace2, Swi5, Mbp1, Swi6, Mbp1:Swi6, Ndd1:Mcm1
6	Fkh2, Mcm1, Ace2, Swi5, Swi6, Mbp1:Swi6
7	Fkh2, Ndd1, Mcm1, Ace2, Swi5, Swi4, Swi6, Fkh1:Fkh2, Mcm1:Swi6, Mbp1:Swi4, Swi5:Swi6
8	Fkh1, Fkh2, Ndd1, Mcm1, Ace2, Swi6, Fkh1:Ndd1, Mcm1:Swi6, Ace2:Swi5
9	Ndd1, Ace2, Swi5, Swi4, Swi6, Mcm1:Swi6, Fkh1:Ndd1, Ndd1:Mcm1
10	Ndd1, Mcm1, Ace2, Swi5, Swi4, Swi6, Ace2:Swi4
11	Fkh1, Mcm1, Swi5, Swi4, Fkh1:Swi4
12	Fkh2, Mcm1, Ace2, Swi5, Swi6, Mcm1:Swi6, Fk2:Ndd1, Fkh2:Ace2, Ndd1:Ace2
13	Ace2, Swi4, Swi6, Ace2:Swi4, Mcm1:Swi4, Ndd1:Swi5
14	Mbp1, Ace2, Swi4, Ace2:Swi4, Ndd1:Mcm1
15	Fkh2, Mcm1, Ace2, Swi5, Swi4, Ace2:Swi4
16	Fkh1, Fkh2, Ndd1, Swi6, Mcm1:Ace2, Fkh2:Ndd1, Ace2:Swi5, Ndd1:Swi5, Ndd1:Swi6
17	Fkh2, Ndd1, Ace2, Swi5, Mbp1, Swi6, Fkh2:Ndd1, Mcm1:Swi6, Mcm1:Ace2
18	Ndd1, Swi5, Swi6, Fkh2:Mcm1, Mcm1:Swi6, Fkh2:Swi6

different transcription factors interact with each other to contribute to the levels of gene expression. The interacting pairs not only include those known to work as a complex or present at the same stage of the cell cycle, they also include other pairs, suggesting that the interactions among these factors may be far more complex than currently thought.

In our analysis, we observed that the variation explained by the nine transcription factors is a function of time in the cell cycle. This indicates the importance of these nine transcription factors, as a group, varies at different stages of the cell cycle.

From the observed gene expression levels for different genes and the binding ratios between each gene and each factor, under a simple assumption, we were able to estimate the relative protein levels of the nine transcription factors studied. We found that although there is good correspondence between expression levels and "protein" levels for some factors, the correspondence is rather weak for others. There is no general relationship, and it appears that the relationship is both gene specific and time specific. The lack of consistency between gene expression data and protein expression data was noted by Ideker *et al.* [18]. However, factors with similar functions, *e.g.* Fkh1 and Fkh2, seem to have similar patterns between the observed gene expression data and the estimated protein expression levels. From the generally weak correlations between gene expression data and the estimated protein levels, we expect that computational models that only use gene expression data to reconstruct biological pathways may have limited power to make precise quantitative predictions. On the other hand, other types of data, such as the binding information, will be very useful in such efforts.

Another question that is of biological interest but has not been addressed in this paper is to examine how much of the gene expression similarities among a group of

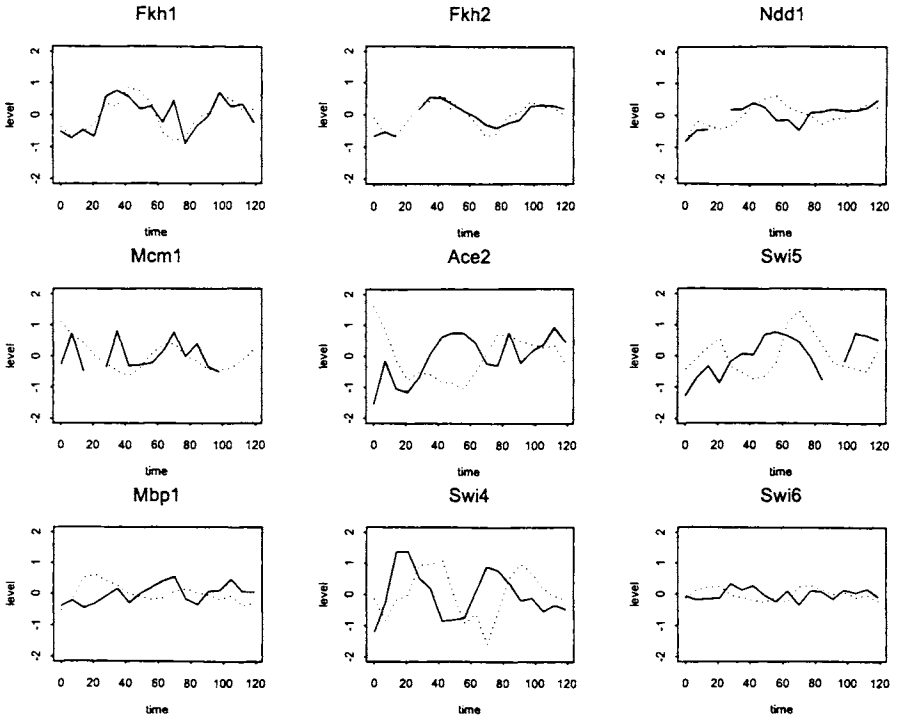


Figure 6: The observed expression levels of the nine transcription factors and their estimated protein expression levels. The data are normalized so that gene expression levels and protein expression levels have the same variance for a given transcription factor

genes can be explained by their regulation through a set of transcription factors. We can study this issue by comparing clusters derived purely from gene expression data and clusters derived purely from DNA binding data. Consistency between the two types of clusters would imply that the studied transcription factors may explain the regulation of these genes well, whereas a poor correlation implies that there are major mechanisms that drive the gene expression patterns but have not been uncovered or included in the study.

We have mainly used AlignACE and ScanACE to identify binding motifs for a group of genes. There are other computer programs available for motif findings and they may offer results better than we have found here. In addition, we have only considered those putative motifs that are similar to known motifs for the nine transcription factors. Although this procedure may exclude some unknown motifs that could play some role in determining DNA-protein binding, the likelihood of missing motifs with strong effects is small: these factors have been under intensive study by yeast geneticists, thus we expect that motifs with strong effects would have been identified. We are currently conducting a more thorough analysis to assess the importance of these

unmatched putative motifs.

Here we have considered the binding as a continuous measurement using the estimated binding ratios from replicate experiments. When we tried to dichotomize the binding data through the p-values reported by Simon *et al.* [29] (0 for the absence of binding and 1 for the presence of binding), the overall fit of the models is not as good as those we reported above (data not shown). This suggests that the continuous measurements do have more information on the regulation and interactions between genes and the transcription factors.

The ultimate goal of genomics studies is to understand biological pathways. In this article, we have shown the limitation of one existing computational method for studying gene regulation and the need to integrate gene expression data with other types data to dissect biological pathways. Incorporating DNA binding data is only the first step to move beyond purely statistical approaches for gene expression analysis.

Acknowledgements

We thank the reviewer for careful reading of this manuscript. Research supported in part by NIH grants GM59507, DK58776, and Grant IRG–58–012–45 from the American Cancer Society.

Hongyu Zhao, Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, hongyu.zhao@yale.edu

Baolin Wu, Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, baolin.wu@yale.edu

Ning Sun, Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, ning.sun@yale.edu

References

[1] T. Akutsu, S. Miyano, and S. Kuhara. Inferring qualitative relations in genetic networks and metabolic pathways. *Bioinformatics*, 16:727–734, 2000.

[2] A. Arkin, P. Shen, and J. Rose. A test case of correlation metric construction of a reaction pathway from measurements. *Science*, 277:1275–1279, 1997.

[3] A. Ben-Dor and Z. Yakhini. Clustering gene expression patterns. *S. Istrail, P. Pevzner, and M. S. Waterman (eds) Recomb 99, ACM Press, Washington, DC*, pages 188–197, 1999.

[4] H. J. Bussemaker, H. Li, and E. D. Siggia. Regulatory element detection using correlation with expression. *Nature Genetics*, 27:167–171, 2001.

[5] R. J. Cho, M. J. Campbell, L. Steinmetz E. A. Winzeler, A. Conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielian, D. Landsman, D. J. Lockhart, and R. W. Davis. A genome wide transcriptional analysis of mitotic cell cycle. *Molecular Cell*, 2:65–73, 1998.

[6] P. D'haeseleer, S. Liang, and R. Somogyi. Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*, 16:707–726, 2000.

[7] S. Dudoit, J. Fridlyand, and T. P. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97:77–87, 2002a.

[8] S. Dudoit, Y. H. Yang, T. P. Speed, and M. J. Callow. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, 12:111–139, 2002b.

[9] B. Efron, R. Tibshirani, J. D. Storey, and V. Tusher. Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96:1151–1160, 2001.

[10] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of National Academy of Sciences USA*, 95:14863–14868, 1998.

[11] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using bayesian networks to analyze expression data. *Journal of Computational Biology*, 7:601–620, 2000.

[12] D. R. Goldstein, D. Ghosh, and E. M. Conlon. Statistical issues in the clustering of gene expression data. *Statistica Sinica*, 12:219–240, 2002.

[13] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

[14] A. J. Hartemink, D. K. Gifford, T. S. Jaakkola, and R. A. Young. Bayesian methods for elucidating genetic regulatory networks. *IEEE Intelligent Systems*, March/April:37–43, 2002.

[15] T. Hastie, R. Tibshirani, D. Botstein, and P. Brown. Supervised harvesting of expression trees. *Genome Biology*, 2:research0003.1–0003.12, 2001.

[16] T. Hastie, R. Tibshirani, M. B. Eisen, A. Alizadeh, R. Levy, L. Staudt, W. C. Chan, D. Botstein, and P. Brown. 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biology*, 1:research0003.1–0003.21, 2000.

[17] J. D. Hughes, P. W. Estep, S. Tavazoie, and G. M. Church. Computational identification of *cis*-regulatory elements associated with functionally coherent groups of genes in *Saccharomyces cerevisiae*. *Journal of Molecular Biology*, 296:1205–1214, 2000.

[18] T. Ideker, V. Thorsson, J. A. Ranish, R. Christmas, J. Buhler, J. K. Eng, R. Bumgarner, D. R. Goodlett, R. Aebersold, and L. Hood. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, 292(5518):929–934, 2001.

[19] V. R. Iyer, C. E. Horak, C. S. Scafe, D. Botstein, M. Snyder, and P. O. Brown. Genomic binding sites of the yeast cell-cycle transcription factors sbf and mbf. *Nature*, 409:533–538, 2001.

[20] M. K. Kerr, M. Martin, and G. A. Churchill. Analysis of variance for gene expression microarray data. *Journal of Computational Biology*, 7:819–837, 2000.

[21] L. Lazzeroni and A. Owen. Plaid models for gene expression data. *Statistica Sinica*, 12:61–86, 2002.

[22] C. Li and W. H. Wong. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of National Academy of Sciences USA*, 98:31–36, 2001.

[23] S. Liang, S. Fuhrman, and R. Somogyi. Reveal, a general reverse engineering algorithm of genetic network architectures. *Pacific Symposium on Biocomputing*, 3:18–29, 1998.

[24] X. Liu, D. L. Brutlag, and J. S. Liu. Bioprospector: Discovering conserved DNA motifs in upstream regulatory regions. *Pacific Symposium on Biocomputing*, 6:127–138, 2001.

[25] Y. Pilpel, P. Sudarsanam, and G. M. Church. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature Genetics*, 29:153–159, 2001.

[26] B. Ren, F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T. L. Volkert, C. J. Wilson, S. P. Bell, and R. A. Young. Genome-wide location and function of DNA binding proteins. *Science*, 290:2306–2309, 2000.

[27] F. P. Roth, J. D. Hughes, P. W. Estep, and G. M. Church. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature Biotechnology*, 16:939–945, 1998.

[28] I. Schmulevich, E. R. Dougherty, S. Kim, and W. Zhang. Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*, 18:261–274, 2002.

[29] I. Simon, J. Barnett, N. Hannett, C. T. Harbison, N. J. Rinaldi, T. L. Volkert, J. J. Wyrick, J. Zeitlinger, D. K. Gifford, T. S. Jaakkola, and R. A. Young. Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell*, 106:697–708, 2001.

[30] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycles regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell*, 9:3273–3297, 1998.

[31] P. Tamayo, P. D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. Golub. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proceedings of National Academy of Sciences USA*, 96:2907–2912, 1999.

[32] M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J. R. Marks, and J. R. Nevins. Predicting the clinical status of human breast cancer using gene expression profiles. *Proceedings of National Academy of Sciences USA*, 98:11462–11467, 2001.

Blind Inversion Needs Distribution (BIND): General Notion and Case Studies

Lei Li

Abstract

A class of scientific measurement problems share a common feature which we refer to as "blind inversion." That is, we can regard a module of measurement instruments as a system with quantities to be measured as input and observations as output. In a blind inversion problem, both the effective system and the input are unknown to us. Due to either experimental design or the nature of scientific problem in question, very often the distributional knowledge of the input can be obtained. Given this piece of information, we apply a two-step scheme – abbreviated by BIND – to solve the blind inversion problem. First, we make use of the distributions of the input and output to estimate the system. Second, we reconstruct the value of each individual input using the system obtained in the first step. From this perspective, we have another look at two measurement problems that are part of Professor Speed's recent research in molecular biology. We also connect the idea with the long-standing predictive deconvolution method used in seismology and discuss assessment issues of BIND.

Keywords: blind inversion; color-correction; DNA sequencing; electrophoresis; microarray; seismology

1 Introduction

Scientific discoveries are based on accurate measurements. The innovation of measurement instruments and invention of conceptual models cross each other's track and lead each other's way throughout the history of science. As instrumental techniques advance and the collected information expands, new tools of data analysis emerge along the way. One such famous historical example is Gauss's use of least squares in astronomy and geodesy. Not only have ingenious algorithms been applied to the practice of data analysis, but probabilistic models such as regression models have also been proposed and widely accepted for the purpose of designing and evaluating measurement processes. Nowadays, it has become common sense that uncertainty is the nature of any measurement processes.

In the area of biology, human beings' understanding of life has experienced great breakthroughs at the molecular level since the last century. Based on new understandings, scientists have developed *in vitro* bio-techniques such as cloning and polymerase

chain reaction (PCR). Even more excitingly, by incorporating cutting-edge technologies from physics, chemistry, mechanics, and computer science, modules of genotyping, DNA sequencing, and monitoring of mRNA abundance have been well integrated. As a result of these engineering efforts, many current biological projects have scaled up to the genome level. Consequently, new problems of experimental design, measurement, and analysis arise to challenge researchers in different areas. In this article, we consider two biological measurement problems relating to laser and dye techniques.

A class of scientific measurement problems share a common feature which we refer to as "blind inversion." As shown in Figure 1, we can regard a module of measurement instruments as a system with quantities to be measured as input and observations as output. A full explanation of the figure can be found in Section 2. Even though in some cases we are, at least approximately, able to describe the system structure by a parametric model, it is sometimes difficult to determine the effective parameters because of uncontrollable internal or external factors that affect the performance of the instrument. Thus, both the effective system and the input are unknown in a blind inversion problem. Without further information, the problem is ill-posed because the solution is not unique. In order to define a well-posed problem, more knowledge is required. The nature of the blind inversion problem does not allow us to inquire for either system parameters or values of individual input. It seems that the only choice goes to the distributional knowledge of the input.

Although we formulate the input distribution in statistical language, the knowledge, if there is any, really comes from considerations of the scientific measurement problem in question. As shown later in our examples, because of either the experimental design or the nature of the scientific problem, quantities to be measured usually demonstrate some kind of canonical distributional form.

Given the information of the input distribution, we apply a two-step scheme – abbreviated by BIND – to solve the blind inversion problem. First, we make use of the distributions of the input and output to estimate the system. Second, we reconstruct the value of each individual input using the system obtained in the first step. It is interesting to notice that we use observations twice yet in two different ways. An analog to this dual perspective of the same dataset is the dual nature of light. Sometimes we adopt the perspective of particles – photons – to understand phenomena such as the photo-electric effect. At other times we adopt the perspective of waves to analyze phenomena such as interference, reflection, and refraction. According to the quantum mechanical explanation, the electromagnetic wave is closely associated with a probability distribution; see Fowles [10].

BIND is more a general notion than a precise solution for a specific problem. We realized its value from two recent biological measurement problems; however, it is certain that researchers have already explored similar ideas, consciously or unconsciously, to solve problems in different scenarios. We point out one such example in the discussion section. Still, we would like to spell it out for a broader awareness.

We arrange the materials in this paper as follows. In Section 2 we illustrate the

idea of BIND by an artificial example. In Section 3 we show how to apply the BIND scheme to achieve an adaptive color-correction for DNA sequencing data proposed by Li and Speed [13]. In Section 4 we have another look at the within-slide normalization procedure proposed by Yang, Dudoit, Luu, and Speed [26], from the perspective of BIND. In Section 5 we connect BIND with the predictive deconvolution method in seismology and discuss assessment issues of BIND.

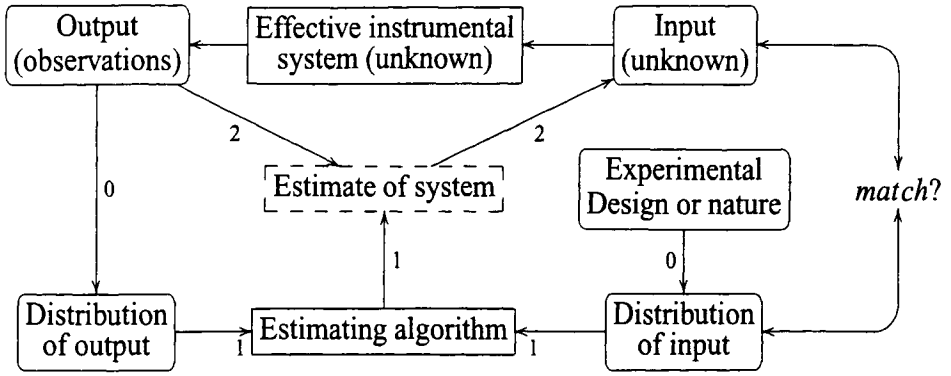


Figure 1: **The schematic representation of the blind inversion problem and BIND.** At the top, each individual input, which is to be measured, goes through the instrumental system and the corresponding output is observed. In a blind inversion problem, both input values and the effective system are unknown. BIND includes three steps. Step 0: identify the distributions of the input and output; Step 1: estimate the system function using the distributional information; Step 2: reconstruct each individual input value.

2 An illustrative example

Consider the following linear system,

$$\begin{bmatrix} y_1(t) \\ y_2(t) \\ y_3(t) \\ y_4(t) \end{bmatrix} = \begin{bmatrix} w_{11} & 0 & 0 & 0 \\ w_{21} & w_{22} & 0 & 0 \\ w_{31} & w_{32} & w_{33} & 0 \\ w_{41} & w_{42} & w_{43} & w_{44} \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \\ x_4(t) \end{bmatrix}, \tag{1}$$

where the input vector $x(t) = (x_1(t), x_2(t), x_3(t), x_4(t))'$ is unknown and to be estimated, the output vector $y(t) = (y_1(t), y_2(t), y_3(t), y_4(t))'$ is observed, and the system matrix $W = [w_{ij}]$ is lower-triangular and non-degenerate. If the system function is given, then the problem is easily solved by inverting the matrix $W = [w_{ij}]$. If the system is not given, then both $[w_{ij}]$ and $x(t)$ are unknown, and this is a blind inversion problem. Without further information, it is an ill-posed problem in the sense that the solution is not unique.

Interestingly, the distributional information of $x(t)$, if it is available somehow, can help solve the blind inversion problem. Let us assume that the distribution of the input

in (1) is **white and normal**, namely, $N(0,\sigma^2 I)$, where I is an identity matrix of order four. Consequently, the distribution of output is normal $N(0,\Sigma)$, where $\Sigma = WW'$. From observations on the output, we construct an empirical estimate of the covariance matrix by the standard method

$$\hat{\Sigma} = \frac{1}{T}\sum_{t=1}^{T}[y(t)-\bar{y}][y(t)-\bar{y}]',$$

where $\bar{y} = \frac{1}{T}\sum_{t=1}^{T} y(t)$ is the average of the observations. Then we estimate the matrix by factorizing $\hat{\Sigma}$. The uniqueness of the factorization is the direct result of the lower-triangular assumption on W. In fact, this is the **Cholesky factorization** for positive definite systems; see [11]. Denote the estimated system matrix by $\hat{W}$. Then for each observation of $y(t)$, we can estimate the corresponding input by $\hat{x}(t) = \hat{W}^{-1}y(t)$.

There is an interesting "cross-talk" interpretation of this model. Suppose a communication system has four channels. The first channel provides perfect transmission and no other channels interfere with it. The second channel is interfered with by the first channel, namely, the first channel leaks some signal to the second. The third channel experiences interference from the first and second channels. The fourth channel is the worst and is interfered with by all the other three channels. This leakage phenomenon can be described by a linear system such as that in (1), in which the signals on the sender's side and receiver's side are respectively represented by $x(t)$ and $y(t)$. The lower triangular cross-talk matrix W is consistent with the above interference structure. In order to reconstruct the original signals from the receiver's side, we need to clear the interference among the channels. If we assume that the signals being transmitted are independent among the four sources and approximately follow a normal distribution, then we can use the above procedure to estimate the signals from the sender's side.

Algorithm 1

(BIND) *The general scheme of blind inversion has three steps as shown in Figure 1:*

Step 0. identify the distributions of the input and output;

Step 1. estimate the system function using the distributional information of the input and output;

Step 2. invert the system and reconstruct each individual input value.

We refer to this idea as BIND (blind inversion needs distribution) hereafter. Measure theory (see Billingsley [3]) sheds some light on the need for the inquiry into the distribution of input. In the absence of singularity, the system function is like the **Radon-Nikodym derivative** of the output distribution with respect to the input distribution. Notice that we have equated terms of distribution and measure in this discussion. The exact meaning of distributional information we refer to here include: first, the support of the measure or the value space; second, the distribution on this space demonstrated by the input. Two general issues ought to be addressed. On the one hand, we expect that the distributional information should be complementary to any partial information

about the system and **sufficient** enough to define a well-posed inversion problem. On the other hand, despite any mathematical formulation, the hypothesis on the input distribution should be based on **scientific considerations** of the problem in question, and we also expect that the hypothesis can be verified to some extent.

3 Color correction of DNA sequencing data

In 1995, I started to do research under the Terry's supervision at UC Berkeley. Around that time, the Human Genome Project was in its accelerating stage; the Lawrence Berkeley and Livermore National Laboratories were part of this joint effort. The key component of this project and of any other genome project is Sanger sequencing; see the book edited by Adams, Fields, and Ventor [1] for background in molecular biology. While working on the crucial problem of physical mapping, David Nelson and Terry [16] initiated research on DNA sequencing and base-calling. The problem interested me and later Simon Cawley. Eventually my thesis [12] and part of Simon's [5] grew out of this research topic. One part of our DNA sequencing work is the correction of the dye cross-talk effect; see Li and Speed [13]. At the time we proposed our algorithm, we did not think much about the underlying principle. Now we explain it according to the BIND scheme. The primary idea of Sanger sequencing lies in its specially-designed **dideoxy enzymatic reactions**. Starting with a target DNA segment, the four dideoxy reactions respectively produce many copies of each possible sub-fragment ending with A, G, C, and T; see Russell [23]. For example, the four kinds of subfragments of a DNA fragment ATTCAGCGT are given by {A, ATTCA}, {ATTCAG, ATTCAGCG}, {ATTC, ATTCAGC} and {AT, ATT, ATTCAGCGT}. These sub-fragments are separated and ordered according to their sizes by electrophoresis, carried out in either a gel or a capillary. A slab gel contains many lanes, yet lane-tracking is required to extract lane signals from raw image data. Capillary electrophoresis, on the other hand, does not require lane-tracking. In order to differentiate the four kinds of sub-fragments from the same electrophoresis lane, each kind of sub-fragment in the enzymatic reaction is labeled with one of four dyes. By design, these four dyes demonstrate different light spectra with respect to a laser of a specific frequency. The problem is to measure the dye concentrations of the four kinds at one region. Excited by the laser, the four dyes emit photons, which are collected in four wavelength bands. However, the observations – four fluorescence intensities – are not direct measurements of the dye concentrations of the four kinds. This is where the complication comes in. The dataset used in this article was from slab gel electrophoresis and was provided by the Human Genome Center at LBNL. In Figure 2 is shown a portion of the fluorescence intensities (top) and the reconstructed dye concentrations (bottom). In the plot of dye concentrations, there is a series of peaks of four colors. The rationale of DNA sequencing and base-calling is: each peak represents one base, and the order of color peaks is consistent with the order of nucleotide bases on the underlying DNA fragment. The color code in Figure 2 is: A – red, G – black, C – green, and T – blue. We notice that adjacent peaks of the same

color overlap and this is where deconvolution is required; see Li and Speed [14]. In comparison with dye concentrations, peaks in the plot of fluorescence intensities are not clean in the sense that they have components in all four colors. Next we explain this phenomenon in some detail.

The spectra of the four dyes used in fluorescence-based DNA sequencing overlap, and thus the cross-talk phenomenon arises. That is, the observed four fluorescence intensities are a transformed version of the four dye concentrations. The transformation is not completely linear because of instrumental limitations. For example, overflow may occur in photon-counters if too many photons are emitted in a short period. Nevertheless, we approximately describe the relationship between the unknowns – four dye concentrations $C(t)$, $G(t)$, $A(t)$ and $T(t)$ – and the observations – four fluorescence intensities $I_1(t), I_2(t), I_3(t), I_4(t)$ – at an electrophoretic time t by the following linear system:

$$\begin{bmatrix} I_1(t) \\ I_2(t) \\ I_3(t) \\ I_4(t) \end{bmatrix} = \begin{bmatrix} 1 & w_{12} & w_{13} & w_{14} \\ w_{21} & 1 & w_{23} & w_{24} \\ w_{31} & w_{32} & 1 & w_{34} \\ w_{41} & w_{42} & w_{43} & 1 \end{bmatrix} \begin{bmatrix} C(t) \\ G(t) \\ A(t) \\ T(t) \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix}, \tag{2}$$

where $[w_{ij}]$ is the cross-talk matrix and (b_1, b_2, b_3, b_4) is the baseline. Note that there are only 12 free parameters in the cross-talk matrix because the spectra are determined by relative fluorescence intensities except for scaling. We parameterize the cross-talk matrix in such a way that its diagonal elements are unity, i.e., $w_{ii} = 1$. We simplify the problem by assuming that the baseline is constant, and we support this assumption by the following argument. First, the baseline refers to the fluorescence background of the measured region. It changes slowly along a lane in a relatively small range with respect to signals. Second, although observations are recorded on a time scale as shown in Figure 2, our view of their distribution ignores their time-dependence. This is equivalent to permuting data. According to our simplification, all kinds of variations except for cross-talk are implicitly aggregated into measurement errors.

The goal of color-correction is to reconstruct the dye concentrations using data of fluorescence intensities. If the cross-talk matrix is known, then a straightforward inversion solves the problem. However, the **effective cross-talk matrix is unknown** and needs to be estimated. Thus we are facing a blind inversion problem, or more specifically, an **adaptive color-correction** problem. According to the general scheme of blind inversion, first of all, we need to consider the distribution of the input – dye concentrations.

The following **non-overlapping hypothesis** is crucial for understanding the problem. Although dye concentrations change from lane to lane, and from gel to gel, we discovered that their **distributional pattern changes little across lanes**. The distribution can be graphically displayed by **pairwise scatter plots**; one such example is shown in Figure 3 (the data shown in the figure is explained after Algorithm 2). The first sub-plot only includes concentrations of C and G fragments. Two distinct cluster

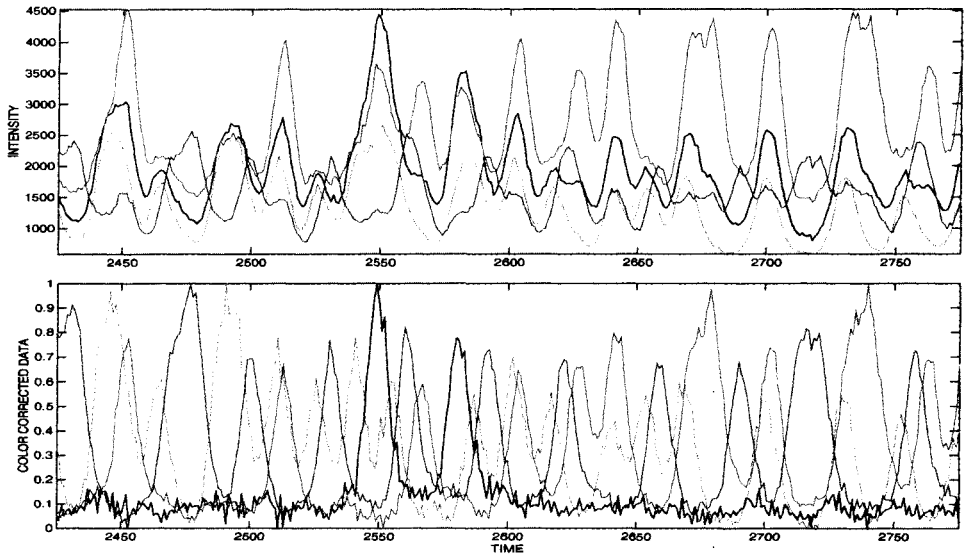


Figure 2: Top: a segment of raw sequencing data from slab gel electrophoresis; Bottom: the color-corrected data, or the estimates of the dye concentrations except for a scale. Color code: A – red, G – black, C – green, and T – blue.

directions are seen along the axes, and almost all other points are in the right-upper quadrant generated by the two cluster directions. We also observe similar patterns in the other five 2-D scatter plots. In fact, such a distributional pattern is determined by the design of Sanger sequencing. Suppose we are in an ideal case by assuming

- **Spectrally non-overlapping hypothesis**: the four dyes are cross-talk free and we observe dye concentrations directly;
- **Spatially non-overlapping hypothesis**: at least in a fairly large range of each trace, the effective mobilities of the four dyes are approximately identical and thus we observe non-overlapping peaks of all four kinds.

In the following, we illustrate how these two hypotheses explain the pattern of scatter plots as shown in Figure 3. For example, we map those observations from non-overlapping C-peaks in Figure 2 (bottom) to points on the cluster directions along the C-concentration axes in the first, second, and third subplots and non-significant points close to the apexes of the fourth, fifth, and sixth in Figure 3. We map those observations from overlapping regions of C-peaks and G-peaks to inner points in the first quadrant, to points on C-concentration axes in the second and third subplots, to points on G-concentration axes in the fourth and fifth subplots, and to non-significant points close to the apex of the sixth quadrant. We map those observations from overlapping regions of more than two kinds of peaks in the same fashion. This key distributional pattern can also be verified empirically using data obtained from a specially designed

experiment. That is, the four differently dye-labeled sub-fragments generated from the four dideoxy reactions are placed into four different yet adjacent lanes of a slab gel. Fluorescence intensities are collected in the same four wavelength bands as those in standard sequencing. This setup uses the same equipment as that in standard sequencing. In this experiment, the four fluorescence intensities obtained from one lane are contributed by only one kind of dye, and their sums are expected to be proportional to the dye concentrations. Here we have ignored the minor baseline issue. The pairwise scatter plots of these "substitutes" of dye concentrations, obtained from one such a cross-talk free experiment, demonstrate the exact pattern in Figure 3; see Figure 1 in [13]. This feature of the distribution provides the basis for our estimation of W and evaluation of color-correction. An appropriate cross-talk matrix is expected to make the distribution of the reconstructed dye concentrations match the pattern shown in Figure 3.

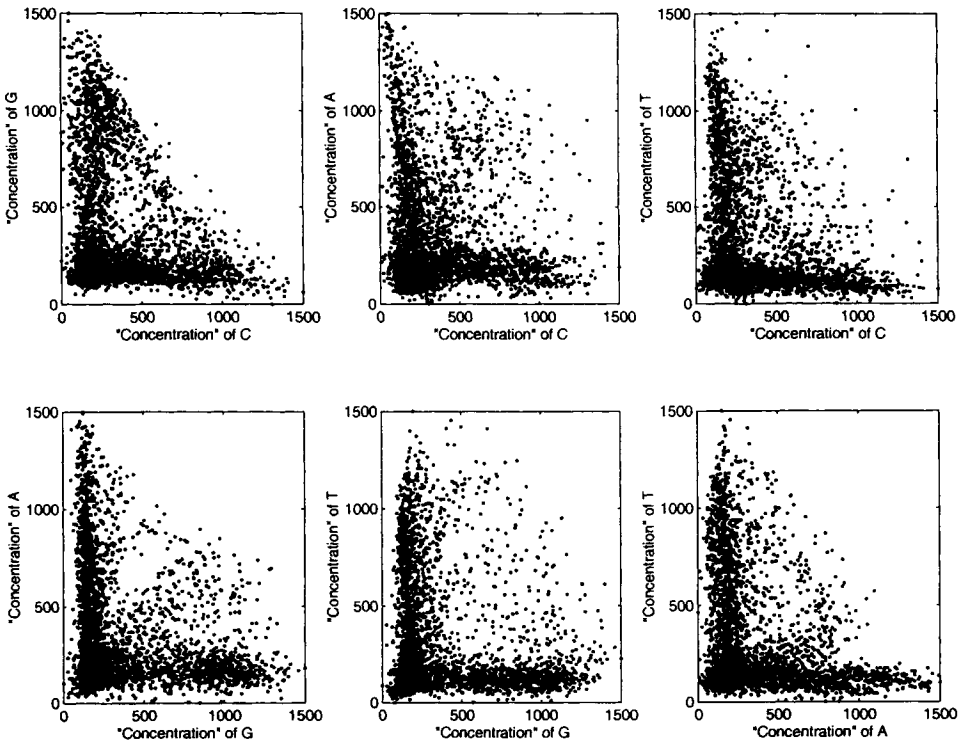


Figure 3: Pairwise scatter plots of the reconstructed dye concentrations.

The non-overlapping hypothesis on distribution of dye concentrations is **sufficient** for estimating the cross-talk matrix. Let us examine the distribution of the raw data – four fluorescence intensities – the output of the system (2). Once again we visualize it with pairwise scatter plots. Figure 4 depicts the six scatter plots of the 3400 observations from one slab gel lane. Let us ignore points in the bottom-left corners of the plots,

which correspond to measurements in valley regions between peaks, or to peaks with low intensities at a pair of wavelength bands; *cf.* Figure 2. Most of the other points lie in a region spanned by two cluster directions – two arms – though they are not as distinct as those in Figure 3. The upper arm of the 3rd, 5th and 6th scatter plots are even more vague because the fourth dye is not as stable as others. If we imagine the complete picture of the distribution in the four-dimensional space, we would find four cluster directions, each corresponding to one column of the cross-talk matrix and almost all other data points lie in the convex cone spanned by them. The pairwise scatter plots are the six 2-dimensional projections of the 4-D scatter plot.

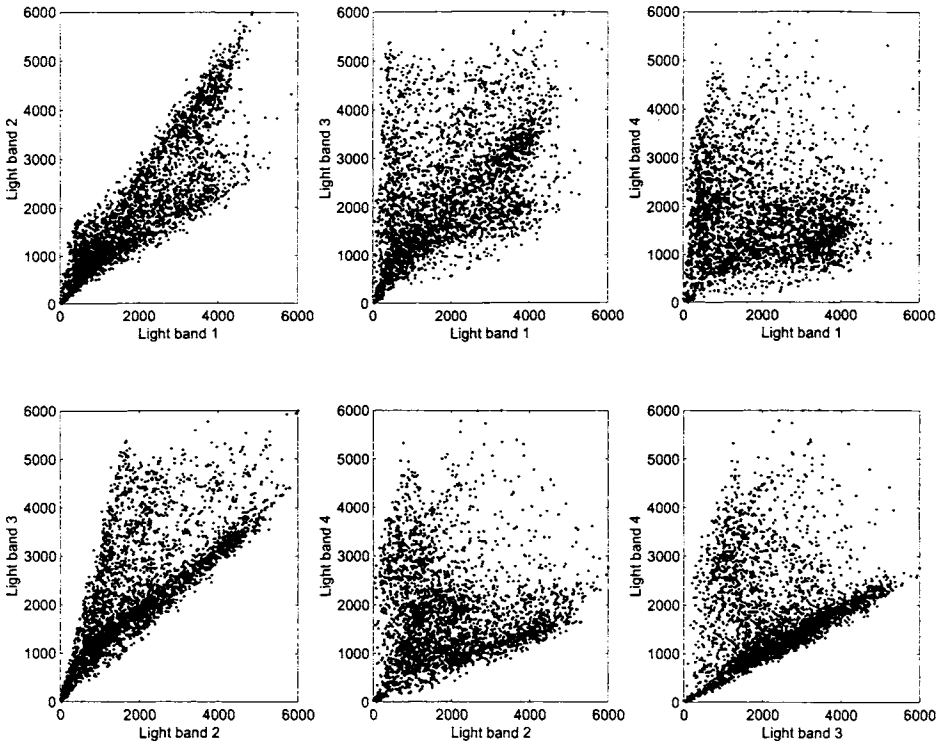


Figure 4: Pairwise scatter plots of the four fluorescence intensities of a slab gel dataset.

The data along each of the 12 arms in the pairwise scatter plot contain the information for estimating one off-diagonal parameter in the cross-talk matrix. Later, we refer to them as "typical points". For example, in the first scatter plot in Figure 4, the slope of the lower boundary should be close to w_{21}, while the slope of the upper boundary should be close to $1/w_{12}$. In other words, the information relevant to the parameter w_{12} can be found in the lower "arm" and that relevant to $1/w_{12}$ can be found in the upper "arm" in the first subplot. Our focus is thus reduced to the 12 slopes. An analog to these "typical points" is the concept of sufficient statistics in statistical modeling. The connection between the data and the parameters leads us to a natural algorithm of esti-

mating the cross-talk matrix: first identify the typical points along the 12 boundaries in the six scatter plots; second, estimate the 12 slopes based on the selected sample points. Technically, we use a **binning** technique to select typical boundary points and a**robust regression** to estimate boundary slopes. These statistical considerations are necessary for handling measurement errors and potential outliers. We notice that some inner points in the 4-D convex cone could possibly mapped to regions close to boundaries when being projected onto 2-D planes, and they would confound with useful information, namely, the typical points. In order to get over this complication, we iterate the above procedure. As iterations are carried out, we expect our estimates to get closer to the target cross-talk matrix. The description of the algorithm is given as follows.

Algorithm 2
(Adaptive color-correction)

0. **Initialization**. *Let $i = 1$. Set the raw data of fluorescence intensities to be the working dataset and the initial estimate $W^{(0)}$ to be the identity matrix. Also set a small positive number α as the threshold of color-correction and a positive integer M as the maximum number of iterations.*

1. **Sampling**. *Consider the first component. It is helpful to look at the first scatter plot consisting of the first and second components in Figure 4.*

   - **Selecting informative range**. *Choose one quantile for the first component. The two bounds of the informative range for w_{21} are defined by this quantile and the largest value of the first component. Those points in the current working dataset with their first components in the range are selected in this iteration for the estimation of w_{21}. For example, if we choose 50%, then, it says we will use those points whose first coordinates are in the upper half; cf. Figure 2, 3, 4.*
   - **Binning**. *Divide the range between these two bounds into bins of the same width.*
   - **Selecting extreme points.** *Among those points whose first component falls into a given bin, find the one having the minimum value in the second component.*

2. **Robust regression**. *Take the points obtained from last step, and run a robust regression of their second components against the first. The estimated slope is taken to be the next estimate of w_{21}. Similarly, the estimate of the slope of the other arm in the same scatter plot is taken to be the next estimate of w_{12}.*

3. **Estimating other parameters**. *Apply the steps similar to 1 and 2 to estimate the other five pairs $\{w_{13}, w_{31}\}$ $\{w_{14}, w_{41}\}$, $\{w_{23}, w_{32}\}$, $\{w_{24}, w_{42}\}$, $\{w_{34}, w_{43}\}$ and assemble them in $\tilde{W}$.*

4. **Checking the color-correction quality**. *Calculate the maximum of the absolute values of the 12 estimated slopes obtained in step 1, 2 and 3. We hereafter refer to this number as cc-number (initials of color correction). If the cc-number is below the threshold α, stop; otherwise, go to step 5.*

5. **Updating**. *Apply the inverse of this matrix to the working dataset (pointwise) and call this the new working dataset. Set $W^{(i)} = W^{(i-1)} * \tilde{W}$ and normalize each column of $W^{(i)}$ to make the diagonal elements unity. Increase i by 1. If $i > M$, stop; otherwise, go back to step 1.*

The algorithm is stopped once we recognize a satisfactory color correction by checking the *cc-number*; see [13] for more details. On exit, $W^{(i)}$ is the estimate of the cross-talk matrix and the working dataset contains the reconstructed dye concentrations. Thus, the procedure does bind the two problems: estimating the cross-talk matrix and color-correcting the measurement of fluorescence intensities. In fact, the dye concentrations in Figure 3 were reconstructed using this algorithm from the fluorescence intensities shown in Figure 4, and the results have been examined. We have experimented with different regression methods in step 2. We observe that the samples obtained in step 1 are not always on the boundaries. Least squares does not work well because of its sensitivity to outliers, and in [13] we proposed the use of a robust procedure – least absolute deviations. Later we adopted the **least trimmed squares** method (LTS) because of its high breakdown point and relatively high efficiency; see Rousseeuw and Leroy [22]. Denote the typical samples obtained from step 1 by $(x_1, y_1), \cdots, (x_s, y_s)$. The least trimmed squares method estimate a straight line with intercept b (an equivalent term to the baseline in (2)) and slope w by

$$\min_{b,w} \sum_{k=1}^{q} |y - b - w \cdot x|^2_{(k)},$$

where $|y - b - w \cdot x|^2_{(k)}$ represents the k-th ordered squared residual, and the sum only takes the smallest q squared residuals into account. We have tested LTS with $q = [n/2] + 1$ and found that five iterations offered a satisfactory solution. The cross-talk matrix used in Figure 3 is obtained in this way. **Least median squares** method (LMS) [21] is another robust procedure and is statistically inefficient with a convergence rate $O(1/\sqrt[3]{N})$ under the normal assumption. On the other hand, algorithms requiring only $O(N^2)$ running time do exist to compute its exact solution in our univariate regression case; see Souvaine and Steele [25]. Another remark is that bins in step 1 do not have to be non-overlapping. However, the bin-width, like the width parameter in kernel smoothing, is the most important and sensitive tuning parameter in this algorithm.

4 Within-slide normalization of gene expression data from microarrays

With the rapid progress of genomic-scale sequencing, complete DNA sequences of some organisms are available, and other genomes can be sequenced in a fairly reasonable time period. Genes on DNA sequences – the blueprint of the life – are the basic biological elements. However, understanding genomic information is much more challenging. A further study of functionalities of genes necessitates the tracking of their dynamic expressions in living organisms. The current method to measure the abundance of mRNA for a specific gene makes use of reverse transcription to its complementary DNA (cDNA), followed by hybridization. The cDNA microarray technique prints thousands of genes on a microscope slide and produces snapshots of gene expression profiles at specific times for specific samples; see Schena, Shalon, Davis, and Brown [24]. A comparison strategy is adopted in cDNA microarray; that is, relative gene expression levels of one sample are measured with respect to a reference. The idea is implemented by a dye technique: label cDNAs from a sample and its reference by two different fluorescent dyes, typically Cy3 (green) and Cy5 (red). Our focus is the difference on the logarithm (base 2) scale of every pair of expression levels corresponding to the same spot on a slide (probe). Let us denote the logarithm of expression levels of the sample and reference at the i-th spot by the pair (U_j, V_j), and denote the logarithm of their measured fluorescence intensities by $(\tilde{U}_j, \tilde{V}_j)$. Ideally, we expect that $(\tilde{U}_j, \tilde{V}_j)=(U_j, V_j)$ except for an offset constant. In practice, non-constant measurement bias occurs because of factors such as physical properties of dyes (heat and light sensitivity, relative half-life), efficiency of dye incorporation, experimental variability in probe coupling and processing procedure, and scanner settings at the data collection step. In order to improve the quality of microarray data, a normalization procedure to adjust the measurement is required.

Sources of variability can be classified into two categories: internal and external with respect to each slide. The effects of external factors are potentially detectable and estimable with multiple-slide data if the experiment is well designed. However, the effects of internal factors are confounded with each slide and thus an adjustment procedure adaptive to each slide is indispensable for the reconstruction of the raw expression levels. Yang, Dudoit, Luu, and Speed [26] proposed an ingenious method – within-slide normalization – to solve the problem. In the following, we have another look at the problem and their normalization procedure from the perspective of BIND. Consider a system with (U_j, V_j) as input and $(\tilde{U}_j$ and $\tilde{V}_j)$ as output. Let $\mathbf{h} = (h_1, h_2)$ be the transformation function; namely,

$$\begin{cases} \tilde{U}_j &= h_1(U_j, V_j) \\ \tilde{V}_j &= h_2(U_j, V_j). \end{cases} \tag{3}$$

The goal is to reconstruct the input variables (U_j, V_j) based on the output variables $(\tilde{U}_j, \tilde{V}_j)$. The system function $\mathbf{h} = (h_1, h_2)$ represents the effect caused by all internal

factors. In fact, we can also include external factors if we have no other better way to estimate them. Obviously this is a blind inversion problem. The BIND scheme leads us to the question: what is the distribution of input, the true expression levels? First, let us suppose that the sample and the reference are identical and that the difference of their expression levels is purely caused by random and uncontrolled effect. In this ideal case, we assume that the random variables $\{(U_j, V_j), j = 1, \cdots\})$ are independent among pairs and within each pair, and they are distributed according to $F(u_j - a_j)$ and $F(v_j - a_j)$, where $F(\cdot)$ is a distribution symmetric about zero and a_j is the average expression level of the j-th gene. If we look at their joint distribution by the scatter plot of U versus V, then we should see that the points cluster around the straight line $V = U$. The average deviation of the points from the straight line measures the precision of the experiment. We denote this joint distribution by Ψ. If the effective measurement system $\mathbf{h}$ is not an identity one, then the distribution of the output, denoted by Ψ', could be different from Ψ. This is exactly what is reported by Dudoit, Yang, Callow and Speed; see Figure 2 in [7].

Next we go back to real practice. Nowadays each slide contains more than a few thousand genes. Suppose that only a small proportion α of the genes are differentially expressed while expressions of the other genes are unchanged except for random fluctuations. Consequently, the distribution of the input in the blind inversion story is a mixture of the two components. One component consists of those unchanged genes, and its scatter plot is similar to Ψ. The other component consists of the differentially expressed genes and is denoted by Γ. Although the cloud shape of Γ in its scatter plot is difficult to find out, its contribution to the input is at most α. The scatter plot of the input variables (U_j, V_j) is a superimposition of those of Ψ and Γ weighted respectively by $1 - \alpha$ and α. We assume that the system function $\mathbf{h}$ is a 1-1 transform. Under $\mathbf{h}$, Ψ and Γ are transformed into distributions denoted respectively by Ψ' and Γ'; that is, $\Psi' = \mathbf{h}(\Psi)$, $\Gamma' = \mathbf{h}(\Gamma)$. This implies that the distribution of the output $(\tilde{U}_j, \tilde{V}_j)$ is $(1-\alpha)\Psi' + \alpha\Gamma'$. If we can separate the two components Ψ' and Γ', then the transform $\mathbf{h}$ of some specific form could be estimated from the knowledge of Ψ and Ψ'. An appropriate estimate $\hat{\mathbf{h}}$ of the transform should satisfy the following: the distribution of $\hat{\mathbf{h}}^{-1}(\Psi')$ is similar to that of Ψ, which centers around the line $V = U$. In other words, the right transform straightens out the distribution cloud of Ψ'. Yang, Dudoit, Luu, and Speed [26] first rotate the coordinate system by 45° as follows,

$$\begin{cases} X &= (U+V)/2 \\ Y &= U-V. \end{cases}$$

After the rotation, the conditional distribution of Y given X should be symmetric about zero. In the scatter plot of (X, Y), the cloud should horizontally center around zero. Each measurement pair $(\tilde{X}_j, \tilde{Y}_j)$ is a transformed version of (X_j, Y_j); we denote $\tilde{X}_j = g_1(X_j, Y_j)$ $\tilde{Y}_j = g_2(X_j, Y_j)$ as in (3). To make the system function estimable, we let $g_1(x,y) = x$ and $g_2(x,y) = g_2(x)$, a free function with some kind of smoothness. Thus the problem becomes a regression problem of $\tilde{Y}$ versus $\tilde{X}$, either in a parametric or in

a nonparametric form. Yang, Dudoit, Luu, and Speed [26] proposed to use **lowess**, a robust local linear regression technique (see Cleveland [6]), to remove the component of Γ' and estimate the transform function $\mathbf{g}$. Once it is estimated, we apply its inverse $\hat{\mathbf{g}}$ to the observations and obtain a reconstruction of the expression difference for each probe. Another **stratification** strategy is adopted in combination with **lowess** smoothing in [26]. That is, the data in one microarray are grouped according to the spatial setup of array printing so that data within each group share a more similar bias pattern. Next, the above normalization procedure is applied to each group; this is referred to as within-print-tip-group normalization in [26]. The above argument provides an interpretation of the within-slide normalization from the perspective of BIND. As a consequence, we see that one justification of the procedure lies in **the hypothesis on the joint distribution of the true gene expression levels of a sample and its reference**.

5 Discussion

5.1 A BIND story in seismology: predictive deconvolution

Various cases of blind deconvolution are reported in the literature; see Li [15] for a recent example and for references. We note that they belong to the class of blind inversion problems, and it is the input distribution that comes to help. We briefly discuss one example, the method of predictive deconvolution used in seismic trace processing, because of its scientific merit. The seismic reflection method aims to determine the distance and directions of remote and inaccessible bodies within the Earth, which is of great importance to oil exploration and other geophysical applications. The basic scheme of the seismic data collection process is the following: active sources of energy such as dynamite, air guns, and chirp signals generators at the surface of the Earth are used to produce waves of some form; the waves propagate downward from the sources into the Earth; at the interfaces between geologic layers in the Earth's crust, part of the waves are transmitted while the rest are reflected; eventually some waves propagate upward to the surface of the Earth and can be detected by receivers located at various distances from the source. The recorded traces of the received waves make up the seismogram. More background can be found in Robinson [17, 18], and Robinson and Durrani [19]. One important problem in seismic data processing can be formulated as follows. Denote the seismic traces by a time series $u(k)$, $k = 1, \cdots, T$. At a stage of the processing (after signature deconvolution), we can postulate a convolution model:

$$u(k) = f(k) * v(k) = \sum_{i=0}^{+\infty} f(k-i)v(i), \tag{4}$$

where $f(k)$ is the reverberation waveform and $v(k)$ is the reflectivity function or the reflectivity coefficient at each layer. Let $U(z)$, $F(z)$, and $v(z)$ be the z-transforms of $u(k)$, $f(k)$, and $v(k)$, respectively. Then we have $U(z) = F(z)V(z)$. We can regard Equation (4) as a linear system, in which $v(k)$ and $u(k)$ respectively play the role of unknown

input and known output. According to the **feedback hypothesis**, the reverberation filter $f(k)$, which characterizes the system, is not only causal but also minimum delay. Specifically, the feedback hypothesis assumes that the reverberation effect takes a form of finite feedback filter, namely,

$$F(z) = \frac{1}{1 + \alpha_1 z + \alpha_2 z^2 + \cdots + \alpha_p z^p},$$

whose zeros are outside of the unit circle on the complex plane. The reverberation refers to the fact that waves are successively reflected between two interfaces of a layer. In practice, reverberations are often generated by such a complicated physical situation with many layers that the effective filter is impossible to be obtained by direct measurement. Thus in order to unravel the seismic traces, we need to estimate both the system filter $F(z)$ and the input – the reflection coefficient. It is clear that this is a blind inversion problem. According to the scheme of BIND, we first inquire for the distribution of the reflection coefficients. Fortunately, several studies showed that the **random hypothesis** is approximately valid in many cases; see Robinson [17], page 278 and the references mentioned there. The **random hypothesis** assumes that the reflection coefficients $v(k)$ follow a white noise stochastic process. That is, $E[v(k)] = 0$, $E[v(k)v(j)] = E[v(k)]E[v(j)] = \sigma^2\delta_{j,k}$, if $k \neq j$. As a matter of fact, the second order statistical property of a stationary stochastic process is characterized by its autocorrelation coefficients. If we further assume that $v(k)$ is normal, then the distribution of the input is uniquely determined because the higher-order cumulants of a normal distribution are zero; see Rosenblatt [20]. We note that the normal assumption is not required by the algorithm of predictive deconvolution. A consequence of the **random hypothesis** is that the output, seismic traces, is a zero mean stationary stochastic process whose second order statistical property is characterized by its autocorrelations, denoted by $\gamma_{uu}(k)$, $k = 0, 1, \cdots$. This set of statistical correlations relates to the reverberation filter through the Yule-Walker equation:

$$\begin{pmatrix} \gamma_{uu}(0) & \gamma_{uu}(1) & \cdots & \gamma_{uu}(k-1) \\ \gamma_{uu}(1) & \gamma_{uu}(0) & \cdots & \gamma_{uu}(k-2) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{uu}(k-1) & \gamma_{uu}(k-2) & \cdots & \gamma_{uu}(0) \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_k \end{pmatrix} = \begin{pmatrix} \gamma_{uu}(1) \\ \gamma_{uu}(2) \\ \vdots \\ \gamma_{uu}(k) \end{pmatrix}. \quad (5)$$

In practice, we plug into this equation estimates of $\gamma_{uu}(k)$ based on data and apply the Levinson-Durbin algorithm [8] to obtain an estimate of the filter denoted by $\hat{\alpha}_k$. The Burg algorithm alternatively estimates the filter directly from the data. The determination of the order p in the filter is a problem of model selection and the technique of AIC or BIC can be applied. With the estimated filter and starting values $u(1), \cdots, u(q)$ obtained from seismic and numerical considerations, we reconstruct the reflection coefficients using the feedback procedure

$$v(k) = u(k) + \sum_{i=1}^{p} \hat{\alpha}_i u(k-i).$$

This is the so-called predictive deconvolution. Except for the specific time series techniques involved, it is consistent with the BIND philosophy as used in previous examples.

5.2 Statistical assessment

5.2.1 Goodness of match with the input distribution

To assess a BIND procedure, we examine the distribution of the reconstructed input and see if it matches the hypothesis. In the DNA sequencing example, we check the *cc-number* achieved at the exit of Algorithm 2 and recognize a satisfactory color correction if it is below a threshold. A graphical check of the scatter plots of the color-corrected data like Figure 3 might be expected in some cases. For cDNA microarray example, we can similarly define a quality number as the maximum absolute value of the regression line obtained by **lowess**, in which the parameters are chosen as those in [26]. The graphical check could be sufficient for many biologists. For the illustrative example (1), we test the independence of the four components; this can be carried out by the likelihood ratio test or other tests, see Chapter 9, Anderson [2]. For the seismic trace example, we check the whiteness of the reconstructed reflection coefficients by examining the flatness of its periodogram or using other statistical tests; see Brockwell and Davis [4].

5.2.2 System sensitivity analysis by data self-perturbation

The reconstruction of input in the BIND scheme is associated with the problem of system estimation. As in any other estimation problem, it is valuable to assess the accuracy of the estimates of the system. One technique in this regard is the **bootstrap**; see Efron and Tibshirani [9]. In the DNA sequencing example, we can generate bootstrap samples by sampling from the raw dataset with replacement and applying the same color-correction algorithm (with the same set of parameters) to each bootstrap sample. The bias and standard deviation of the bootstrap estimates reflect systematic bias and variability of the algorithm with respect to data self-perturbation. It is possible that these statistics contain some scientific meaning and could provide some guidance for researchers. For the DNA sequencing example, we show the result of a bootstrap study with 200 replicates in Table 1. It includes the bias, standard error, and coefficient of variation for each of the 16 parameter estimates – here we switch from the parameterization of the cross-talk matrix in (2) to the one whose columns sum to unity. The estimates are almost unbiased for all the four dyes. The SDs and CVs measure the stability of the four dyes. For example, the comparatively larger SDs and CVs of the estimates regarding the fourth dye associated with T indicate that its physical and chemical properties are not as stable as others. Our collaborator Dr. Kheterpal (she was with Prof. Mathies' group in the Chemistry Department at UC, Berkeley during our collaboration) verified this observation. This sensitivity analysis by bootstrap applies

to other parametric systems such as the example in (1). However, the technique of data self-perturbation needs special care for systems with a nonparametric form, such as that in the microarray example and for systems with a spatial structure, such as that in the seismology example.

Table 1: Accuracy assessment of the estimate by bootstrap ($\times 10^{-3}$)

	C				G			
	mean	bias	SD	CV	mean	bias	SD	CV
1	333	-12	14	42	203	-1	3	15
2	330	0	7	21	412	1	4	10
3	241	9	10	42	296	0	4	13
4	96	2	6	63	88	0	4	46
	A				T			
	mean	bias	SD	CV	mean	bias	SD	CV
1	70	0	10	143	115	-3	11	96
2	209	5	12	57	139	4	26	187
3	544	-3	16	30	183	-4	17	93
4	176	-2	6	34	562	3	50	89

5.3 Final remarks

Among the many things I learned from Prof. Terry Speed through the years of my study under his supervision, the one that impressed me very much was his conscientious service to the scientific community, especially in genetics and molecular biology, as a statistician. His broad and dynamically-changing research interests are phenomenal. His extensive collaborations with biologists constantly bring research life-blood into his students' study. We also notice that he earned so much respect not only from his statistician colleagues but also from researchers in other fields.

This article is motivated by Terry's advocacy of **considering scientific meanings in mathematical and statistical modeling**. The abstraction of BIND provides a way to think about a scientific problem mathematically as well as a way to think about mathematics scientifically. The BIND scheme hinges on a hypothesis on the distribution of system input. The verification of this hypothesis requires careful consideration, and it varies from one problem to another. No matter how BIND is implemented, either by an algorithm from numerical recipes or by a novel procedure, the bottom line is: the distribution of the reconstructed output should match the hypothesis. It is our hope that the BIND notion can help statisticians apply their toolbox to more scientific measurement problems in the future.

Acknowledgments

Prof. Terence P. Speed and Dr. David O. Nelson provided the author with great help in this work. The research is supported by the NSF grant DMS–9971698, and DOE grant DE–FG03–97ER62387. The author would also like to acknowledge help provided by the Institute of Pure and Applied Mathematics, UCLA.

Lei Li, Molecular and Computational Biology, University of Southern California, 1042 West 36th Place, DRB 289, Los Angeles, CA 90089-1113, lilei@hto.usc.edu

References

[1] M. D. Adams, C. Fields, and J. C. Ventor, editors. *Automated DNA sequencing and analysis*. Academic Press, London, San Diego, 1994.

[2] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. John Wiley & Sons, 2nd edition, 1984.

[3] P. Billingsley. *Probability and Measure*. John Wiley & Sons, 1986.

[4] P. J. Brockwell and R. A. Davis. *Time Series: Theory and Models*. Springer-Verlag, 1991.

[5] S. E. Cawley. *Statistical models for DNA sequencing and analysis*. PhD thesis, University of California, Berkeley, 2000.

[6] W. S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74:829–836, 1979.

[7] S. Dudoit, Y. H. Yang, M. J. Callow, and T. P. Speed. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, 12:111–140, 2002.

[8] J. Durbin. The fitting of time series models. *Rev. Inst. Internat. Statist.*, 28:233–244, 1960.

[9] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall: New York, London, 1993.

[10] G. R. Fowles. *Introduction to Modern Optics*. Halt, Rinehart And Winston, Inc., 2nd edition, 1975.

[11] G. H. Golub and C. F. Van Loan. *Matrix Computations*. The John Hopkins University Press, Baltimore and London, 1996.

[12] L. Li. *Statistical Models of DNA Base-calling*. PhD thesis, University of California, Berkeley, 1998.

[13] L. Li and T. P. Speed. An estimate of the color separation matrix in four-dye fluorescence-based DNA sequencing. *Electrophoresis*, 20:1433–1442, 1999.

[14] L. Li and T. P. Speed. Parametric deconvolution of positive spike trains. *Annals of Statistics*, 28:1279–1301, 2000.

[15] T. H. Li. Blind deconvolution of linear system with multilevel nonstationary input. *Annals of Statistics*, 23:690–704, 1995.

[16] D. O. Nelson and T. P. Speed. Recovering DNA sequences from electrophoresis. In S. E. Levinson and L. Shepp, editors, *Image Models (and their Speech Model Cousins*, pages 141–152. Springer-Verlag, New York, 1996.

[17] E. A. Robinson. *Physical Applications of Stationary Time-Series*. Macmillan Publishing, 1980.

[18] E. A. Robinson. *Seismic Inversion and Deconvolution, Part A: Classical Methods*. Geophysical press, 1984.

[19] E. A. Robinson and T. S. Durrani. *Geophysical Signal Processing*. Prentice Hall, 1986.

[20] M. Rosenblatt. *Stationary Sequences and Random fields*. Birhhäuser, 1985.

[21] P. J. Rousseeuw. Least median of squares regression. *Journal of the American Statistical Association*, 79:871–881, 1984.

[22] P. J. Rousseeuw and A. M. Leroy. *Robust Regression and Outlier Detection*. Wiley, New York, 1987.

[23] P. J. Russell. *Genetics*. Harpercollins College Publisher, New York, 1995.

[24] M. Schena, D. Shalon, R. M. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270:467–470, 1995.

[25] D. L. Souvaine and J. M. Steele. Time- and space-efficient algorithms for least median of squares regression. *Journal of the American Statistical Association*, 82:794–801, 1987.

[26] Y. H. Yang, S. Dudoit, P. Luu, and T. P. Speed. Normalization for cDNA microarray data. In M. L. Bittner, Y. Chen, A. N. Dorsel, and E. R. Dougherty, editors, *Microarrays: Optical Technologies and Informatics*, Proceedings of SPIE. 2001.

Designing Meaningful Measures of Read Length for Data Produced by DNA Sequencers

David O. Nelson and Jane Fridlyand

Abstract

Nearly everyone uses "the number of Q20 bases" as a rough measure of the effective length of a given DNA sequence produced by the base-caller PHRED. This metric simply counts the number of bases in a read in which the PHRED quality score is at least 20. While the number of Q20 bases is a simple, easy to implement rule-of-thumb, it does not have much else going for it: it consistently underestimates the number of usable bases in the read. In this short paper, we develop and evaluate an alternative metric that uses more of the PHRED quality data in a read to predict how many bases from that read would make it into the eventual consensus sequence of an assembly. The metric was developed by evaluating a set of pre-existing, high-quality assembled contigs. The resulting predictor is a simple function of the histogram of PHRED quality values already produced by sequencing software and performs nearly as well as a more complex additive model that uses regression splines.

Keywords: DNA read length; genomics; predicting progress; PHRED; PHRAP

1 Introduction

Large-scale genome sequencing projects have become increasingly common over the last fifteen years. Many recent papers, starting with Lander and Waterman in 1988 [6], have described mathematical models for predicting the progress of such sequencing projects. These different "Lander-Waterman" analyses arise in response to different approaches to sequencing large genomes. They model the sequencing process as a coverage process like those described by Hall [5] and derive predictions of mean coverage, depth, expected number of gaps, and the like, as a function of the number of clones sequenced N, the genome size G, and the length of sequence L obtained from an individual clone chosen for sequencing. These predictions are then used to estimate the number of clones required to obtain an assembled genome to a given depth or coverage. Conversely, statistics on coverage and read length gathered during the sequencing effort are used with these models to track progress, detect problems, and refine estimates of the remaining work required.

The approximate genome size G can be determined in advance, and number of clones sequenced N is easy to obtain from daily production statistics. However, what

about L? The average number of bases sequenced from the end of a clone can be tuned by changing electrophoresis conditions, run-time, and the sequencer chosen to sequence the DNA. In fact, deciding on the desired length L is a major factor in planning sequencing projects, along with the size (or sizes) of sequencing clone and whether or not both ends are to be sequenced.

On one hand, most Lander-Waterman analyses assume L to be a constant (although Lander and Waterman do provide some guidance on the degradation in performance due to random clone sizes) and assume that any overlap between two clones is detected with probability one for overlaps of a certain size or greater. On the other hand, sequencing centers that use the most popular combination of base-caller and assembler, PHRED [4, 3] and PHRAP [7], are faced with a much more complex situation with respect to read length and overlap detection. PHRED can produce *extremely* long reads, but also throttles the process somewhat by providing a probability of error with each base read. This base-specific probability of error is expressed as a "quality value" for each base i: $q_i = -10\log_{10} p_i$, where p_i is (more or less) the probability that base i is called in error. PHRED produces integer quality values ranging from zero to approximately fifty, and those associated with bases at the ends of the read are typically much lower than those in the middle. PHRAP uses these quality values in the assembly process in a complex way. A byproduct of an assembly is a "trimmed" read for each read that entered the assembly, in which some number of bases at the start and end of each read are discarded during the alignment process.

Most sequencing centers finesse the problem of estimating L for a read by the simple expedient of counting the number of bases in a read for which $q_i \geqslant 20$. This Q_{20} rule arose during the initial phases of the Human Genome Project and was adopted by the public consortium as a common measure of read length. However, for planning future projects, it would be desirable to derive a better measure of read length, and preferably one that related to some measure of the useful size of a read.

In this paper, we define the "effective read length" of a read in an assembly as the length of the trimmed read produced by PHRAP. We believe that this definition of effective read length provides a more reasonable model for L in Lander-Waterman analyses of projects that use PHRED and PHRAP as a base-caller and assembler. In this paper, we explore some of the features of this distribution and build predictors of L as a function of the set of q_i. Our goal is to provide a simple algorithm to estimate L that is more accurate and precise than the Q_{20} rule currently in place.

2 Methods

All analyses were done using the statistical computing environment *R* [8].

Source of Reads

We analyzed assemblies from fifty-one sequencing projects produced by the Joint Genome Institute (JGI) in Walnut Creek, California during two time periods spanning the 1998–2002. Forty-eight of the sequencing projects were cosmids completed during the period of November 1998–April 1999. These projects were sequenced on ABI 377 slab gel sequencers [2]. Three of the sequencing projects were bacterial artificial chromosomes (BAC's) completed during the period June 2002–August 2002. These projects were sequenced on a combination of Molecular Dynamics Megabace 1000 [1] and ABI 3700 class capillary sequencers.

All projects were base-called and assembled using the current versions of PHRED and PHRAP with default parameters. From each project, we selected only those contigs which contained at least 300 reads and had coverage between 5 and 60. Five of the projects had two such contigs; the rest had just one "main" contig.

Data Gathering

We excluded reads that contained vector, as they would need a special treatment in order to remove the vector sequence and calculate effective read length. In addition, we excluded those reads that had ends extending outside the trimmed part of the final contig. We obtained statistics on each read from the output files produced by PHRED, as well as the standard output file produced by PHRAP. Data obtained for each read included the length of the untrimmed read; the length of the trimmed read; the insertion, deletion, and substitution error rates in the trimmed part of the read; the expected number of correct bases in the read, defined as $n - \sum_i 10^{-q_i/10}$, where n is the number of bases read and the q_i are the corresponding quality values; the number of bases in the read with PHRED quality values in five different histogram bins (0–9, 10–19, 20–29, 30–39, 40 and above); and the expected number of correct bases in each of those histogram bins.

3 Results

Distribution of Percent Trimmed

Table 1 summarizes, by quintile of average depth, the characteristics of the 52,097 reads from fifty-one contigs obtained from the forty-eight slab gel projects. Each row of the table shows summary statistics for one of five quintiles of depth of coverage in the slab data set. Summary statistics for each quintile include the number of contigs, the median number of reads in the contigs, and the median length of the contigs. Cosmids are around 40,000 bases long, approximately the same size as the median size of each contig in all five depth quintile.

Table 2 describes the characteristics of the 13,539 reads from five BAC contigs obtained from the three capillary electrophoresis projects. BAC's are considerably longer

Table 1: Characteristics of 51 cosmid contigs by depth quintile

Quintile	Depth	Number of Contigs	Median Number of Reads	Median Length of Contig
1	[12,14]	14	772	43,458
2	(14,16]	9	882	39,670
3	(16,18]	10	945	41,473
4	(18,21]	8	1,064	38,790
5	(21,46]	10	1,326	39,890

Table 2: Characteristics of five BAC contigs sequenced by capillary electrophoresis

Project	Contig	Depth	Number of Reads	Contig Length
THW	I	37	3,934	177,944
TKM	I	50	1,967	67,818
	II	50	2,729	96,546
TKP	I	39	1,012	40,502
	I	45	3,897	142,666

than cosmids, ranging from 150,000 to 200,000 bases in length. Adding up the contig sizes, we see that the selected contigs represent approximately the length of their respective clones, and are all sequenced to high depth.

Figure 1 shows the relationship between raw and trimmed read length as a function of sequencing technology and quartile of raw read length within sequencing technology. Note that the read length quartile values differ in slab and capillary technologies, largely because of the superior read length obtained with current capillary machines. The quartile bins of raw read length for slab reads were 107–607, 608–657, 658–832, and 833–2149. For capillary reads, the quartile bins were 187–795, 796–1114, 1115-1213, and 1214–1578. There is considerable similarity between the distributions of proportion trimmed, as a function of read length. The longer the raw read is, the larger the proportion trimmed. Hence, the number of bases trimmed goes up dramatically as the raw read length increases. The larger spread of proportion trimmed in the highest quartile of slab gel reads is likely due to the extremely large size of the bin (833–2149, versus 1214–1578 for the capillary reads). Overall, the median percentage trimmed was slightly over seven percent, approximately twenty-five percent of the reads had less than five percent trimmed, seventy-five percent of the reads had less than fifteen percent of the raw read length trimmed, and 456 reads had over 90 percent trimmed.

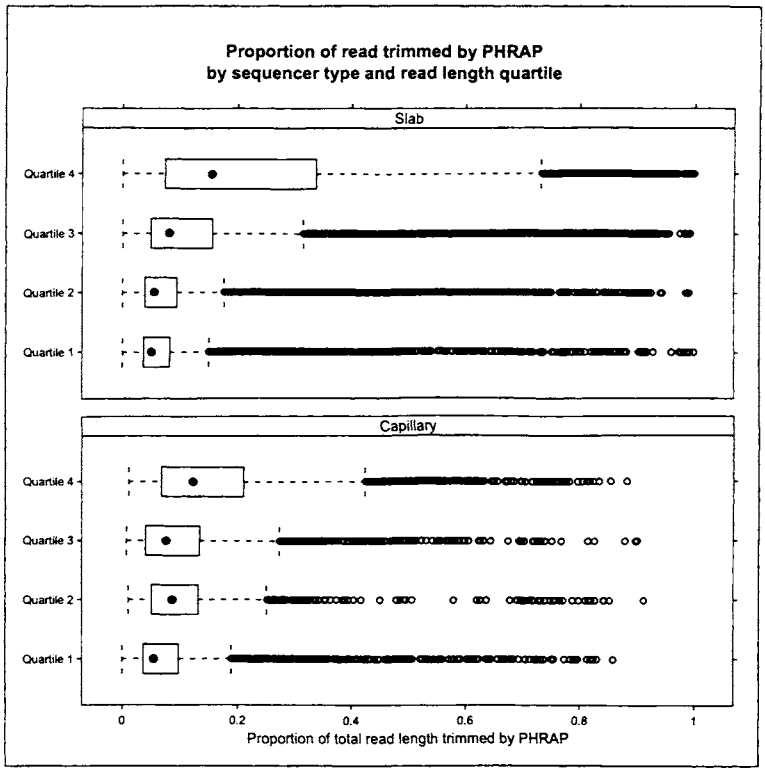


Figure 1: Boxplots showing the distribution of of the proportion of the read trimmed by PHRAP, as a function of the type of sequencer (slab gel vs. capillary) used and the quartile of raw read length. Within each panel, four boxplots are shown: one for each quartile of raw read length. The top panel shows the distribution for the reads in the data set that were produced by a slab gel sequencer, while the bottom panel shows the distribution for reads in the data set that were produced by capillary sequencer. Note that the average proportion trimmed increases with increased read length, but is relatively stable across sequencing technologies.

Current Measures of Effective Read Length

We now examine how well two common measures of read length predict the actual number of bases used by PHRAP: the Q_{20} rule and the "expected correct", or E_c rule. First, we examine the Q_{20} rule.

Recall that the Q_{20} rule simply counts the number of bases with a PHRED quality score of 20 or more. Figure 2 shows a scatter plot of the relationship between Q_{20} and the number of bases actually used. This plot shows clearly the extent to which Q_{20} dramatically underestimates the actual number of bases used by PHRAP. Superimposed on the scatter plot is a scatter plot smoother fit and a line dividing the region where Q_{20} overestimates from the region where it underestimates. Note the almost total lack of

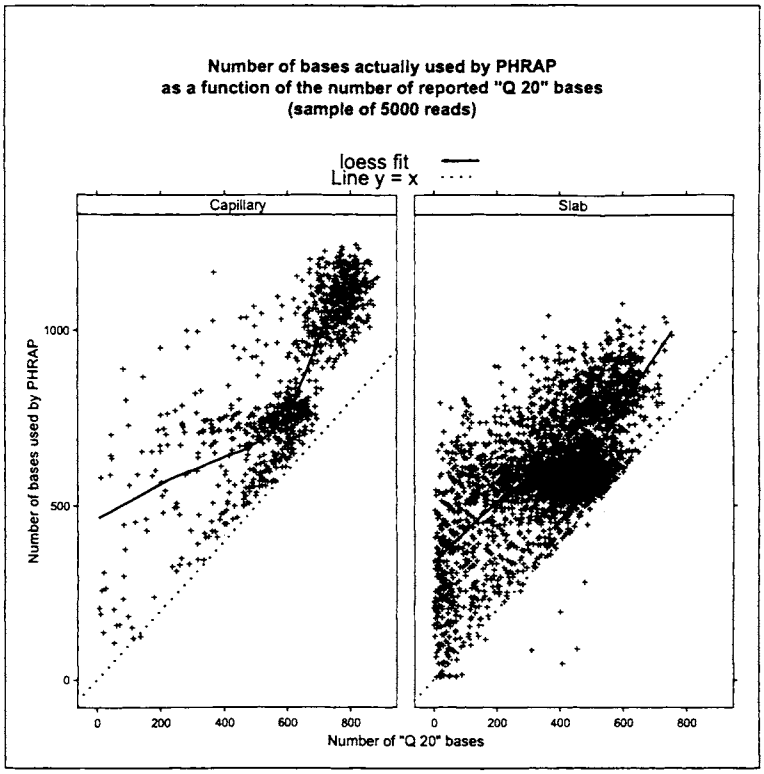


Figure 2: The relationship between the number of "Q 20" bases in a read and the actual number of bases used in the assembly. The Q_{20} rule almost universally underestimates the number of usable bases in a read, irrespective of sequencing technology. Note that for ease in graphing, only 5000 randomly sampled points are plotted.

points where Q_{20} underestimates the read length. This result is not too surprising, as PHRAP goes to great lengths to use the bases at the ends of the read, where the quality scores are typically low. In addition, the graph does indicate that some other metric that uses more of the information in the PHRED histogram cannot help but improve the performance of a read length predictor.

A second obvious choice for read length estimator is the expected number of correct bases, which can be written as

$$E_c = n - \sum_{i=1}^{n} 10^{-q_i/10},$$

where n is the number of bases in the untrimmed read. This estimator subtracts off a read-specific constant from the untrimmed read length in an attempt to estimate the number of trimmed bases.

Figure 3 shows a scatter plot much like Figure 2, only plotting the number of bases

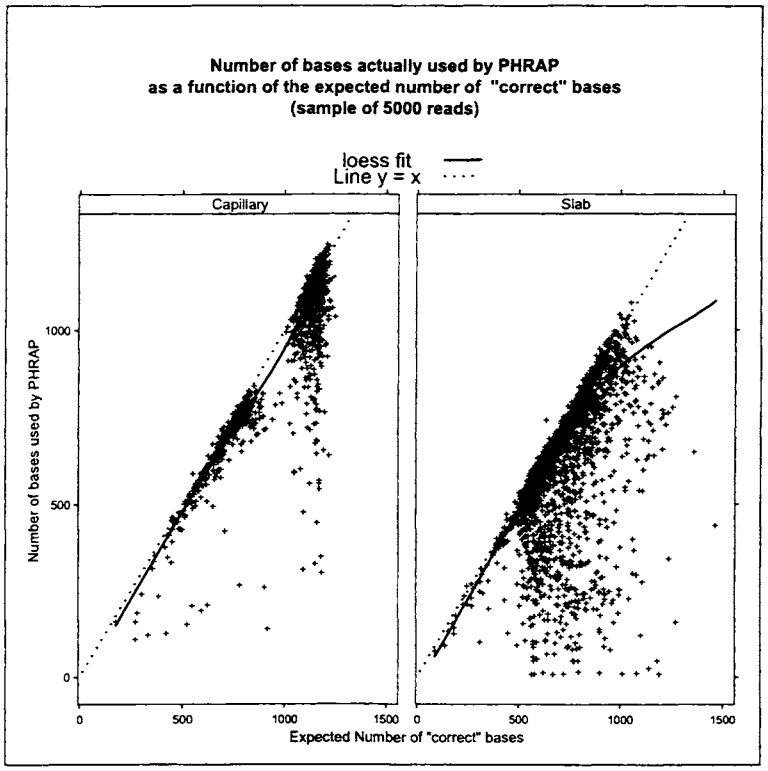


Figure 3: The relationship between the expected number of correct bases, as estimated by PHRED quality scores, and the actual number of bases used in the assembly. Note that the "expected number" rule performs better, but often overestimates the number of usable bases in a read.

used against E_c. Here we see the opposite effect of that observed with the Q_{20} rule: E_c tends to *overestimate* the number of bases used by PHRAP. Despite that overestimation, the tight clustering of points around the line $y = x$ indicates that this estimator is clearly superior to the Q_{20} rule, especially for the capillary reads. Perhaps some combination of the two estimators should be considered. We now examine that possibility.

Additive Combinations of Histogram Values

We now consider a simple generalization of the above two estimators. The Q_{20} rule can be written as a simple affine combination of the histogram counts produced as a byproduct of the sequencing process flow at the JGI:

$$Q_{20} = w_0 + w_1 N_1 + w_{10} N_{10} + w_{20} N_{20} + w_{30} N_{30} + w_{40} N_{40},$$

where $N_1, N_{10}, N_{20}, N_{30}$, and N_{40} are the number of bases in the read with PHRED quality values in $[0,9)$, $[10,19)$, $[20,29)$, $[30,39)$, and $[40,50]$, respectively, and w_0, w_1, w_{10},

w_{20}, w_{30}, and w_{40} are weights. For the Q_{20} rule, $w_0 = w_1 = w_{10} = 0$, and $w_{20} = w_{30} = w_{40} = 1$. The E_c estimator can also be easily approximated by linear combination of these histogram counts, with the weights w_j approximating error probabilities for bases in histogram bin j. We will now generalize to additive combinations of smooth functions of histogram counts as predictors of read length.

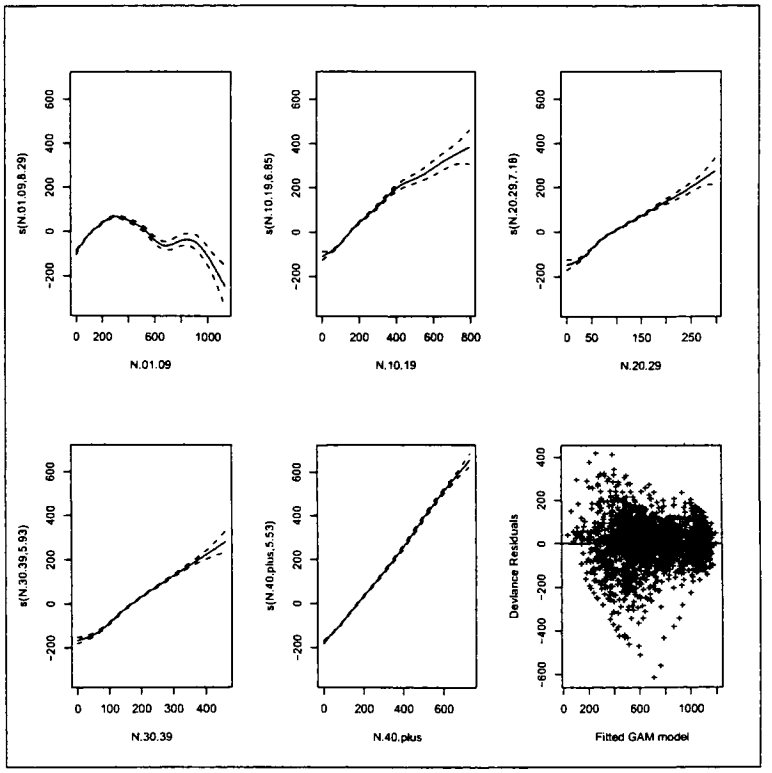


Figure 4: The five smooth terms in a GAM estimate of the number of bases used by PHRAP, along with a plot of the deviance residuals versus fit. The five terms correspond to smooth functions of the number of bases in each of the five histogram bins described in **Methods**. In addition, a two-factor term adjusting for sequencing technology was also included. The labels "N.01.09" through "N.40.plus" correspond to the five histogram bins of PHRED values produced by local sequencing software (0–9, 10–19, 20–29, 30–39, and 40 or more). Each y axis label is of the form $s(x,d)$, where $s()$ is an estimated smoothing spline, x is the count in the corresponding histogram bin, and d is the approximate degrees of freedom in the estimated spline. Only the 5000 random points plotted in Figures 2 and 3 were used in the fit.

Figure 4 shows the result of fitting a generalized additive model to the read length data described above. The model fit was of the form

$$y_i = \alpha_0 + \alpha_1 x_i + \sum_j f_j(n_{ij}) + \varepsilon_i$$

where y_i is the number of bases used by PHRAP for read i, x_i is an indicator variable that is one whenever the read was performed on a slab instrument, n_{ij} is the number of bases in histogram bin j for read i, the f_j are penalized regression splines [9] with knots to be determined by cross-validation, and ε_i is Gaussian error. Five of the panels in Figure 4 show estimates of the f_j, along with standard errors, while the lower right panel shows a plot of deviance residuals versus fitted values. All of the terms in the model were highly significant, and the adjusted R^2 of the fit was 0.88.

From Figure 4 we see that, except for the histogram bin corresponding to bases with $q < 10$, the resulting splines are reasonably linear. The spline for the $q < 10$ bin, on the other hand, looks more complex. The conclusion we draw is that more detail about the structure of the $q < 10$ quality values will be needed to make a substantial improvement. However, as a first approximation, the spline looks like it might be adequately approximated by a quadratic.

Consequently, we refit the data to a linear model with a quadratic term for the $q < 10$ histogram term:

$$y_i = \alpha_0 + \alpha_1 x_i + \beta_0 \frac{n_{i0}}{100} + \beta_1 \left(\frac{n_{i0}}{100}\right)^2 + \sum_{j>0} \gamma_j n_{ij} + \varepsilon_i \tag{1}$$

(In order to keep the quadratic coefficient to a reasonable size, we scaled the $q < 10$ histogram value by dividing by 100). The results of the fit are shown in Table 3. We

Table 3: Results of linear model (quadratic term for N.01.09)

Term	Estimate	S.E.	t Ratio	$\Pr(>\|t\|)$
Intercept	-101.90	7.58	-13.45	$< 10^{-15}$
Slab	35.00	4.36	8.03	$< 10^{-14}$
N.01.09/100	79.00	2.19	36.07	$< 10^{-15}$
$(N.01.09/100)^2$	-9.59	0.29	-32.98	$< 10^{-15}$
N.10.19	0.79	0.02	40.04	$< 10^{-15}$
N.20.29	1.57	0.03	51.83	$< 10^{-15}$
N.30.39	1.01	0.02	61.52	$< 10^{-15}$
N.40.plus	1.15	0.01	107.61	$< 10^{-15}$
Residual S.D.	76			
Adjusted R^2	0.87			

see that the linear coefficients for bases with $q < 20$ are around 0.79, while the bases with $q \geqslant 20$ have coefficients somewhat above one. We also see that a large number of bases with low quality decreases the effective read length.

The fit in Table 3 was based on a training of 5000 points. In order to evaluate the prediction error, we examined the distribution of the absolute value of the difference

between the predicted number of bases and actual number of bases on a test set consisting of the other 60,636 reads in the data set. Table 4 summarizes that distribution, and compares it with the prediction error for four other estimators: the generalized additive model described above, a linear model without a quadratic term for the $q < 10$ bases, E_c, and Q_{20}.

Table 4: Absolute prediction error quantiles for estimators of effective read length

	Quantile of Prediction Error				
Model	25%	50%	75%	95%	99%
Additive Model	10	24	52	159	298
Linear Model					
(with quadratic term)	12	26	56	161	302
(no quadratic term)	19	39	71	173	323
E_c	10	22	62	365	703
Q_{20}	115	201	309	462	594

We see that, except for extremely large errors, the Q_{20} estimator is dominated by each of the other estimators analyzed. We see that the E_c estimator is quite competitive with the linear model, at least until the read length gets extremely large.

4 Conclusions

We can draw several conclusions from the above analyses. First, the Q_{20} predictor grossly underestimates the effective length of a sequencing read. Except for the extreme cases, all of the other predictors discussed dominate it under all circumstances in which they were compared. Second, the E_c and the linear model predictors have comparable prediction error: on average, about sixty bases. However, the E_c estimator has two disadvantages when compared to the estimators derived from a linear model. First, the errors for E_c appear to be biased: on average, E_c overestimates the read length. Second, E_c requires the entire set of PHRED quality scores. If we restrict our attention to estimators based on histograms only, Table 4 shows that the best estimator based only on a linear combination of histogram values is dominated by the linear model with an added quadratic term, as expected. Finally, we note that the appropriate simple linear combination of PHRED histogram bins is quite competitive with the much more complex generalized additive model. The main benefit of adding a quadratic term seems to be to decrease prediction error in the extreme case of a long, low-quality read.

The boxplots in Figure 1 show a considerable amount of skewness in the distribution of percent trimmed. Although this skewness is transmitted somewhat to the number of bases in the consensus, log-transforming the outcome y_i in Equation 1 does not improve the prediction error at all.

In these analyses, we have not explored any effects due to mis-called bases. Other statistics gathered for these analyses include the percent of indels (insertions/deletions) and substitution errors in the trimmed read. Our analysis (not shown) indicates that this component of effective read length is small (under a few percent), and is dominated by PHRAP's trimming process.

It seems clear that the relationship between the PHRED quality values and the size of the region PHRAP trims from the raw read is both simple and quite complex. It is simple in the sense that, in most cases, the expected number of bases E_c closely matches what PHRAP uses. However, an examination of Figure 3 shows that as the raw read gets longer, the situation becomes quite complex, and the size of the region trimmed becomes more of a function of serial correlations between quality values. This situation is exactly what various kinds of "moving window" trimming algorithms try to capture. It would be interesting to explore the extent to which statistically-based moving window algorithms might outperform the marginal approach outlined above in the situation of long, low-quality reads.

Acknowledgments

This work was performed under the auspices of the U. S. Department of Energy by the University of California, Lawrence Livermore National Laboratory under contract number W–7405–ENG–48.

David O. Nelson, Joint Genome Institute, Lawrence Livermore National Laboratory,
daven@llnl.gov

Jane Fridlyand, Comprehensive Cancer Center, University of California, San Francisco,
janef@cc.ucsf.edu

References

[1] Amersham Biosciences. Megabace web site. http://www.megabace.com.

[2] Applied Biosystems. http://www.appliedbiosystems.com/products.

[3] Brent Ewing and Phil Green. Basecalling of automated sequencer traces using Phred. II. Error probabilities. *Genome Research*, 8:186–194, 1998.

[4] Brent Ewing, LaDeana Hillier, Michael C. Wendt, and Phil Green. Basecalling of automated sequencer traces using Phred. I. Accuracy assessment. *Genome Research*, 8:175–185, 1998.

[5] Peter Hall. *Introduction to the Theory of Coverage Processes*. John Wiley and Sons, 1988.

[6] Eric S. Lander and Michael S. Waterman. Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics*, 2:231–239, 1988.

[7] University of Washington. Phrap web site. `http://www.phrap.org`.

[8] R Project. R web site. `http://www.r-project.org`.

[9] Simon N. Wood. Modelling and smoothing parameter estimation with multiple quadratic penalties. *Journal of the Royal Statistical Society B*, 62(4):413–428, 2000.

Extensions to a Score Test for Genetic Linkage with Identity by Descent Data

Sandrine Dudoit and Darlene R. Goldstein

Abstract

Genetic analysis aims to determine which underlying genes affect traits, their chromosomal locations and variants, and, ultimately, their modes of action at the biochemical level. Linkage analysis is an initial step in elucidating the genetic mechanisms affecting a trait of interest. This paper reviews genetic linkage analysis, with an emphasis on the score test approach developed by Dudoit and Speed [8, 10]. Two extensions of the test under current investigation are also presented: use of the test with larger sets of relatives than pairs, and generalization to allow for missing DNA identity by descent (IBD) information.

Keywords: allele sharing; complex traits; identity by descent; linkage analysis; pedigree; score test

1 Introduction

A central problem in genetic analysis is to determine which gene(s), if any, affect particular phenotypes, the chromosomal locations of these genes, their different alleles and, ultimately, their biochemical modes of action. Linkage analysis is an initial step in elucidating the genetic mechanisms affecting a trait of interest. Its goal is to determine the chromosomal location of the gene(s) influencing the trait. Linkage analysis proceeds by tracking patterns of coinheritance of the trait of interest and other traits or genetic markers, relying on the varying degree of recombination between trait and marker loci to map the loci relative to one another.

Mendel's second law of inheritance hypothesizes that different "factors" (traits or genes) segregate to gametes (sperm or egg) independently. Actually, independent assortment of gene pairs only occurs when the genes are on different chromosomes or are so far apart on the same chromosome that there is the same chance of recombination as nonrecombination. Such pairs of genes are said to be *unlinked*. Two genes are *linked* when they do not segregate independently. A measure of the degree of linkage is the *recombination fraction*, the chance of recombination occurring between two loci, denoted almost universally in the genetics literature as θ. For unlinked genes, $\theta = 1/2$; for linked genes, $0 \leqslant \theta < 1/2$. The following gives a brief introduction to linkage analysis; more substantial detail is provided by Ott [30], McPeek [27], and Speed [32].

Data for linkage analysis consist of sets of related individuals (*pedigrees*) and information on the genetic marker and/or trait *genotypes* (the two alleles at a locus) or *phenotypes* (the outward manifestation of a trait), usually selected on the basis of phenotype (*e.g.* a disease, such as diabetes, or a quantitative trait, such as glucose tolerance). For this setup, the recombination fraction is most commonly estimated by the method of maximum likelihood, the likelihood being determined by an appropriate genetic model for the coinheritance of the loci. The conventional measure of support for the hypothesis of linkage between two loci at recombination fraction θ versus that of no linkage is given by the *lod score*

$$Z(\theta) = log_{10}\left[\frac{L(\theta)}{L(1/2)}\right],$$

where $L(\theta) \propto f(\underline{X} \mid \theta)$ denotes the likelihood for θ given the observed data $\underline{X}$. Positive values of Z are evidence of linkage, while negative values indicate no linkage. With lod score linkage analysis, the null hypothesis of no linkage ($H_0 : \theta = 1/2$) is rejected for sufficiently large values of $Z(\hat{\theta}_{MLE})$, often taken to be 3. Linkage analysis based on the lod score is referred to in the genetics literature as "parametric" or "model-based" linkage analysis, as the mode of inheritance must be specified using some parametric model.

Genetic linkage mapping has been successful at mapping genes for traits following Mendelian inheritance patterns, typically recessive or dominant diseases. Identifying genes affecting complex traits, or traits not following these simple modes of inheritance, has proven to be more challenging. Lod-score linkage analysis for complex traits is difficult to carry out due to many complicating factors. Chief among these is that the mode of inheritance is rarely known. "Nonparametric," or "model-free," approaches thus have appeal, since they do not require a genetic inheritance model to be specified. Such methods usually focus on identical by descent (IBD) allele sharing at a locus between a pair of relatives. DNA at a locus is shared by two relatives *identical by descent* if it originated from the same ancestral chromosome. In families of individuals possessing the trait of interest, there is association between allele sharing at loci linked to trait susceptibility loci and the trait (see *e.g.* Dudoit and Speed [9] for examples). This association may be used to localize trait susceptibility genes. For loci unlinked to trait susceptibility loci, IBD sharing of DNA is not associated with the occurrence of the trait. Early work on linkage analysis using IBD data from sib-pairs can be found in Day and Simons [6] for qualitative traits, and in Haseman and Elston [20] for quantitative traits.

Testing for linkage with IBD data has developed along different lines, depending on the type of trait. For qualitative traits, the test is based on *IBD sharing conditional on phenotypes*, *e.g.* affected sib-pair methods (see [21] for a review). On the other hand, for quantitative trait loci (QTL), linkage analysis is based on examination of *phenotypes conditional on sharing*. A very widely used procedure in QTL mapping in humans is the Haseman-Elston method [20], implemented for sib-pairs and other

relative pairs in the SIBPAL and RELPAL programs of the computer package S.A.G.E. [11]; many extensions of it are also available [1, 2, 3, 12, 16, 17, 24, 28, 29]. In this method, the squared difference in phenotype values for the two relatives is regressed on the (estimated) proportion of alleles they share IBD. The method can also be used with qualitative traits (binary coding), but is clearly not appropriate for analysis of relatives where the phenotypic difference is fixed by design (*e.g.* affected sib-pairs). A disadvantage of the standard Haseman-Elston method is that it uses only differences in the phenotypes rather the full joint phenotypic data, incurring possible information loss [36].

The pattern of IBD sharing at a locus within a pedigree is summarized by an *inheritance vector*, which completely specifies the ancestral source of DNA [25]. For sibships of size k, it is convenient to label paternally derived alleles at the locus (1, 2) and maternally derived alleles (3, 4). The inheritance vector at a given locus is the vector $x = (x_1, x_2, ..., x_{2k-1}, x_{2k})$, where for sib i, x_{2i-1} is the label of the paternally inherited allele (1 or 2) and x_{2i} is that of the maternally inherited allele (3 or 4) at the locus. Note that the labels 1, 2, 3, and 4 for the parental DNA only have meaning within a sibship, and may therefore correspond to different sequences of DNA in different sibships.

Inheritance vectors for sibships may be grouped into IBD configurations which can be thought of as orbits of groups acting on the set of possible inheritance vectors (Dudoit and Speed [8], Ethier and Hodge [13]). For a pair of sibs, when paternal and maternal allele sharing are not distinguished, the 16 possible inheritance vectors give rise to three IBD configurations C_j: the sibs may share 0, 1, or 2 alleles IBD at the locus (Table 1). In the case of *affected* sib-trios, that is, all three sibs are affected with the trait under study, there are four IBD configurations (Table 2); in the case of a quantitative trait on sib-trios, the number of IBD configurations is 10 (Table 3).

Table 1: Sib-pair IBD configurations

Alleles IBD	Inheritance vectors	$\lvert C_j \rvert$
0 IBD	(1, 3, 2, 4), (1, 4, 2, 3), (2, 3, 1, 4), (2, 4, 1, 3)	4
1 IBD	(1, 3, 1, 4), (1, 4, 1, 3), (2, 3, 2, 4), (2, 4, 2, 3)	8
	(1, 3, 2, 3), (1, 4, 2, 4), (2, 3, 1, 3), (2, 4, 1, 4)	
2 IBD	(1, 3, 1, 3), (1, 4, 1, 4), (2, 3, 2, 3), (2, 4, 2, 4)	4

2 Score Test for Linkage

2.1 General Form of the Score Test

Dudoit and Speed [8, 10] proposed a score test to detect linkage with IBD data on sets of relatives. This approach represents a unified likelihood-based approach to the linkage

Table 2: Affected sib-trio IBD configurations

IBD configuration C_j	Pair-wise IBD sharing[a]	Representative inheritance vector	$\lvert C_j \rvert$
1	2, 2, 2	(1, 3, 1, 3, 1, 3)	4
2	2, 1, 1	(1, 3, 1, 3, 1, 4)	24
3	1, 1, 0	(1, 3, 1, 4, 2, 3)	24
4	2, 0, 0	(1, 3, 1, 3, 2, 4)	12

[a]Number of alleles shared IBD between sibs 1 and 2, 1 and 3, 2 and 3, respectively for the representative vector; this order may not be the same for each vector in the configuration

Table 3: Sib-trio IBD configurations for quantitative traits

IBD configuration C_j	Pair-wise IBD sharing[a]	Representative inheritance vector	$\lvert C_j \rvert$
1	2, 2, 2	(1, 3, 1, 3, 1, 3)	4
2	2, 1, 1	(1, 3, 1, 3, 1, 4)	8
3	2, 0, 0	(1, 3, 1, 3, 2, 4)	4
4	1, 1, 0	(1, 3, 1, 4, 2, 3)	8
5	1, 0, 1	(1, 3, 1, 4, 2, 4)	8
6	1, 1, 2	(1, 3, 1, 4, 1, 4)	8
7	0, 0, 2	(1, 3, 2, 4, 2, 4)	4
8	0, 2, 0	(1, 3, 2, 4, 1, 3)	4
9	1, 2, 1	(1, 3, 1, 4, 1, 3)	8
10	0, 1, 1	(1, 3, 2, 4, 1, 4)	8

[a]Number of alleles shared IBD between sibs 1 and 2, 1 and 3, 2 and 3, respectively

analysis of qualitative and quantitative traits using IBD data on pedigrees. The likelihood for the recombination fraction θ, conditional on the phenotypes of the relatives, is used to form a score test of the null hypothesis of no linkage ($\theta = 1/2$).

The probability vector of IBD configurations, conditional on pedigree phenotypes, at a *marker* locus linked to a trait susceptibility locus at recombination fraction θ can be written as

$$\rho(\theta,\pi)_{1\times m} = \pi_{1\times m} T(\theta)_{m\times m},$$

where π represents the conditional probability vector for IBD configurations at the *trait* locus and the number of IBD configurations is m. $T(\theta)$ denotes the transition matrix between IBD configurations at loci separated by recombination fraction θ, and has infinitesimal generator Q. The probability vector π will in general depend on (possibly very many) unknown genetic parameters. Under the null hypothesis that the marker

and trait susceptibility loci are unlinked, the IBD sharing distribution at the marker is given by the stationary distribution of $T(\theta)$, which is

$$\alpha = \rho(1/2, \pi) = \frac{1}{K}(|\, C_1 \,|, \ldots, |\, C_m \,|),$$

where $|\, C_j \,|$ is the number of inheritance vectors in IBD configuration C_j and K is the number of inheritance vectors. For general pedigrees, K is 2 raised to the number of relevant meioses, *e. g.*, for sib-pairs, $K = 2^4$.

For a given pedigree type, the form of the score test statistic is determined by the second largest eigenvalue λ_2 and corresponding eigenvector(s) of Q. The eigenvalues and their multiplicities give information regarding the form the score statistic takes. The eigenvalues are negative even integers. If $\lambda_2 = -2\kappa$, the score test is based on the κ^{th} derivative of the log-likelihood. If λ_2 has multiplicity 1, then the score statistic is independent of the genetic model for the trait. In sibships, $\lambda_2 = -4$, with multiplicity depending on the group that defines the IBD configurations (Dudoit and Speed [8]).

For sibships, because the first derivative in the Taylor series expansion of the log-likelihood about the null value $\theta = 1/2$ is 0, the score statistic is based on the second derivative $T''(1/2) = 8P_{-4}$, where P_{-4} is the projection matrix for the eigenvalue -4 and having rank the multiplicity of -4.

The score test approach is motivated by a large number of advantages, including: it is locally most powerful for alternatives close to the null; unlike a number of tests for linkage, the score test does not depend on assumptions such as population genotypes being in Hardy-Weinberg equilibrium – any genotype distribution can be used; conditioning on phenotypes eliminates selection bias introduced by nonrandom ascertainment, which is how samples are commonly obtained in practice; combining differently ascertained pairs is straightforward, which is important because otherwise some portion of the data may not be used. And as is seen below, the power and apparent robustness properties make the test an attractive alternative to nonparametric tests.

2.2 Score Test for Pairs of Relatives

The linkage information from IBD and phenotype data on n sib-pairs is combined into the score statistic

$$S_{sib}(\nu) = 16 \sum_{i=1}^{n} (\pi_{2i} - \pi_{0i})(N_{2i} - N_{0i}),$$

where $\pi_{ji} = \pi_j(\phi_{1i}, \phi_{2i}; \nu)$ is the conditional probability, given phenotypes (ϕ_{1i}, ϕ_{2i}) and genetic model parameters ν, that sib-pair i shares j alleles IBD ($j = 0, 1, 2$) at the *trait* locus (which could be one of several unlinked loci contributing to the trait); N_{ji} is 1 if sib-pair i shares j alleles IBD at the *marker* locus and 0 otherwise; and the sum is over all sib-pairs in the sample. The null IBD distribution at the marker is (1/4, 1/2, 1/4) for sharing (0, 1, 2) alleles.

For half-sib, avuncular, and grandparental pairs, the form of the test statistics is

$$S_{rel}(\mathrm{v}) \propto \sum_{i=1}^{n} (\pi_{1i} - \pi_{0i})(N_{1i} - N_{0i});$$

the constant of proportionality and its sign differs for these relative types, but cancels in standardization. For these relative pairs, the null IBD distribution at the marker is (1/2, 1/2) for sharing (0, 1) alleles. For pairs of cousins, the score test statistic is given by

$$S_{cousin}(\mathrm{v}) = 12 \sum_{i=1}^{n} (\pi_{1i} - \pi_{0i}/3)(N_{1i} - N_{0i}/3).$$

In this case, the null IBD sharing probabilities are (3/4, 1/4) for (0, 1) alleles.

For these types of relative pairs, the form of the score test statistic is fairly simple and readily interpretable. The statistic can be viewed as a weighted combination of IBD scores for each pair type, where the weights are given by differences in sharing probabilities conditional on phenotypes. For qualitative traits in sib-pairs, the weights depend on the genetic model but are constant in the phenotype and hence factor out. Thus, no genetic model is required. In general, however, this is not the case and a genetic model must be assumed in order to compute the weights.

The power and robustness properties of the score test were extensively studied via simulation of sib-pair and general relative pair data on a quantitative trait (Goldstein, Dudoit and Speed [18, 19]). For these studies, data were generated under a biallelic major gene model for the quantitative trait ϕ consisting of a single gene effect g with residual variation e, so that $\phi = \mu + g + e$, with μ the overall mean. Genotypic effect values are $g = a$ (> 0) for an A_1A_1 individual, $g = d$ for an A_1A_2 individual, and $g = -a$ for an A_2A_2 individual (see *e.g.* Falconer and Mackay [14]). The error term e has mean 0 and variance σ_e^2, constant across genotypes. The joint distribution of the error terms for a pair of relatives was assumed to be bivariate normal, with correlation ρ. Thus, in the population the trait is distributed as a mixture of bivariate normals, with mixing probabilities equal to the genotype frequencies. The *heritability* of a trait due to the genetic locus is the proportion of genetic variance to total variance: $H = \sigma_g^2/(\sigma_g^2 + \sigma_e^2)$. The parameters d, p, ρ, and H were varied, along with the selection strategy used to obtain pairs. Each simulated data set was analyzed with every model under consideration (one correct, the others wrong). This set of models was chosen as it is widely used in simulation studies of methods for analyzing quantitative traits.

In many realistic simulation scenarios, the score test approach showed large power gains over commonly used nonparametric tests, even when the assumed model for analysis deviated greatly from the true generating model. Based on the simulations, a generic additive model was recommended when little is known about the true underlying model.

Although the focus here has been on pairs of relatives, Dudoit [7] showed that the same score test approach is more generally applicable to any set of relatives. In practice, families included in studies of genetic traits often consist of more than a single

pair of relatives. In addition, these simulation studies, like most, have considered only the case when complete IBD information is available. Yet realistically, the genetic information necessary to determine IBD status may be incomplete for some individuals. Thus, further investigation of test properties and feasibility of implementation for these situations is warranted.

3 Some Extensions of the Score Test

The score test approach is quite general, and its implementation for pairs of relatives may be generalized in a number of ways. For example, the model for phenotypes may be expanded to include covariates. We consider here a few other extensions that we are currently researching: first, derivation and implementation of the test for larger pedigrees and an examination of test feasibility and properties in this case; second, modification of the test to accomodate data with incomplete IBD information.

3.1 Score Test for Sib Trios

The next largest pedigree to consider, after small "pedigrees" of pairs of individuals, would contain three individuals. Here we consider sib-trios, in both the case of a qualitative trait and for a quantitative trait.

For a qualitative trait, sib-trios may have the same trait values, as do affected sib-trios (ASTs), or they may instead be discordant (DSTs), where one has a different value than the other two. The IBD configurations for ASTs are given in Table 3. The infinitesimal generator for the IBD configuration transition matrix $T(\theta)$ is

$$Q_{AST} = \begin{bmatrix} -6 & 6 & 0 & 0 \\ 1 & -4 & 2 & 1 \\ 0 & 2 & -4 & 2 \\ 0 & 2 & 4 & -6 \end{bmatrix},$$

which has eigenvalues $\lambda = 0, -4, -8, -8$ (Dudoit and Speed [8]). The score statistic is

$$S_{AST} = \frac{16}{3}(3\pi_1 + \pi_2 - \pi_3 - \pi_4)(3N_1 + N_2 - N_3 - N_4),$$

where N_j denotes the number of ASTs with IBD configuration C_j at the marker. Although the form of the statistic here is a little more complicated than that for sib-pairs S_{sib}, it is not overly so. For DSTs, however, there are seven IBD configurations, and the eigenvalue $\lambda_2 = -4$ has multiplicity two, leading to a score statistic that is the sum of two statistics similar to S_{AST}, but with seven rather than four terms in each factor.

To complete the picture for sib-trios, we have derived the score test statistic in the case of a quantitative trait (QST) as well. In this case, there are 10 IBD configurations

(Table 3); the infinitesimal generator Q here is

$$Q_{QST} = \begin{bmatrix} -6 & 2 & 0 & 0 & 0 & 2 & 0 & 0 & 2 & 0 \\ 1 & -6 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 2 & -6 & 2 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & -6 & 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & -6 & 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & -6 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 2 & 2 & -6 & 0 & 0 & 2 \\ 0 & 0 & 0 & 2 & 0 & 0 & 0 & -6 & 2 & 2 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & -6 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & -6 \end{bmatrix},$$

with eigenvalues 0 (multiplicity 1), -4 (multiplicity 3), and -8 (multiplicity 6). The three orthonormal (unit norm with respect to the inner product $\langle\ ,\ \rangle_\alpha$) right eigenvectors corresponding to $\lambda_2 = -4$ are

$$u = (0,-1,-2,-1,0,1,2,0,0,1)$$
$$v = (2,0,-2,-1,-1,1,0,0,1,0)$$
$$w = \sqrt{2}(-1,0,1,1,0,-1,-1,1,0,0).$$

Thus, the score statistic for n QSTs is based on the second largest eigenvalue of Q and is given by

$$S_{QST} = 8\sum_{i=1}^{n}\left(\sum_{j=1}^{10} u_j\pi_{ji}\right)\left(\sum_{k=1}^{10} u_k N_{ki}\right) + 8\sum_{i=1}^{n}\left(\sum_{j=1}^{10} v_j\pi_{ji}\right)\left(\sum_{k=1}^{10} v_k N_{ki}\right) + 8\sum_{i=1}^{n}\left(\sum_{j=1}^{10} w_j\pi_{ji}\right)\left(\sum_{k=1}^{10} w_k N_{ki}\right)$$

$$= 8\sum_{i=1}^{n}(-\pi_{2i} - 2\pi_{3i} - \pi_{4i} + \pi_{6i} + 2\pi_{7i} + \pi_{10i})(-N_{2i} - 2N_{3i} - N_{4i} + N_{6i} + 2N_{7i} + N_{10i})$$
$$+ 8\sum_{i=1}^{n}(2\pi_{1i} - 2\pi_{3i} - \pi_{4i} - \pi_{5i} + \pi_{6i} + \pi_{9i})(2N_{1i} - 2N_{3i} - N_{4i} - N_{5i} + N_{6i} + N_{9i})$$
$$+ 8\sqrt{2}\sum_{i=1}^{n}(-\pi_{1i} + \pi_{3i} + \pi_{4i} - \pi_{6i} - \pi_{7i} + \pi_{8i})(-N_{1i} + N_{3i} + N_{4i} - N_{6i} - N_{7i} + N_{8i}).$$

So for quantitative traits, even with only one extra individual, the form of the score statistic is already much more complicated, and correspondingly much less interpretable, than it is for pairs. In addition, specification of a joint phenotypic model is more cumbersome for larger pedigrees, and may also be unstable due to the larger number of IBD

configurations. Furthermore, even once a model is specified, exact calculation of the statistic also becomes more difficult.

In nonparametric linkage analysis, the problem of dealing with larger sets of relatives than pairs has been approached in a number of different ways [26, 33, 34, 35]; for reviews, see [5, 15]. A widely used method to handle the issue is to consider the set of relatives only pairwise, typically by considering all possible pairs [22]. We have begun to compare exact treatment of QSTs with approximations based on pairwise score statistics. We hope to arrive at a weighting scheme based on pairs which will provide a good approximation to the exact treatment, yet is simpler and faster to compute and interpret.

3.2 Score Test with Missing IBD Information

Computing the score statistic relies on availability of complete inheritance vectors, so that there is sufficient genotypic information to determine IBD allele sharing status. In practice, however, the available genotype data may be limited to information on the allele states (identity by state, or IBS) and thus there is some information missing. IBD status may also be missing due to failure of the genotyping method for some individuals or unavailability of connecting individuals in the pedigree. It is therefore desirable to modify the score test to allow for the case of incomplete genotypic information.

When IBD information is incomplete, partial information obtained from marker data may be summarized by the *inheritance distribution*, a conditional probability distribution over possible inheritance vectors at the marker locus [23, 24]. Now, rather than counting the number of pedigrees with IBD configuration C_j, let

$$r_{ji} = P(\text{Pedigree has IBD configuration } C_j \text{ at the marker} \mid M_i),$$

where M_i denotes available marker information. Then a natural test statistic $\widetilde{S}(\nu)$ may be obtained from the complete data score statistic $S(\nu)$, by replacing the IBD indicators by their expectation given the marker data. When the trait and marker loci are in linkage equilibrium, then

$$\widetilde{S}(\nu) = E_0[S(\nu) \mid M, \phi]$$

for marker data M and phenotypes ϕ. Kruglyak *et al.* [23] use a similar statistic with a "perfect data" approximation, which consists of substituting the null variance of the complete data statistic, $Var_0[S(\nu) \mid \phi]$, for the null variance $Var_0[\widetilde{S}(\nu) \mid \phi]$ of the incomplete data statistic. This approximation is conservative, as $Var_0[\widetilde{S}(\nu) \mid \phi] \leqslant Var_0[S(\nu) \mid \phi]$.

In fact, the true inheritance distribution $\{r_{ji}\}$ will rarely be known; rather, it must be estimated from the data, for example with the program GENEHUNTER [23]. Call these estimated probabilities $\{\hat{r}_{ji}\}$. Then the incomplete data statistic for sib-pairs (ignoring

the multiplicative constant 16) is

$$\widetilde{S}(\nu) = \sum_{i=1}^{n} (\pi_{2i} - \pi_{0i})(\hat{r}_{2i} - \hat{r}_{0i}).$$

The null expectation and variance of $\widetilde{S}(\nu)$ may be estimated using sample moments of $\hat{r}_{2i} - \hat{r}_{0i}$ from the data. This approach may be problematic, though, as there must be a sufficient number of sib-pairs with the same missing genotype pattern to give reliable estimates. This aspect is even worse with larger pedigrees.

We believe that *multiple imputation* provides a more promising approach to estimation of the linkage score statistic with missing IBD data. Rubin [31] details multiple imputation procedures in the context of survey nonresponse; for multiple imputation in genetics problems, see Clayton [4]. With single imputation, one value is chosen for the missing information. With multiple imputation, missing data are replaced with at least two values representing the distribution of possibilities. Multiple imputation methods allow standard complete data methods to be applied, have increased efficiency over single imputation methods, and also more realistically reflect the increase in uncertainty due to the missing information. Thus, we are currently working to extend the applicability of the score test using multiple imputation to estimate missing IBD sharing.

Sampling from the imputation distribution $[M_{missing} \mid M_{observed}]$ of the marker information under the null T times yields multiple copies of "complete" data. Each of these produces a statistic $S^{(t)}$, $t = 1, ..., T$ (we now suppress the dependence of S on genetic model ν to avoid cumbersome notation below). Then we can define the multiple imputation "score" statistic S^* as the average value of $S^{(t)}$ over the T copies:

$$S^* = \frac{1}{T} \sum_{t=1}^{T} S^{(t)}.$$

Under the null, $E(S^*) = 0$ and

$$Var(S^*) = V - \frac{(T-1)}{T} E_{observed} \left\{ Var_{[M_{missing} | M_{observed}]}(S) \right\},$$

where $V = Var_0[S \mid \phi]$ is the complete data score statistic variance under the null [4]. The second term may be estimated using the sample variance of the imputation statistics $S^{(t)}$.

This method of extending the score test for linkage may be viewed as a compromise between the conservative "perfect data" approximation and exact calculation by enumeration of all possible states for the missing marker data. Such evaluation quickly becomes infeasible when several markers, each with large numbers of alleles, are used, as is common in linkage studies. There will be a reduction of power attributable to missing information, but preliminary simulations using the multiple imputation approach are encouraging. We are working toward a more complete implementation with the aim of broadening the class of problems to which the score test approach may be applied.

Sandrine Dudoit, Division of Biostatistics, University of California, Berkeley,
sandrine@stat.berkeley.edu

Darlene R. Goldstein, Bioinformatics Core Facility, Institut Suisse de Recherche Expérimentale sur le Cancer (ISREC), 1066 Epalinges, Switzerland; and Institut de mathematiques (IMA), École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland,
darlene.goldstein@isrec.unil.ch

References

[1] C. I. Amos. Robust variance-components approach for assessing genetic linkage in pedigrees. *American Journal of Human Genetics*, 54:535–543, 1994.

[2] C. I. Amos and R. C. Elston. Robust methods for the detection of genetic linkage for quantitative data from pedigrees. *Genetic Epidemiology*, 6:349–360, 1989.

[3] C. I. Amos, R. C. Elston, A. F. Wilson, and J. E. Bailey-Wilson. A more powerful robust sib-pair test of linkage for quantitative traits. *Genetic Epidemiology*, 6:435–449, 1989.

[4] D. Clayton. Tests for genetic linkage and association with incomplete data. 2001. Invited talk; available at http://www-gene.cimr.cam.ac.uk/clayton/talks/enar01.pdf.

[5] S. Davis and D. E. Weeks. Comparison of nonparametric statistics for detection of linkage in nuclear families: single-marker evaluation. *American Journal of Human Genetics*, 61:1431–1444, 1997.

[6] N. E. Day and M. J. Simons. Disease-susceptibility genes — their identification by multiple case family studies. *Tissue Antigens*, 8:109–119, 1976.

[7] S. Dudoit. *Linkage analysis of complex human traits using identity by descent data.* Ph.D. thesis, Department of Statistics, University of California, Berkeley, 1999.

[8] S. Dudoit and T. P. Speed. A score test for linkage using identity by descent data from sibships. *Annals of Statistics*, 27:943–986, 1999.

[9] S. Dudoit and T. P. Speed. Triangle constraints for sib-pair identity by descent probabilities under a general multilocus model for disease susceptibility. In M. E. Halloran and S. Geisser, editors, *Statistics in Genetics*, volume 112 of *IMA Volumes in Mathematics and its Applications*, pages 181–221. Springer-Verlag, New York, 1999.

[10] S. Dudoit and T. P. Speed. A score test for linkage analysis of qualitative and quantitative traits based on identity by descent data on sib-pairs. *Biostatistics*, 1:1–26, 2000.

[11] S. A. G. E. Statistical analysis for genetic epidemiology. *Genetic Epidemiology*, 1998.

[12] R. C. Elston, S. Buxbaum, K. B. Jacobs, and J. M. Olson. Haseman and elston revisited. *Genetic Epidemiology*, 19:1–17, 2000.

[13] S. N. Ethier and S. E. Hodge. Identity-by-descent analysis of sibship configurations. *American Journal of Medical Genetics*, 22:263–272, 1985.

[14] D. S. Falconer and T. F. C. Mackay. *Introduction to Quantitative Genetics*. Longman, Essex, England, 4th edition, 1996.

[15] E. Feingold, K. K. Song, and D. E. Weeks. Comparison of allele-sharing statistics for general pedigrees. *Genetic Epidemiology*, 19 Suppl 1:S92–S98, 2000.

[16] D. W. Fulker and L. R. Cardon. A sib-pair approach to interval mapping of quantitative trait loci. *American Journal of Human Genetics*, 54:1092–1103, 1994.

[17] D. W. Fulker, S. S. Cherny, and L. R. Cardon. Multipoint interval mapping of quantitative trait loci, using sib pairs. *American Journal of Human Genetics*, 56:1224–1233, 1995.

[18] D. R. Goldstein, S. Dudoit, and T. P. Speed. Power of a score test for quantitative trait linkage analysis of relative pairs. *Genetic Epidemiology*, 19 Suppl 1:S85–S91, 2000.

[19] D. R. Goldstein, S. Dudoit, and T. P. Speed. Power and robustness of a score test for linkage analysis of quantitative traits using identity by descent data on sib pairs. *Genetic Epidemiology*, 20:415–431, 2001.

[20] J. K. Haseman and R. C. Elston. The investigation of linkage between a quantitative trait and a marker locus. *Behavior Genetics*, 2:3–19, 1972.

[21] E. R. Hauser and M. Boehnke. Genetic linkage analysis of complex genetic traits by using affected sibling pairs. *Biometrics*, 54:1238–1246, 1998.

[22] S. Hodge. The information contained in multiple sibling pairs. *Genetic Epidemiology*, 1:109–122, 1984.

[23] L. Kruglyak, M. J. Daly, M. P. Reeve-Daly, and E. S. Lander. Parametric and nonparametric linkage analysis: a unified multipoint approach. *American Journal of Human Genetics*, 58:1347–1363, 1996.

[24] L. Kruglyak and E. S. Lander. Complete multipoint sib-pair analysis of qualitative and quantitative traits. *American Journal of Human Genetics*, 57:439–454, 1995.

[25] E. S. Lander and P. Green. Construction of multilocus genetic maps in humans. *Proceedings of the National Academy of Sciences, USA*, 84:2363–2367, 1987.

[26] K. Lange. A test statistic for the affected sibset method. *Annals of Human Genetics*, 50:283–290, 1986.

[27] M. S. McPeek. An introduction to recombination and linkage analysis. In T. P. Speed and M. S. Waterman, editors, *Genetic Mapping and DNA Sequencing*, volume 81 of *IMA Volumes in Mathematics and its Applications*. Springer-Verlag, New York, 1996.

[28] J. M. Olson, S. Rao, K. B. Jacobs, and R. C. Elston. Linkage of chromosome 1 markers to alcoholism related phenotypes by sib pair linkage analysis of principal components. *Genetic Epidemiology*, 17 Suppl 1:S271–S276, 1999.

[29] J. M. Olson and E. Wijsman. Linkage between quantitative trait and marker locus: methods using all relative pairs. *Genetic Epidemiology*, 10:87–102, 1993.

[30] J. Ott. *Analysis of Human Genetic Linkage*. Johns Hopkins University Press, Baltimore, 3rd edition, 1999.

[31] D. B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York, 1987.

[32] T. P. Speed. What is a genetic map function? In T. P. Speed and M. S. Waterman, editors, *Genetic Mapping and DNA Sequencing*, volume 81 of *IMA Volumes in Mathematics and its Applications*. Springer-Verlag, New York, 1996.

[33] P. J. Ward. Some developments on the affected-pedigree-member method of linkage analysis. *American Journal of Human Genetics*, 52:1200–1215, 1993.

[34] D. E. Weeks and K. Lange. The affected-pedigree-member method of linkage analysis. *American Journal of Human Genetics*, 42:315–326, 1988.

[35] A. S. Whittemore and J. Halpern. A class of tests for linkage using affected pedigree members. *Biometrics*, 50:118–127, 1994.

[36] F. A. Wright. The phenotypic difference discards sib-pair QTL linkage information. *American Journal of Human Genetics*, 60:740–742, 1997.

Cost Efficiency of Genetic Linkage Studies Using Mixtures of Selected Sib-pairs

Jian Han and Rudy Guerra

Abstract

Sib-pairs are relatively easy to collect and use of extreme quantitative phenotypes provide high statistical power. Thus, selected sib-pair (discordant, concordant) study designs are among the most useful in quantitative genetic linkage analysis. Dudoit and Speed [5, 6] proposed a score test for linkage that allows analysis of any sample, random or selected, by conditioning on phenotype and analyzing genotype. Selected sampling strategies have largely focused on studies collecting data on a single type of sib-pair. Using the score test statistic, we demonstrate that sampling designs based on a mixture of sib-pair types are more cost efficient than the traditional single selection scheme. In particular, there is no need to discard a large fraction of screened individuals. Cost efficient designs are based on a mixture of concordant and discordant sib-pairs, with the selection threshold of concordant sib-pairs more stringent than that of discordant pairs. General guidelines for the thresholds are given as a function of mode of inheritance, allele frequency, and residual correlation, as well as the cost ratio of phenotyping to genotyping. Since in many cases the mode of inheritance is not completely known, robustness with respect to assumed genetic models is also addressed.

Keywords: linkage; sample size; selected sample; sib-pair; study design

1 Introduction

It is well known that quantitative genetic linkage analysis based on random sampling of sib-pairs usually has low statistical power to detect non-Mendelian, quantitative, or complex disease loci. For example, Blackwelder and Elston [3] showed in simulations that even when heritability is moderate (30%) at a single locus, the power of the Haseman-Elston [11] linkage test based on random sampling of sib-pairs is low. Significant improvement in power can be achieved when an unselected sibling is regressed on the proportion (π) of alleles shared identical-by-descent (IBD) with a selected sibling [4]. Eaves and Meyer [7] provide evidence of additional power increases depending on the types of sib-pairs selected: discordant, with the sib-pair representing both tails of the phenotypic distribution; or concordant high (low), where both siblings are selected from the upper (lower) tail of the phenotypic distribution. However, see Allison *et al.* [1] for some limitations on the general utility of selected sib-pairs. Risch and Zhang

[13] also argue for selected sib-pair designs, but unlike previous authors they propose conditioning on the sampled (observed) phenotypes to analyze IBD sharing among the sib-pairs. Their discussion, however, is limited to study designs sampling a fixed type of sib-pair.

Dudoit and Speed [5, 6] generalize the work of Risch of Zhang [13] in three respects. First, although Dudoit and Speed also analyze IBD data, they specifically test for linkage in the traditional sense of evaluating a null hypothesis involving a recombination fraction ($H_0 : \theta = 0.5$), whereas Risch and Zhang evaluate a null hypothesis involving average allele-sharing ($H_0 : \pi = 0.5$). Second, Dudoit and Speed condition on observed phenotypes, while Risch and Zhang condition on phenotypic deciles. Lastly, the mean IBD statistic of Risch and Zhang is interpretable only for a fixed sib-pair type (*e.g.*, all discordant sib-pairs using a fixed threshold); the Dudoit and Speed score test statistic is not restricted to a fixed sampling scheme. By definition, both approaches reflect the actual sampling (conditioning on phenotype) and stochastic nature of the outcome (allele-sharing). In this sense they depart from making assumptions that are clearly violated under methods that model or analyze the phenotype, while viewing allele-sharing as "fixed" design variables. Both approaches are seemingly limited by having to specify knowledge of the gene action underlying the phenotype-genotype association. Robustness studies by Risch and Zhang [13], Zhao, Zhang, and Rotter [15], and Goldstein, Dudoit, and Speed [8], however, show that various characteristics of the approaches are fairly insensitive to misspecifying the mode of inheritance.

One drawback of selected study designs is that a large number of sib-pairs usually need to be screened in order to obtain the minimum sample size (number of sib-pairs) for the desired power. The more stringent the selection thresholds, the more screening that has to be done. Zhao *et al.* [15] evaluate cost efficiency across extremely discordant (ED), concordant high (CH), and concordant low (CL) sib-pair study designs. Considering the three types of designs separately, they conclude that ED sib-pair studies are the most cost efficient and robust against incorrect mode of inheritance and allele frequencies. They note, however, that more cost efficient studies may be possible by using all three types of selected sib-pairs.

Gu *et al.* [10] and Gu and Rao [9] report increased power over ED designs by combining all three types of sib-pairs into a single test statistic. In addition, they show that using all three types of sib-pairs is more cost effective than using just ED pairs. One issue that has yet to be fully addressed, in these and other investigations, is that the three types of sib-pairs may not be equally available in the population. Indeed, their prevalence is highly dependent on the underlying mode of inheritance. Ignoring this fact may lead to inefficient study designs, especially at screening where a lot of time and resources may be required to obtain certain extreme phenotypes that are relatively rare under the true gene action. Related to this idea is that better power and cost efficiency may be achieved by allowing different thresholds for the various sib-pair types.

In summary, there are not yet available general optimal sampling designs, defined by power or cost efficiency, for genetic linkage studies using selected sib-pairs. In this

paper we use the score test of Dudoit and Speed [5, 6] to develop such optimal sampling strategies.

2 Methods

Following standard major locus models (*e.g.*, Haseman and Elston [11]; Amos and Guerra [2]), we assume a locus A with two alleles, A_1 and A_2, with population allele frequencies p, $q\ (=1-p)$, respectively. Let x_{1i} and x_{2i} be the sib-pair phenotypic values of sib-pair i. The sib-pair phenotypes are modeled as:

$$\begin{aligned} x_{1i} &= \mu + g_{1i} + e_{1i}, \\ x_{2i} &= \mu + g_{2i} + e_{2i}, \end{aligned}$$

where μ is an overall mean of x; g_{ji} is the genetic effect due to trait locus A; e_{ji} represents combined residual genetic and environmental contributions with variance σ_e^2. The genetic effect g_{ji} equals a, d, and $-a$ according to genotypes A_1A_1, A_1A_2, and A_2A_2, respectively. To account for residual genetic and environmental correlations, we assume that the sib-pair model error (e_{1i}, e_{2i}) is distributed as a bivariate normal distribution with zero mean vector and correlation coefficient ρ. The additive and dominant components of genetic variation at locus A are defined as $\sigma_a^2 = 2pq[a-d(p-q)]^2$ and $\sigma_d = (2pqd)^2$. The heritability due to locus A is defined as $H = \sigma_g^2/(\sigma_g^2 + \sigma_e^2)$, where $\sigma_g^2 = \sigma_a^2 + \sigma_d^2$.

Under the null hypothesis of no linkage between a marker locus and trait locus, the proportion of genes shared IBD at the marker locus is expected to be $1/2$ regardless of the type of sib-pairs collected. When linkage is present mean IBD sharing among ED (CH/CL) pairs is expected to be less (more) than $1/2$. The test statistic used by Risch and Zhang [13] is the sample average for IBD sharing; it has an (asymptotic) null Gaussian distribution with mean zero and variance $1/8$. The Gaussian distribution under the alternative hypothesis of linkage depends on the selection scheme.

Gu *et al.* [10] proposed the extremely discordant and concordant (EDAC) test statistic, which combines ED, CH, and CL sib-pairs. It is defined as

$$T = \frac{1}{2(n_2+n_0)}\left\{\sum_{i=1}^{n_2}[X_{1i}(h,h)+X_{2i}(h,h)] + \sum_{i=1}^{n_0}[X_{1i}(l,l)+X_{2i}(l,l)]\right\} - \frac{1}{2n_1}\sum_{i=1}^{n_1}[X_{1i}(h,l)+X_{2i}(h,l)],$$

where h and l are indices of high and low tail thresholds [*e.g.* (10%,10%)=(10,10)], respectively; n_0 is the number of CL pairs, n_1 the number of ED pairs, and n_2 the number of CH pairs. X_{1i} is the number of alleles shared IBD from the father and X_{2i} is the number of alleles shared IBD from the mother. Consequently, $X_{1i}+X_{2i}$ is the number of allele shared IBD from the parents. The test statistic is thus a difference

between average proportion IBD-sharing among concordant sib-pairs and that among discordant sib-pairs. Under the null hypothesis, T is asymptotically distributed as a Gaussian random variable with mean 0 and variance $\sigma^2(T) = (n_1 + n_2 + n_0)/(8n_1(n_2 + n_0))$. A one-sided test statistic is given by $T/\sigma(T)$. Formulas for calculating sample size for ED pairs (and therefore for CH and CL pairs) are provided in Gu *et al.* [10].

In general, when a (test) statistic is formed through a linear combination of several available statistics, the number of observations entering each individual statistic and the weights in the combination typically have practical meaning and interpretation. In the present context, both factors may be motivated by mode of inheritance considerations. Risch and Zhang [13], for example, note that ED pairs are universally most useful among all possible types of sib-pairs; whereas CH or CL pairs are useful depending on mode of inheritance and allele frequency, when only a single type of sib-pair is used. Risch and Zhang [13, 14] and Zhao *et al.* [15] also discuss appropriate thresholds for selection sampling. Consider, for example, sampling top 10% and bottom 10% discordant sib-pairs. Under moderate to high positive residual correlation these extreme discordant pairs are relatively more difficult to find than concordant sib-pairs. It is possible to take advantage of the positive correlation by requiring a more stringent selection threshold to recruit more informative concordant pairs. Continuing with our example, we might set a threshold of (10,10) for ED pairs, while selecting top 5% CH pairs and bottom 5% for CL pairs. This flexibility may allow for increased statistical power and better cost efficiency by better selecting more informative sib-pairs. The score test of Dudoit and Speed allows for a broad range of selection strategies.

Dudoit and Speed [5, 6] proposed a score test for evaluating a null hypothesis of no linkage, $H_0 : \theta = 1/2$, against an alternative, $H_1 : 0 \leqslant \theta < 1/2$. The test can be used with the major gene model defined above. The statistic is

$$S(\nu) = 16 \sum_{i=1}^{n} (\pi_{2i} - \pi_{0i})(N_{2i} - N_{0i}),$$

where ν represents genetic parameters, such as values of a, d, p, σ_e^2, ρ, and mode of inheritance (recessive, dominant, additive); π_{2i} is the conditional probability that the *ith* sib-pair shares 2 genes IBD at the trait locus, given sib-pair phenotype (x_{1i}, x_{2i}); π_{0i} is similarly defined as the probability of sharing 0 genes IBD at the trait locus. N_{ji} is an indicator variable, equal to 1 if sib-pair i shares j ($j = 0 or 2$) genes IBD and 0 otherwise.

Under the null hypothesis, S is asymptotically normal with mean 0 and variance $\sigma^2(S) = \frac{1}{2}\sum_{i=1}^{n}(\pi_{2i} - \pi_{0i})^2$. The null hypothesis is rejected at level α when $S/\sigma(S) > z_\alpha$. Under the alternative hypothesis, S is asymptotically distributed as a normal random variable, $N(\mu_A, \sigma_A^2)$, where

$$\begin{aligned} \mu_A &= \sum_{i=1}^{n} (\pi_{2i} - \pi_{0i})(\tau_{2i} - \tau_{0i}), \\ \sigma_A^2 &= \sum_{i=1}^{n} (\pi_{2i} - \pi_{0i})^2 (\tau_{2i}\bar{\tau}_{2i} + \tau_{0i}\bar{\tau}_{0i} + 2\bar{\tau}_{0i}\bar{\tau}_{2i}); \end{aligned}$$

τ_{2i} is the conditional probability that the *ith* sib-pair shares 2 genes IBD at the *marker* locus, given sib-pair phenotype (x_{1i}, x_{2i}); τ_{0i} is similarly defined as the probability of sharing 0 genes IBD at the marker locus; $\bar{\tau} = 1 - \tau$.

The conditional asymptotic power of S, given the phenotypes, is denoted as

$$\Gamma(\theta, \nu; \mathbf{X}) = 1 - \Phi\left(\frac{z_\alpha \sqrt{\frac{1}{2}\sum_{i=1}^{n}(\pi_{2i} - \pi_{0i})^2} - \mu_A}{\sigma_A}\right), \tag{1}$$

where $\mathbf{X}$ is a $n \times 2$ matrix representing the n sib-pair phenotypes, and Φ is standard normal cumulative distribution function such that $\Phi(z_\alpha) = 1 - \alpha$. The unconditional power may be estimated by the average of a set of conditional powers generated under the same model.

The score test is derived through an approximation of the maximum likelihood ratio test and is locally most powerful [5, 6]. One criticism of the test is that it requires specification of a mode of inheritance model for the weights (π) to be determined. (The observed data are the counts, N.) As has been noted, however, the test appears to be sufficiently robust with respect to mode of inheritance assumptions (Goldstein, Dudoit and Speed [8]). The important feature of the test is that it faithfully reflects the non-random sampling that is typical of most genetic epidemiologic studies. We refer readers to the original papers for more technical details.

3 Sample Size Approximation

Although "randomly selected" sib-pairs may be of some utility - and would likely be available through the screening process - in this article we focus on using only ED, CH, and CL sib-pairs. Since the score test is conditional on observed data, there are no closed-form formulas for sample size calculations associated with the unconditional power based on (1); however, it is possible to approximate the power function for a given sampling selection. Under the assumption that the trait and marker loci are in complete linkage ($\theta = 0$), we have $\pi_{ji} = \tau_{ji}$ and expression (1) becomes

$$\Gamma(\nu; \mathbf{X}) = 1 - \Phi\left(\frac{Z_\alpha \sqrt{\frac{1}{2}\mu_A} - \mu_A}{\sigma_A}\right), \tag{2}$$

with

$$\mu_A = \sum_{i=1}^{n}(\pi_{2i} - \pi_{0i})^2, \tag{3}$$

$$\sigma_A^2 = \sum_{i=1}^{n}(\pi_{2i} - \pi_{0i})^2(\pi_{2i}\bar{\pi}_{2i} + \pi_{0i}\bar{\pi}_{0i} + 2\bar{\pi}_{0i}\bar{\pi}_{2i}). \tag{4}$$

To determine the sample sizes, the probability parameters (π) need to be estimated. To this end, define selection schemes (S), $TxBy$, $TxTx$, $BxBx$, where T and B indicate "top" (upper) and "bottom" (lower) tails of the phenotypic distribution, x and y tail areas. For example, the selection scheme $T10B10$ requires sib-pair phenotype (x_1,x_2) to satisfy $max(x_1,x_2) > p_{90}$ and $min(x_1,x_2) < p_{10}$, where p_h is the h^{th} percentile of the (marginal) phenotypic distribution.

By definition, $\pi_2 = P(\text{sib-pair shares 2 trait genes IBD} \mid x_1,x_2)$. Under a given selection scheme (S), π_2 can be estimated by

$$\hat{\pi}_2 \doteq \int\int_S \pi_2 \, dx_1 dx_2.$$

This is equivalent to the estimation of D_2 in equation (1) of Risch and Zhang [13]. Similarly, $\hat{\pi}_0$ can also be used to estimate π_0.

Let n be the total selected sample size, $n = n_{ED} + n_{CH} + n_{CL}$, where n_{ED}, n_{CH} and n_{CL} are the sample sizes of selected ED, CH and CL sib-pairs, respectively. For a specified genetic model and selection scheme, let P_{ED}, P_{CH}, and P_{CL} be the probability of randomly selecting an ED, CH, and CL sib-pair from the phenotype distribution; define $r_{ED} = P_{ED}/(P_{ED} + P_{CH} + P_{CL})$, the proportion of ED pairs in the population of ED, CH, and CL sib-pairs. Proportions r_{CH} and r_{CL} are similarly defined. Lastly, let $\hat{\pi}_{ED2}$ and $\hat{\pi}_{ED0}$ be estimates (as defined above) of π_2 and π_0, respectively, for ED sib-pairs. Denote similar estimates for CH and CL sib-pairs.

The mean (μ_A) of the estimated score statistic can thus be estimated as

$$\begin{aligned}
\mu_A &= \sum_{i=1}^{n_{ED}} (\pi_{2i} - \pi_{0i})^2 + \sum_{i=1}^{n_{CH}} (\pi_{2i} - \pi_{0i})^2 + \sum_{i=1}^{n_{CL}} (\pi_{2i} - \pi_{0i})^2 \\
&\approx n_{ED}(\hat{\pi}_{ED2} - \hat{\pi}_{ED0})^2 + n_{CH}(\hat{\pi}_{CH2} - \hat{\pi}_{CH0})^2 + n_{ED}(\hat{\pi}_{CL2} - \hat{\pi}_{CL0})^2 \\
&\approx n[r_{ED}(\hat{\pi}_{ED2} - \hat{\pi}_{ED0})^2 + r_{CH}(\hat{\pi}_{CH2} - \hat{\pi}_{CH0})^2 + r_{ED}(\hat{\pi}_{CL2} - \hat{\pi}_{CL0})^2] \\
&\stackrel{\text{def}}{=} nW.
\end{aligned}$$

In a similar way, the variance (σ_A^2) of the test statistic can be estimated, say nU. Substituting parameter estimates in the conditional power function (2) we obtain

$$Power = 1 - \beta \stackrel{\text{def}}{=} 1 - \Phi\left(\frac{z_\alpha\sqrt{\frac{1}{2}nW} - nW}{nU}\right),$$

and the corresponding sample size n is given by

$$n = \left[\frac{z_\alpha\sqrt{\frac{1}{2}W} - z_\beta\sqrt{U}}{W}\right]^2.$$

The sample sizes for ED, CH, and CL sib-pairs are then calculated as nr_{ED}, nr_{CH}, nr_{CL}, respectively.

We emphasize that when the selection schemes are determined, the number of ED, CH and CL sib-pairs to be selected are calculated according to their selection probability under an assumed genetic model. This method of selection makes use of the extreme sib-pairs relatively readily available in the population and minimizes wasting resources attempting to find sib-pairs that may be difficult to collect under the genetic model. Although the sample size calculation is an approximate one, simulations show that the observed power is always at least as great as the nominal power; see below.

Since the score test weights ED, CH and CL sib-pairs according to a working genetic model, it may be more powerful than the EDAC test whereby the three types of sib-pairs are treated equally. Therefore, for fixed power and type I error probability, the score test may require smaller sample sizes than the EDAC approach.

***Example* 1**

Table 1 shows sample sizes corresponding to $H = 0.3$, $1 - \beta = 0.8$ and $\alpha = 0.001$ under a $T10B10$ selection scheme for ED, $T5T5$ for CH, and $B5B5$ for CL sib-pairs. Under all parameter configurations considered ($\rho = 0.2, 0.4$; $p = 0.1, \ldots, 0.9$; recessive, dominant, and additive models), the two tests indicate the same qualitative pattern of sample size requirements. For example, when $\rho = 0.4$ and $p = 0.3$ under a recessive model, both tests require $n_{ED} \leqslant n_{CL} \leqslant n_{CH}$. However, the score test always requires smaller sample sizes in terms of the total (n) and specific sib-pairs (n_{ED}, n_{CH} or n_{CL}. Table 2 gives the average percent reduction in total sample size of the score test relative to the EDAC test. Higher reductions are obtained in the presence of higher residual correlation. The smallest average reduction (10%) is observed under additive gene action with lower residual correlation. Table 1 shows that ED sib-pairs are less informative when the sib-pairs have a relatively higher degree of (positive) residual correlation; compare n_{ED} sample sizes at $\rho = 0.2$ to $\rho = 0.4$ under each test. This makes sense since a higher degree of (positive) correlation would tend to make the phenotypes more similar. Indeed, both the score test and EDAC test have n_{CH} and n_{CL} each larger than n_{ED} when $\rho = 0.4$. The score test is also less sensitive then EDAC with respect to the given increase in residual correlation. Associated with an increase from $\rho = 0.2$ to $\rho = 0.4$ under the recessive model, the average (across p) percent increase in total sample size (n) for the score test is 28%; under the dominant model the average increase is 24% and under the additive it is 20%. The corresponding results for the EDAC test are 74%, 49%, and 50% under recessive, dominant, and additive models, respectively. Relatively larger sample sizes for extreme discordant sib-pairs are generally observed under a dominant model with lower residual correlation ($\rho = 0.2$), and $n_{ED} \approx n_{CH} > n_{CL}$ under an additive model with $\rho = 0.2$. In most other cases concordant sib-pairs are required more so that discordant pairs. The observed pattern of overall results remained the same when other selection schemes were considered (data not shown). ■

Table 1: Sample sizes requirements for score test and EDAC test. $H = 0.3$, power = .8 and $\alpha = 0.001$, with selection scheme T10B10 for ED, T5T5 for CH and B5B5 for CL sib-pairs. Number of ED, CH and CL sib-pairs denoted by n_{ed}, n_{ch} and n_{cl}, respectively; $n = n_{ed} + n_{ch} + n_{cl}$.

		Score Test	EDAC Test	Score Test	EDAC Test
		$\rho = 0.2$	$\rho = 0.2$	$\rho = 0.4$	$\rho = 0.4$
	p	$n(n_{ed}, n_{ch}, n_{cl})$	$n(n_{ed}, n_{ch}, n_{cl})$	$n(n_{ed}, n_{ch}, n_{cl})$	$n(n_{ed}, n_{ch}, n_{cl})$
	0.1	229(99,75,55)	580(251,190,139)	329(52,145,132)	1036(165,456,415)
	0.3	60(19,27,14)	102(32,46,24)	64(10,31,23)	152(23,73,56)
Rec	0.5	128(38,51,39)	146(44,58,44)	160(18,74,68)	257(28,120,109)
	0.7	165(46,60,59)	165(46,60,59)	241(22,110,109)	311(29,141,141)
	0.9	111(29,32,50)	139(36,40,63)	134(13,54,67)	245(23,99,123)
	0.1	103(36,41,26)	116(41,46,29)	118(18,55,45)	164(25,77,62)
	0.3	138(52,43,43)	143(54,44,45)	194(28,83,83)	211(31,90,90)
Dom	0.5	117(45,31,41)	131(50,35,46)	141(23,56,62)	183(30,73,80)
	0.7	63(24,13,26)	93(35,20,38)	64(14,22,28)	118(25,40,53)
	0.9	231(101,55,75)	578(253,137,188)	330(57,130,143)	970(168,382,420)
	0.1	90(30,38,22)	108(36,46,26)	100(14,49,37)	148(21,72,55)
	0.3	105(34,39,32)	109(35,41,33)	128(16,59,53)	155(19,71,65)
Add	0.5	112(34,39,39)	112(34,39,39)	139(15,62,62)	166(18,74,74)
	0.7	112(31,36,45)	119(33,38,48)	139(14,60,65)	184(18,79,87)
	0.9	94(24,25,45)	126(32,34,60)	110(11,43,56)	209(20,82,107)

4 Relative Importance of ED, CH, CL Sib-pairs

It is well known [13, 6] that extreme discordant sib-pairs are generally most powerful when a single selection scheme is used. Gu *et al.* [10] argue that concordant sib-pairs available in the screening pool provide an important additional source of linkage information and should be included in the selected sample. However, several practical questions remain unanswered, including the following. What are the relative merits of the various sib-pair types in a given study design? More specifically, how are the individual sample sizes (n_{ED}, n_{CH}, n_{CL}) related to linkage information. For a given level of power, are studies carried out with ED pairs alone more or less cost efficient than those that include mixtures of concordant and discordant sib-pairs? How should the thresholds for the different sib-pair types be chosen?

***Example* 2**

As a motivating example, consider an additive model with heritability $H = 0.3$, allele (A_1) frequency $p = 0.2$, and sib-pair residual correlation $\rho = 0.4$. Under this model a selection scheme of T15B15 for ED, T10T10 for CH, and B10B10 for CL pairs

Table 2: Percent decrease in score test total sample size (n) relative to EDAC test. Model parameters as in Table 1.

	Recessive		Dominant		Additive	
p	$\rho = 0.2$	$\rho = 0.4$	$\rho = 0.2$	$\rho = 0.4$	$\rho = 0.2$	$\rho = 0.4$
.1	61	78	11	28	17	32
.3	41	58	3	8	4	17
.5	12	38	11	23	0	16
.7	0	23	32	46	6	25
.9	20	45	60	66	25	47
Ave	27	48	23	34	10	27

corresponds to selection probabilities $P(ED) = 0.0103, P(CH) = 0.030$, and $P(CL) = 0.027$. At 80% power and type I error probability $\alpha = 0.001$, the required sample sizes are $n_{ED} = 28$, $n_{CH} = 81$, and $n_{CL} = 73$. Figure 1 shows the relative importance of each kind of pair. In plot (a) the number of ED pairs is fixed at 28, and numbers of CH and CL pairs vary from zero to the require sample size. When both n_{CH} and n_{CL} are zero, the power is 38%. The power gradually increases as more CH pairs are added, but the CL pairs do not affect power very much. Plot (b) clearly shows the importance of ED pairs with n_{CH} fixed at 81. Using the required 81 CH pairs alone yields a power of 26%. The relative importance of both ED and CH pairs is jointly exhibited in plot (c) where there are 73 concordant low pairs. Under this particular model, ED pairs affect power the most; CH pairs contribute as well, but the usefulness of CL pairs is very limited (73 CL pairs alone has power of nearly zero). As shown in plot (d), the sample size needed to achieve power of 0.8 using only ED pairs is 52. The expected number of randomly screened sib-pairs to obtain 52 ED pairs is $52/(0.0103) = 5048$; the expected number of randomly screened sib-pairs to obtain the mixed sample is $182/(0.0103 + 0.03 + 0.027) = 2704$. Note that the last calculation is not based on an optimal selection scheme, which may further reduce the screening size. ■

Example 2 makes evident that adding more sib-pairs (concordant or discordant) in the sample provides an increase in the power of the test, albeit possibly small. This is generally true for the score test regardless of the mode of inheritance, allele frequency and residual correlation. The EDAC test, however, occasionally loses power when CH pairs are combined with ED pairs.

***Example* 3**

In Example 2, the CL sib-pairs were least important in their contribution to the power of test, but this is not always the case. Consider a dominant model with $H = 0.3$, allele (A_1) frequency $p = 0.8$, and $\rho = 0.4$. The selection scheme is as in Example 2,

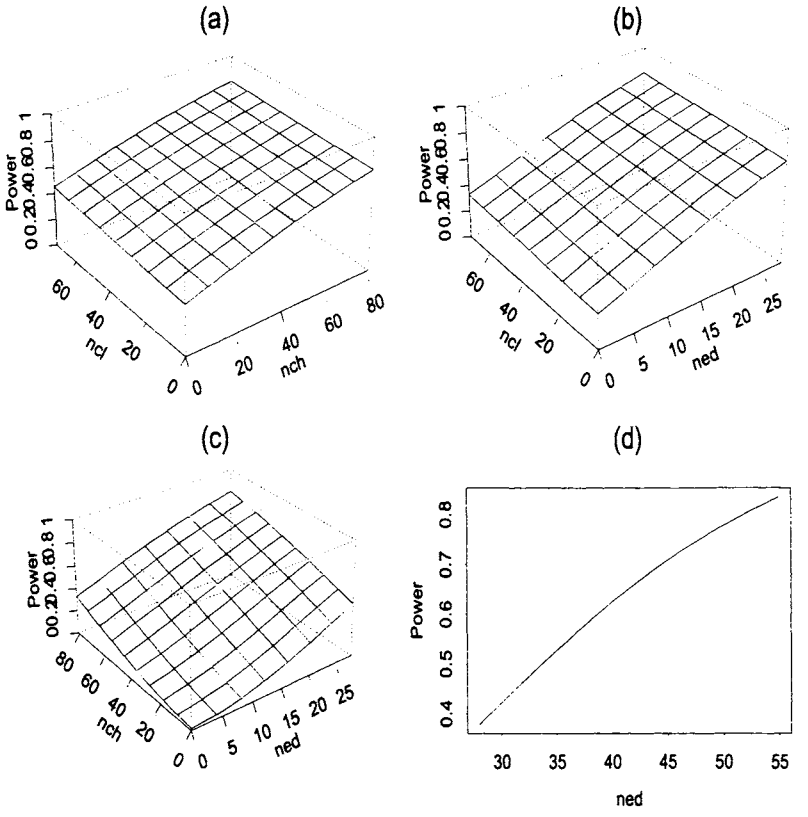


Figure 1: Power of score test for various combinations of sample sizes under selection scheme T15B15 for ED, T10T10 for CH and B10B10 for CL pairs. Nominal power = 0.8 and α = 0.001. True genetic parameters include H=0.3, p=0.2, ρ = 0.4, additive gene action. (a) Power of test for fixed number of ED pairs, $n_{ed} = 28$. (b) Power of test for fixed number of CH pairs, $n_{ch} = 81$. (c) Power of test for fixed number of CL pairs, $n_{cl} = 73$. (d) Power of test using only ED pairs.

$T15B15$ for ED, $T10T10$ for CH and $B10B10$ for CL pairs. Under this dominant model the selection probabilities for ED, CH and CL pairs are 0.0146, 0.0265, and 0.0295, respectively. At 80% power with $\alpha = 0.001$, the sample sizes for ED, CH and CL pairs are 43, 78, and 87, respectively. Contrary to the results of the previous example, the CL pairs are the most important in terms of power contribution, whereas both ED and CH pairs have a limited role; see plots (a), (b) and (c) of Figure 2. Using CL sib-pairs alone the test has moderate power at 68%. As shown in plot (d), a study with only $T15B15$ ED sib-pairs requires approximately 225 pairs to achieve the desired power of 80%. The expected number of sib-pairs screened for this ED-only study is $223/0.0146 = 15273$, compared with $208/(0.0146 + 0.0265 + 0.0295) = 2946$ for a mixture study. ■

The examples illustrate the potential savings in screening by using mixed sib-pair types in the score test. We also see that some sib-pair types are less useful than others in determining the power of the test.

5 Optimal Mixture of Sib-pairs

In this section we address the issue of optimal selection thresholds for ED, CH and CL sib-pairs in the selected sample in order to minimize the total cost of phenotyping and genotyping. Given a desired power and type I error probability, our goal is to find the optimal selection scheme for the score test such that the cost of the test is minimized.

We assume that sib-pairs are randomly chosen from the population and that the ratio (R) of phenotyping-to-genotyping cost ranges as 0.02, 0.1, 1, 10, 50. Eight selection thresholds for ED sib-pairs are considered: T10B10, T10B20, T10B25, T15B15, T15B25, T20B20, T25B25 and T30B70. Since moderate to high (positive) residual correlation makes it difficult to find an ED sib-pair, more stringent thresholds for this type of sib-pair are not considered here. Among these selection schemes, some are symmetric (*e.g.*, T10B10) and some are asymmetric (*e.g.*, T10B25). We consider seven symmetric selection schemes for CH (CL) pairs: T1T1, T3T3, T5T5, T10T10, T15T15, T20T20, T25T25 (B1B1, B3B3, B5B5, B10B10, B15B15, B20B20, B25B25). For now we assume equivalent tail areas for CH and CL sib-pairs (*e.g.*, T5T5 and B5B5). The more stringent thresholds for concordant pairs are chosen since they are relatively easier to recruit than ED pairs under positive residual correlation. Thus, the total number of selection schemes considered is 56 (8×7). This seems to be wide enough coverage to be practically useful. Heritability H is fixed at 0.1 or 0.3, allele frequency p ranges from 0.1 to 0.9 (by 0.2), residual correlation (ρ) takes values 0.1 or 0.4. The total number of genetic models considered is $2 \times 2 \times 5 \times 3 = 60$ (heritability $\times$ correlation $\times$ p $\times$ gene action).

The total cost of interest is the sum of the cost for phenotyping all screened sib-pairs required to obtain the total sample size and the cost of genotyping the selected sib-pairs. The total cost (TC) is calculated as $TC = 2RN + 2n$ (Zhao *et al.* [15]), where $n = n_{ED} + n_{CH} + n_{CL}$. N is the expected total number of screened sib-pairs calculated as $n/[P(ED) + P(CH) + P(CL)]$; R the cost ratio of phenotyping to genotyping. Without loss of generality, the cost of genotyping one individual is assumed to be 1 unit in the calculation of total cost. For each of the 60 genetic models, the optimal (minimum cost) sampling is obtained by searching all 56 stated selection schemes for a fixed cost ratio (R) of phenotyping to genotyping. For purposes of comparison with what might be considered accepted convention, we also report results for ED-only study designs; minimum cost is found among the eight ED selection schemes.

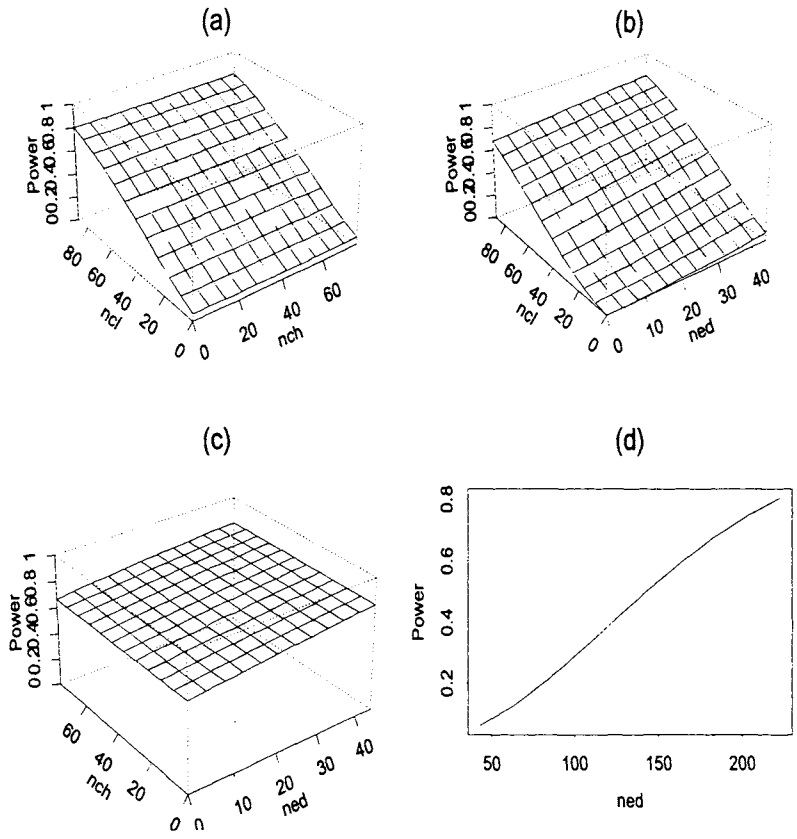


Figure 2: Power of score test for various combinations of sample sizes under selection scheme T15B15 for ED, T10T10 for CH and B10B10 for CL pairs. Nominal power = 0.8 and α = 0.001. True genetic parameters include H=0.3, p=0.8, ρ = 0.4, dominant gene action. (a) Power of test for fixed number of ED pairs, n_{ed} = 43. (b) Power of test for fixed number of CH pairs, n_{ch} = 78. (c) Power of test for fixed number of CL pairs, n_{cl} = 87. (d) Power of test using only ED pairs.

***Example* 4**

We first consider the case where phenotyping cost is very low compared to the cost of genotyping, $R = 0.02$, and $H = 0.3$. Optimal selection schemes with corresponding sample sizes and total costs (in thousand units) are listed in Table 3. Analogous results for an ED-only study are given in the right part of the table. Column 3 shows the optimal selection scheme for the given parameters; since both types of concordant pairs have equal tail areas their selection schemes are summarized as Tx/Bx. Column 4 shows the required sample sizes for the optimal sampling; column 5, the total cost (thousand units) when mixed (m) sib-pairs are used. Columns 6 – 8 show results for ED-only optimal studies. The dominant case is not reported since it is equivalent to a recessive case with upper and lower thresholds switched and p replaced by $1 - p$.

We first consider the recessive model. Here mixed samples require smaller total sample sizes and are generally more cost efficient than ED-only samples. At lower allele frequencies (approximately 0.3 or less), mixed samples are much more cost efficient than ED-only samples, independent of residual correlation. At the higher allele frequencies, discordant sib-pairs are generally more informative than are concordant pairs in that the ED pairs constitute the majority of the total sample size. Concordant high pairs are slightly more (less) informative than concordant low pairs at the lower (higher) allele frequencies; they are equally informative at $p \approx 0.5$. At higher residual correlation ($\rho = 0.4$), the optimal thresholds for discordant pairs become less stringent with increasing allele frequency. This relationship holds less so under weaker residual correlation ($\rho = 0.1$). This is what we might expect, since at higher degrees of (positive) residual correlation, ED pairs are observed as such because of linkage effects overriding the residual correlation. On the other hand, with higher residual correlation it is less clear whether concordant pairs are phenotypically similar because of genes or residual factors, which may or may not reflect genetic factors. Thus, ED pairs are relatively more important and the relaxing thresholds under $\rho = 0.4$ reflect the need to collect them. Conversely, the concordant thresholds are relatively more extreme in order to distinguish the genetic signal from the "noise" in the residual correlation. Under the additive case, the mixed and ED-only samples are about equally cost efficient, with the exception at very high allele frequencies (0.9 or higher). And, as in the recessive case, a higher degree of residual correlation is associated with less (more) stringent optimal thresholds for discordant (concordant) sib-pairs. The recessive and additive cases also share in common the fact that in most cases the concordant pairs represent a small fraction of the total sample size. ■

***Example* 5**

When phenotyping and genotyping costs are about the same ($R \approx 1$, Table 4), the total costs increase compared to the case $R < 1$ not only because of higher costs per individual, but also because the total sample sizes increase as well. Characteristics of study

Table 3: Optimal selection schemes found from all 56 possible selection combinations for recessive (top) and additive (bottom) model with $H = 0.3$. The cost ratio of phenotype-to-genotype is $R = 0.02$. Power = 0.8, $\alpha = 0.001$.

Recessive Model

ρ	p	ED, CH/CL	$n(n_{ed}, n_{ch}, n_{cl})$	COST	ED	n	COST
0.1	0.1	T10B10, T1/B1	53(43,9,1)	0.234	T10B25	8604	26.4
	0.3	T10B10, T5/B5	63(26,26,1)	0.237	T10B25	163	0.58
	0.5	T10B25, T5/B5	125(85,23,17)	0.412	T10B25	113	0.441
	0.7	T15B15, T1/B1	113(109,2,2)	0.477	T15B15	111	0.476
	0.9	T15B15, T5/B5	134(84,17,33)	0.444	T15B15	138	0.565
0.4	0.1	T10B10, T1/B1	24(12,9,3)	0.172	T10B20	1004	5.92
	0.3	T10B25, T5/B5	68(25,24,19)	0.216	T10B25	76	0.387
	0.5	T15B25, T1/B1	86(78,4,4)	0.353	T15B25	82	0.355
	0.7	T25B25, T1/B1	114(108,3,3)	0.357	T25B25	110	0.352
	0.9	T20B80, T3/B3	105(68,16,21)	0.359	T20B20	91	0.38

Additive Model

ρ	p	ED,CHCL	$n(n_{ed}, n_{ch}, n_{cl})$	COST	ED	n	COST
0.1	0.1	T10B25,T3B3	95(79,11,5)	0.307	T10B25	106	0.366
	0.3	T10B20,T3B3	100(81,11,8)	0.367	T10B20	96	0.389
	0.5	T15B15,T3B3	114(95,9,10)	0.413	T15B15	108	0.425
	0.7	T15B15,T5B5	126(80,19,27)	0.423	T15B15	112	0.464
	0.9	T10B10,T5B5	92(33,19,40)	0.374	T15B15	153	0.617
0.4	0.1	T10B25,T1B1	50(43,4,3)	0.229	T10B25	50	0.251
	0.3	T15B25,T1B1	69(63,3,3)	0.263	T15B25	67	0.264
	0.5	T20B20,T1B1	75(68,3,4)	0.284	T20B20	71	0.283
	0.7	T20B20,T1B1	77(69,4,4)	0.307	T20B20	74	0.312
	0.9	T15B15,T3B3	77(33,18,26)	0.236	T20B20	97	0.404

design and costs when $R = 1$, compared to $R = 0.02$, include a more prominent role of concordant sib-pairs, less stringent optimal thresholds, and higher gains in sample sizes and costs by the mixed sampling scheme. By relaxing the thresholds, we are able to recruit the desired number of sib-pair types without the need to screen prohibitively large numbers of sib-pairs. ■

The impact of residual correlation on the optimal mixture selection scheme is summarized in Table 5. The selection schemes for ED sib-pairs are listed in column 1 from most stringent (top) to least stringent (bottom); selection schemes for CH and CL are given in the first row from most stringent (left) least stringent(right). The top (bottom) panel gives results for $\rho = 0.1$ ($\rho = 0.4$). The entry is the frequency of the intersecting combination of ED and CHCL pairs defining an optimal design among 56 choices. For

Table 4: Optimal selection schemes found from all 56 possible selection combinations for recessive (top) and additive (bottom) models with $H = 0.3$. The cost ratio of phenotype-to-genotype is $R = 1$. Power = 0.8, $\alpha = 0.001$.

Recessive Model

ρ	p	ED,CHCL	$n(n_{ed}, n_{ch}, n_{cl})$	COST	ED	n	COST
	0.1	T10B10,T1B1	53(43,9,1)	6.44	T10B25	8604	475
	0.3	T10B25,T10B10	105(42,40,23)	3.5	T10B25	163	13
0.1	0.5	T15B25,T20B20	311(73,129,109)	4.78	T30B30	445	9.03
	0.7	T30B30,T25B25	492(178,157,157)	4.67	T30B30	306	6.92
	0.9	T25B25,T15B15	308(150,64,94)	4.9	T25B25	352	10.7
	0.1	T10B20,T1B1	43(31,9,3)	6.27	T10B25	1255	177
	0.3	T15B25,T10B10	134(38,52,44)	3.64	T15B25	125	11.3
0.4	0.5	T30B30,T15B15	300(120,93,87)	4.18	T30B30	208	6.63
	0.7	T30B30,T25B25	450(95,178,177)	4.11	T30B30	157	5.58
	0.9	T30B30,T15B15	308(118,89,101)	4.22	T30B30	229	7.44

Additive Model

ρ	p	ED,CHCL	$n(n_{ed}, n_{ch}, n_{cl})$	COST	ED	n	COST
	0.1	T10B25,T15B15	197(53,85,59)	4.28	T10B25	106	7.87
	0.3	T25B25,T20B20	380(153,118,109)	4.64	T30B30	381	6.99
0.1	0.5	T25B25,T20B20	379(144,117,117)	4.77	T30B30	468	7.32
	0.7	T25B25,T20B20	394(141,121,132)	5.1	T30B30	381	8.16
	0.9	T25B25,T15B15	337(166,71,100)	5.31	T25B25	405	12.1
	0.1	T15B25,T10B10	151(43,58,50)	3.97	T15B25	75	6.51
	0.3	T30B30,T15B15	285(122,84,79)	3.72	T30B30	176	4.91
0.4	0.5	T30B30,T15B15	287(116,85,86)	3.88	T30B30	170	5.18
	0.7	T30B30,T15B15	301(114,92,95)	4.2	T30B30	176	5.18
	0.9	T30B30,T15B15	338(131,97,110)	4.6	T30B30	259	8.28

example, among the 150 parameter configurations defining a genetic model, there were 16 that had as an optimal design T10B10 ED, T1T1 CH and B1B1 CL sib-pair types. When residual correlation increases (decreases), the marginal counts of discordant pairs shift toward less (more) stringent thresholds. This general pattern corroborates the specific results seen in Tables 3 and 4. Considering discordant and concordant selection jointly, we observe that the majority of counts occur along the diagonal at $\rho = 0.1$, while most of the counts are located below the diagonal at $\rho = 0.4$. Consequently, in the presence of positive residual correlation we should not plan studies that combine extreme discordant pairs with less extreme concordant sib-pairs.

Similar summary counts in Table 6 are stratified by low ($R = 0.02, 0.1$) and high ($R = 1, 10, 50$) phenotype-to-genotype costs . When phenotyping cost is relatively low,

Table 5: Counts of optimal mixture selection schemes among all 150 possible selection combinations when residual correlation $\rho = 0.1$ (top) and $\rho = 0.4$ (bottom). Other parameters are $H = 0.1, 0.3$, $p = 0.1, 0.3, 0.5, 0.7, 0.9$, the cost ratio of phenotype-to-genotype $R = 0.02, 0.1, 1, 10, 50$, under recessive, dominant, and additive gene action. Power = 0.8, α is 0.001. Under CHCL is shown the selection scheme for CH and CL pairs; for example, T10B10 means T10T10 for CH pairs, and B10B10 for CL pairs.

	$\rho = 0.1$							
	CHCL							
ED	T1B1	T3B3	T5B5	T10B10	T15B15	T20B20	T25B25	Sum
T10B10	16	0	14	2	0	0	0	32
T10B20	0	1	2	0	0	0	0	3
T10B25	4	1	2	14	1	0	0	22
T15B15	2	1	3	11	0	0	0	17
T15B25	0	0	0	2	5	8	0	15
T20B20	0	0	1	5	0	0	0	6
T25B25	0	0	0	3	5	11	0	19
T30B30	0	0	0	0	1	5	30	36
Sum	22	3	22	37	12	24	30	150
	$\rho = 0.4$							
	CHCL							
ED	T1B1	T3B3	T5B5	T10B10	T15B15	T20B20	T25B25	Sum
T10B10	14	0	1	0	0	0	0	15
T10B20	1	0	0	0	0	0	0	1
T10B25	11	1	3	0	0	0	0	15
T15B15	1	4	1	0	0	0	0	6
T15B25	2	0	6	5	0	0	0	13
T20B20	8	1	3	3	0	0	0	15
T25B25	2	0	10	3	1	0	0	16
T30B30	0	1	0	6	27	12	23	69
Sum	39	7	24	17	28	12	23	150

fairly stringent discordant and concordant sib-pairs (upper left region) should be collected. Conversely, when phenotyping cost is relatively high, less stringent conditions are indicated. Of course, these observations are general guidelines; more specific designs are possible with more information other than just the phenotype-to-genotype cost ratio. However, in cases when very little is known about the underlying genetic factors, one may not know more than the costs involved.

Lastly, we summarize the comparison of costs between the optimal mixed sample and optimal ED-only sample; Figure 3 gives an overview. The y-axis represents ED:mixed cost ratio, and the x-axis indexes an ordered set of parameters as given below:

for R=(0.02, 0.1, 1, 10, 50)
  for H=(0.1, 0.3)

Table 6: Counts of optimal mixture selection schemes among all 150 possible selection combinations when phenotype-to-genotype cost ratio $R = 0.02, 0.1$ (top) and $R = 1, 10, 50$ (bottom). Other parameters are $H = 0.1, 0.3$; $p = 0.1, 0.3, 0.5, 0.7, 0.9$; $\rho = 0.2, 0.4$; recessive, dominant, and additive gene action. Nominal power = 0.8 and $\alpha = 0.001$. Under CHCL is shown the selection scheme for CH and CL pairs; for example, T10B10 means T10T10 for CH pairs, and B10B10 for CL pairs.

	$R = 0.02, 0.1$							
				CHCL				
ED	T1B1	T3B3	T5B5	T10B10	T15B15	T20B20	T25B25	Sum
T10B10	16	0	15	2	0	0	0	33
T10B20	0	1	2	0	0	0	0	3
T10B25	6	2	5	8	0	0	0	21
T15B15	3	5	4	8	0	0	0	20
T15B25	2	0	6	2	0	0	0	10
T20B20	8	1	4	4	0	0	0	17
T25B25	2	0	10	3	0	0	0	15
T30B30	0	1	0	0	0	0	0	1
Sum	37	10	46	27	0	0	0	120
	$R = 1, 10, 50$							
				CHCL				
ED	T1B1	T3B3	T5B5	T10B10	T15B15	T20B20	T25B25	Sum
T10B10	14	0	0	0	0	0	0	14
T10B20	1	0	0	0	0	0	0	1
T10B25	9	0	0	6	1	0	0	16
T15B15	0	0	0	3	0	0	0	3
T15B25	0	0	0	5	5	8	0	18
T20B20	0	0	0	4	0	0	0	4
T25B25	0	0	0	3	6	11	0	23
T30B30	0	0	0	6	28	17	53	104
Sum	24	0	0	27	40	36	53	180

for Mode=(recessive, additive)
  for ρ=(0.1, 0.3)
    for p=(0.1, 0.3, 0.5, 0.7, 0.9)

For example, the first 5 points correspond to R=0.02, H=0.1, recessive gene action, ρ=0.1 and p=0.1, 0.3, 0.5, 0.7 or 0.9; the next 5 points correspond to R=0.02, H=0.1, recessive gene action, ρ=0.3 and p=0.1, 0.3, 0.5, 0.7 or 0.9, and so on. Although the costs for both designs increase as R increases, the ratio ED:mixed is generally between 1 and 2. In plot (a), the 10 pairs of high-low peaks correspond to the 20 combinations of R, mode-of-inheritance, and ρ. Within each pair the decrease reflects increases in ρ; across pairs the peak magnitudes reflect changes in mode-of-inheritance. When the model is recessive with infrequent allele $(1 - p)$, the optimal cost from the test with

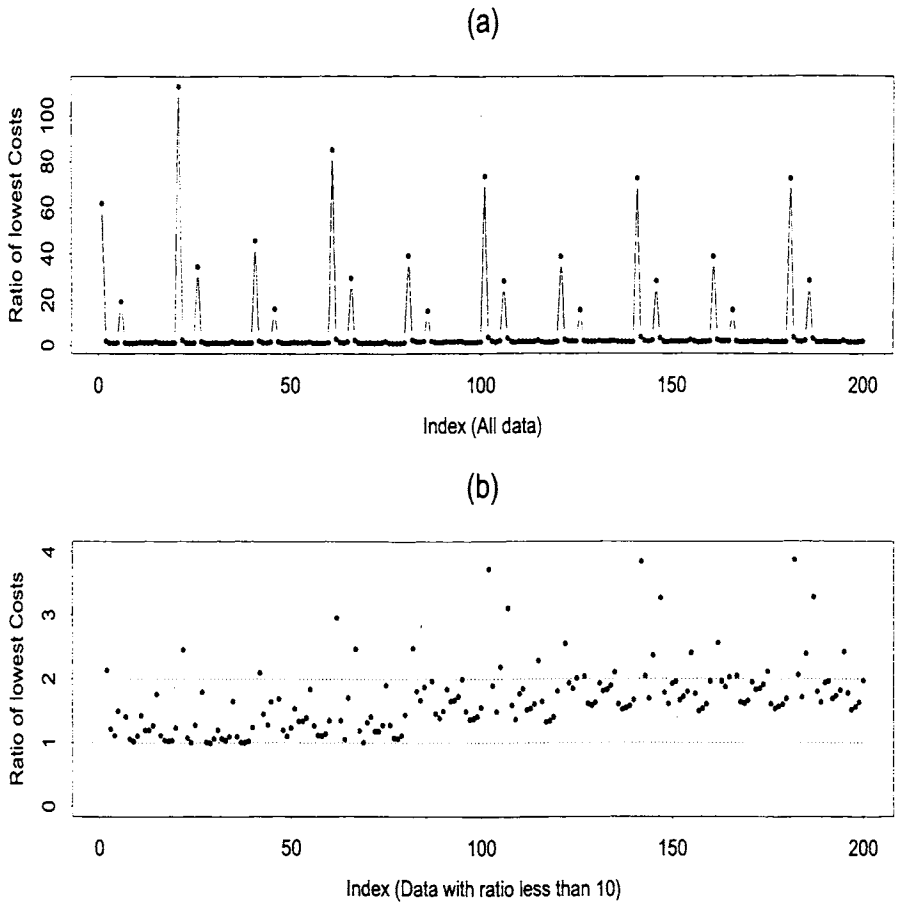


Figure 3: The ratio of lowest cost from the test using ED sib-pairs alone to the lowest cost of the mixture test. The lowest cost is chosen from all possible selection schemes under each genetic models: $H = 0.1, 0.3$, recessive, additive, $\rho = 0.1, 0.4$, and $p = 0.1, 0.3, 0.5, 0.7, 0.9$. Power = 0.8 and $\alpha = 0.001$. (a) Plot of ratios of all available data. (b) Plot of ratios below 10.

ED pairs alone can attain 20 – 100-fold increases over the mixed sample. A similar conclusion holds for a dominant model with frequent allele (p). More information about the low ratios is shown in plot (b), where the high ratios (greater than 10) are not shown. In some cases the ratios are close to 1, especially when $R = 0.02$ or $R = 0.1$.

6 Discussion

Extremely discordant and extremely concordant (high or low) sib-pairs are among the most useful sib-pairs in genetic linkage analysis of quantitative trait loci (Risch and

Zhang [13]). Essentially, selecting sib-pairs mimics a designed experiment whereby known genotypes are typically compared by phenotypic averages (*e.g.*, analysis of variance). Comparing averages (first moments) is much more powerful than analyzing variances (second moments), as is the basis of the Haseman and Elston [11] robust sib-pair method for linkage. By selecting extreme discordant or concordant sib-pairs we are enriching the sample with individuals that are more likely to be in the tails of the phenotypic distribution because of genotype rather than chance.

Gu *et al.* [10] extended the mean IBD test of Risch and Zhang to incorporate the three types of pairs into a single test (EDAC). In this paper, we show the advantages of the score test for linkage, developed by Dudoit and Speed [5, 6], when multiple selection schemes are possible. Under the mixed selection strategy, the score test provides more power than the EDAC test by weighting each kind of sib-pair according to its linkage information under an assumed genetic model. As the basis for inclusion is the underlying biological mechanisms, it is not surprising that the score test performs better than an alternative that combines test statistics largely on the basis of statistical principles, although the latter has been shown to significantly increase power over unselected sib-pairs.

Compared with the ED-only selection scheme, the mixture selection scheme not only makes better use of the screening process it is also more cost efficient. Considerable savings in cost are seen under recessive and dominant modes of inheritance. Residual correlation between sib-pairs plays a key role in the optimal design of selected samples. At higher degrees of correlation (perhaps larger than 0.3-0.4) discordant pairs become increasingly difficult to obtain. Therefore, the threshold for ED pairs should be accordingly relaxed. Conversely, the threshold for concordant pairs may be more stringent. The results shown in Tables 3, 4 and 5 provide some useful guidelines when the cost ratio (R) of phenotype-to-genotype is known. More specific guidelines are possible when there is knowledge of residual correlation.

We have assumed that the trait locus and marker locus are in complete linkage, but this is not an unrealistic assumption as more and more genetic markers are available for many organisms. Also, we have set a conservative type I error probability of $\alpha = 0.001$, as discussed Lander and Kruglyak [12], to more closely resemble a "search" for trait loci, whether by a scan or a relatively large panel of candidate genes. An error rate of $\alpha = 0.01$ or $\alpha = 0.0001$ gives the same basic patterns in optimal designs as discussed in the text. The IBD mean test of Risch and Zhang and more general methods as developed by Dudoit and Speed are needed to more faithfully reflect the reality of genetic epidemiology studies. Analyzing genotypes conditional on phenotypes provides a realistic framework under which to study genetic traits. The specific assumptions underlying the Dudoit-Speed score test allow one to evaluate the appropriate use of the method in any given situation. This is an important step when assessing the validity of study results, especially in observational studies.

Dedication

This paper is dedicated to Terry Speed. As thesis advisor he provided unending support and motivation; learning from him was truly inspirational. At this particular time in the development of statistical methods for statistical genetics and bioinformatics we are indeed fortunate to have Terry play a major role in shaping the field. As an occasional lone voice in the desert, he reminds us that we are trying to solve real problems that more often than not require thinking outside the box. This is perhaps the most important thing I learned from him – solve the problem. I am privileged to have his attention as friend, colleague, and advisor. – *Rudy Guerra*

Rudy Guerra, Program in Biostatistics, Department of Statistics, Rice University, Houston rguerra@rice.edu

Jian Han, Biostatistics, PPD, Inc., Austin jian.han@austin.ppdi.com

References

[1] D. B. Allison, M. Heo, N. J. Schork, S-L Wong, and R. C. Elston. Extreme selection strategies in gene mapping studies of oligogenic traits do not always increase power. *Human Heredity*, 48:97–107, 1998.

[2] C. I. Amos and R. Guerra. Statistics in human genetics. In S. Kotz, D. Banks, and C. Read, editors, *Encyclopedia of Statistical Science, Update*, volume 3, pages 334–346. Wiley, New York, 1998.

[3] W. C. Blackwelder and R. C. Elston. A comparison of sib-pair linkage tests for disease suscpetibility loci. *Genetic Epidemiology*, 2:85–97, 1985.

[4] G. Carey and J. A. Williamson. Linkage analysis of quantitative trait: increased power by using selected samples. *American Journal of Human Genetics*, 49:786–796, 1991.

[5] S. Dudoit and T. P. Speed. A score test for linkage using identity by descent data from sibships. *Annals of Statistics*, 27:943–986, 1999.

[6] S. Dudoit and T. P. Speed. A score test for the linkage analysis of quantitative and qualitative traits based on identity by descent data from sib-pairs. *Biostatistics*, 1:1–26, 2000.

[7] L. Eaves and J. Meyer. Locating human quantitative trait loci: Guidelines for the selection of sibling pairs for genotyping. *Behavior Genetics*, 24:443–455, 1994.

[8] D. R. Goldstein, S. Dudoit, and T. P. Speed. Power and robustness of a score test for linkage analysis of quantitative traits using identity by descent data on sib pairs. *Genetic Epidemiology*, 20:415–431, 2001.

[9] C. Gu and D. C. Rao. Linkage strategy for detection of human quantitative-trait loci. ii optimization of study design based on extreme sib pairs and generalized relative risk ratios. *American Journal of Human Genetics*, 61:211–222, 1997.

[10] C. Gu, A. Todorov, and D. C. Rao. Combining extremely concordant sibpairs with extremely discordant sibpairs provides a cost effective way to linkage analysis of quantitative trait loci. *Genetic Epidemiology*, 13:513–533, 1996.

[11] J. K. Haseman and R. C. Elston. The investigation of linkage between a quantitative trait and a marker locus. *Behavior Genetics*, 2:3–19, 1972.

[12] E. Lander and L. Kruglyak. Genetic dissection of complex traits: Guidlines for interpreting and reporting linkage results. *Nature Genetics*, 11:241–247, 1995.

[13] N. Risch and H. Zhang. Extreme discordant sib pairs for mapping quantitative trait loci in humans. *Science*, 268:1584–1589, 1995.

[14] N. Risch and H. Zhang. Mapping quantitative trait loci with extreme discordant sib pairs: sampling considerations. *American Journal of Human Genetics*, 58:836–843, 1996.

[15] H. Zhao, H. Zhang, and J. Rotter. Cost-effective sib-pair designs in the mapping of quantitative-trait loci. *American Journal of Human Genetics*, 60:1211–1221, 1997.

Multipoint Fine-scale Linkage Disequilibrium Mapping: Importance of Modeling Background LD

Andrew L. Strahs and Mary Sara McPeek

Abstract

In linkage disequilibrium (LD) mapping, use of information on multiple markers simultaneously is expected to lead to greater power to detect association and smaller confidence intervals (CIs) for the location of the variant of interest than would be obtained from single-point analysis. Among the important challenges facing case-control LD mapping methods are (i) even when an appropriate control sample is available, there may be background LD in the control sample which must be taken into account in the analysis, especially when fine-scale data are collected, and (ii) in practice, genotype rather than haplotype data are often available, limiting the applicability of some methods. Furthermore, in cases when genotype data can, in principle, be incorporated, it can be computationally challenging. We focus on simultaneous solution of these problems in the context of the Decay of Haplotype Sharing (DHS) method. We develop a computationally efficient method that allows for genotype or haplotype data on many loci and incorporates background LD based on a Markov model of order η. The case of a Markov model of order 2 is implemented in free software. In addition, we demonstrate that failure to adequately model background LD can potentially have a major effect on the analysis, and we develop and apply methods for assessing the adequacy of the model for background LD.

Keywords: Decay of Haplotype Sharing; linkage disequilibrium; fine-scale mapping; background linkage disequilibrium; cystic fibrosis; hidden Markov model

1 Introduction

Linkage disequilibrium (LD) has been shown to be useful for fine-mapping of trait-associated variants [6, 10, 11, 15]. While early approaches generally treated each marker separately, haplotype-based LD mapping methods have the potential to provide considerable additional information when dense marker data are available in a region. There are several approaches that combine results across loci in various ways without explicitly modeling dependence among loci [4, 7, 17, 23, 31, 32]. Among approaches that explicitly model dependence across loci, Service *et al.* [29] and MacLean *et al.*

[20] perform haplotype-based tests for association, in which they use multilocus models for haplotypes descended from an ancestor, taking into account recombination and mutation. They require haplotype data, assume background linkage equilibrium, and do not consider effects of population structure. There are several methods that perform haplotype-based tests of association in trios consisting of parents and an affected offspring, conditional on the transmitted and non-transmitted haplotypes in the parents (*e.g.* [2, 3, 36]). Lam *et al.* [16] use a parsimony method to build an evolutionary tree of disease haplotypes assuming a disease mutation occurs in an intermarker interval. They then compute the likelihood of the tree using a model for recombination and mutation. They obtain a posterior distribution for the location of the variant. Their method assumes haplotype data. Background linkage disequilibrium is taken into account by a Markov-type method in which the lag at any stage is chosen to coincide with the longest match in the control database.

McPeek and Strahs [22] form a confidence interval (CI) for the location of the variant, in which they make use of a multilocus decay-of-haplotype-sharing (DHS) model for haplotypes descended from an ancestor, taking into account recombination and mutation. They propose a quasi-likelihood approach to take into account population structure in the affecteds. Assuming a conditional coalescent model for the population structure, McPeek and Strahs [22] derive an approximate correction factor for the likelihood, and they model background LD by a Markov chain with lag 1. Morris *et al.* [25] concentrate on biallelic markers in a Bayesian framework and obtain the posterior distribution of the location of the trait-associated variant using Markov chain Monte Carlo (MCMC). They use a similar approach to that of McPeek and Strahs [22] to correct for population structure. Rannala and Reeve [28] also employ a Bayesian framework and obtain the posterior density of the position of the trait-associated variant by employing MCMC to integrate over coalescent genealogy trees, using biallelic marker data and information about candidate genes from an annotated human genome sequence. Neither Morris *et al.* [25] nor Rannala and Reeve [28] consider background LD.

Liu *et al.* [18] also obtain the posterior distribution of the location of the trait-associated variant using MCMC. Their model for population structure groups the disease haplotypes into clusters corresponding to different ancestral haplotypes and assumes a star-shaped genealogy for the haplotypes within each cluster given the ancestral haplotype. They model background LD by a Markov chain with lag 1. Zhang and Zhao [34] extend McPeek and Strahs [22] by implementing a stepwise mutation model for mutation in microsatellite markers. They extend the conditional coalescent model of McPeek and Strahs [22] to allow variable population size. Morris *et al.* [24] obtain the posterior distribution of location of the trait-associated variant and incorporate a shattered coalescent model for genealogies of the disease haplotypes using MCMC. They also model background LD using a Markov chain with lag 1.

In this study, we simultaneously tackle two of the major difficulties that arise in multipoint LD mapping with data on random samples of cases and controls: (1) LD is generally present in the controls as well as in the affecteds, and this background LD

can have a major impact on the analysis; (2) data are typically in the form of genotypes with unknown phase, rather than haplotypes. Dealing with both of these issues simultaneously presents particular computational challenges, and a focus of our work has been development of an efficient algorithm to handle them.

To deal with the second problem, that data are typically in the form of (unphased) genotypes, one possible solution is to try to reconstruct haplotypes based on population information, using one of the available methods [9, 12, 18, 19, 30]. We prefer instead to incorporate uncertainty about the haplotypes into the analysis. The likelihood framework of DHS makes an extension from haplotype to genotype data straightforward in principle: one need only sum the likelihoods of all possible sets of haplotypes compatible with the observed genotype data. This approach is implemented by Zhang and Zhao [35]. However, with more than a small number of loci, this straightforward approach quickly becomes computationally infeasible. We introduce a more computationally efficient approach using a hidden Markov model (HMM), in which we incorporate a Markov model, with lag η, for LD in the controls.

Methods for LD mapping generally consider two statistical problems, detection of association (*i.e.* hypothesis testing) and localization (*i.e.* construction of a CI for the variant of interest). If background LD, *i.e.* LD present in the controls as well as in the affecteds, is not adequately captured by the model, it may be falsely attributed to the presence of a variant associated with the trait. For the detection problem, unmodeled background LD could result in excess false positive detections of association. For the localization problem, unmodeled background LD could result in CIs that fail to have the appropriate probability of covering the true location. For the detection problem, a number of approaches have been developed that aim to produce valid hypothesis tests in the presence of background LD [2, 8, 27, 36]. Here we instead focus on the localization problem, which is not treated by these papers. In this context, McPeek and Strahs [22] model background LD in control haplotypes by use of a Markov chain model of lag $\eta = 1$. In analyzing the data set of Kerem *et al.* [15], we find that a Markov chain of lag $\eta = 2$ is preferable, as shown in the subsection "Importance of modeling background LD" of the Results section. Incorporation of Markov models for background LD is more challenging when genotype data are used instead of haplotype data, because implementation of a Markov model requires one to keep track of phase information. In this study, we devise a HMM to simultaneously incorporate genotype data and a Markov model with lag $\eta = 2$ for background LD.

These new methods make it feasible to perform multipoint LD mapping on data sets consisting of (unphased) genotypes for a large number of markers. We use simulated examples to compare fine-mapping based on genotype and haplotype data, and we use the CF data set [15] to demonstrate the importance of the improved modeling of background LD.

2 Methods

DHS method for haplotype data

As developed in McPeek and Strahs [22], the DHS likelihood under the model for a single observed haplotype $\mathbf{h}_{\text{obs}}$ drawn from the population of affecteds, when there is one ancestral haplotype $\mathbf{h}_{\text{anc}}$, is

$$L(x, \mathbf{h}_{\text{anc}}, \tau^{-1}, p; \mathbf{h}_{\text{obs}}) = (1-p)\tilde{L}(x, \mathbf{h}_{\text{anc}}, \tau^{-1}; \mathbf{h}_{\text{obs}}) + pP_{\text{null}}(\mathbf{h}_{\text{obs}}), \tag{1}$$

where x is the location of the variant; τ is number of generations to the ancestor, or, equivalently, τ^{-1} is a measure of the amount of linkage disequilibrium and is equal to the expected genetic distance from the variant to either edge of the ancestral segment in an observed haplotype; p is a heterogeneity parameter representing the probability that the haplotype $\mathbf{h}_{\text{obs}}$ is not descended from the ancestral haplotype $\mathbf{h}_{\text{anc}}$; and $P_{\text{null}}(\mathbf{h})$ is the frequency of haplotype $\mathbf{h}$ in the control population. Furthermore,

$$\begin{aligned}\tilde{L}(x, \mathbf{h}_{\text{anc}}, \tau^{-1}; \mathbf{h}_{\text{obs}}) \;=\; & \sum_{i=0}^{l_{re}} \sum_{j=0}^{l_{le}} \Bigg[g(\tau^{-1}, -j, i) \times \prod_{k=-j}^{i} m(k, \tau, \mathbf{h}_{\text{anc}}(k), \mathbf{h}_{\text{obs}}(k)) \times \\ & P_{null}(\mathbf{h}_{\text{obs}}(i+1), \mathbf{h}_{\text{obs}}(i+2), \ldots, \mathbf{h}_{\text{obs}}(l_{re})) \times \\ & P_{null}(\mathbf{h}_{\text{obs}}(-l_{le}), \mathbf{h}_{\text{obs}}(-l_{le}+1), \ldots, \mathbf{h}_{\text{obs}}(-j-1)) \Bigg]\end{aligned}$$

is the likelihood assuming that observed haplotype $\mathbf{h}_{\text{obs}}$ is a τth-generation descendent of ancestral haplotype $\mathbf{h}_{\text{anc}}$. In the above expression, $m(k, \tau, \alpha, \beta)$ models the mutation process; it is the probability that allele β is observed at locus k, given that the haplotype's τth generation ancestor at locus k had allele α. i and j index markers, with marker 0 corresponding to the putative location x of the variant, and with markers on, say, the distal side of x numbered with consecutive negative integers decreasing in the direction away from the centromere and with markers on the proximal side of x numbered with consecutive positive integers increasing in the direction of the centromere. (Note that the integer labels for the markers are defined relative to the putative position x of the variant, which varies across the region during the analysis.) Here, $-l_{le}$ is the index corresponding to the "left edge" of the data set (*i.e.* the marker farthest from the centromere), and l_{re} is the index corresponding to the "right edge" of the data set (*i.e.* the marker closest to the centromere). In the above expression, we sum over all possible choices of the marker intervals containing the two (unobserved) breakpoints of the ancestral segment. Moreover,

$$g(\tau^{-1}, -j, i) = e^{-\tau d_{-j,i}}(1 - e^{-\tau d_{-j-1,-j}})(1 - e^{-\tau d_{i,i+1}})$$

is the probability that $\mathbf{h}_{\text{obs}}$ inherits the variant and the ancestral segment, intact, between loci $-j$ and i inclusive but that it is no longer intact at locus $-j-1$ nor at locus $i+1$.

Here, $d_{i,j}$ is the genetic distance between loci i and j. McPeek and Strahs [22] discuss how to incorporate multiple ancestral haplotypes into this likelihood expression.

To combine likelihoods across observed haplotypes, one must make some assumptions about how the haplotypes are related. Under the assumption of independent recombinational histories (*i.e.* a star-shaped phylogeny), one can multiply the likelihoods across haplotypes. This approach is generally anti-conservative when this assumption does not hold. McPeek and Strahs [22] propose a quasi-likelihood approach to take into account population structure, which in principle could be applied to any chosen population model. For the case in which the affecteds are presumed to be only very distantly related with little else known about the population structure, McPeek and Strahs [22] propose a conditional coalescent model for the phylogeny relating the affected individuals, conditional on the time to the common ancestor. With complete data, they calculate and maximize a quasi-likelihood with respect to this model, and with incomplete data, they calculate and maximize a similar expression with the complete data likelihoods replaced by incomplete data likelihoods. We employ the same approach here. In the case of the conditional coalescent model or any other exchangeable population model, the parameter estimates obtained in this way are the same as under the assumption of independence, but with the standard errors for the parameters inflated and the log-likelihood deflated. When DHS is used to fine-map, this widens the CI for the location of the trait-associated variant. In practice, then, to implement the conditional coalescent model, we proceed as if the observations were independent, and then implement the appropriate correction to the log-likelihood and standard errors at the end. The approximate correction factor in the conditional coalescent case is

$$\sum_{k=1}^{n-2}\left(2(n-2)!(n+1)/[(n-1)(n-k+1)(n-k+2)(n-k)(k-1)!(n-k-2)!]\times \sum_{i=1}^{\infty}(-1)^{i+1}\binom{n+i-1}{n-k}^{-1}\right),$$

which corrects a typo in McPeek and Strahs [22] (factor of $(-1)^i$ vs. $(-1)^{i+1}$). The DHS model can also be extended to the case when population structure is known [33]. In that case, the shape of the likelihood curve and, in particular, the maximum, will generally not be the same when population structure is taken into account as when independence is assumed.

The formulae of this section give a mathematical representation of the likelihood. However, for computational efficiency in calculating and maximizing the likelihood, we reformulate the probability model as a hidden Markov model (HMM) in the subsection "HMM for haplotype data, with Markov(η) model for background LD" below.

Uncertainty in ancestral haplotype is incorporated in CI construction

To construct a CI for the location of the variant, McPeek and Strahs [22] invert an

(approximate) likelihood-ratio test. At each putative location x, their approximate log-likelihood is maximized over $\mathbf{h}_{\text{anc}}$, $1/\tau$ and p, assuming that the variant is located at x. (The properties of the profile likelihood are discussed in McCullagh and Nelder [21].) The CI is then based on a comparison of the highest maximized log-likelihood to the maximized log-likelihoods at other locations. We emphasize that inference about variant location is **not** performed conditional on the maximizing value of ancestral haplotype. The mapping approach of McPeek and Strahs [22] does, in fact, take into account the uncertainty in ancestral haplotype.

Mode of inheritance, mutation, and background LD

Implicit in the method given in the previous sub-section is the assumption of a multiplicative model for the mode of inheritance, similar to that described by Morris *et al.* [25]. Where β_k corresponds to allele k,

$$P\{\text{affected} \mid (G_1, G_2) = (i, j)\} = \beta_i \beta_j$$

for an individual with genotype (i, j) at the variant. The multiplicative model has the recessive model as a special case, but also allows heterogeneity. This model is convenient when one does not have the information of how the haplotypes are paired. When that information is available, one could easily implement some other mode of inheritance in the analysis.

The mutation model we use is the same as that given by McPeek and Strahs [22]. For biallelic loci, this amounts to assuming the same rate of mutation between the two alleles. We assume the same mutation rate at all markers. These assumptions can be easily modified [34].

When choosing mutation rates to use in the DHS analysis, it may not be appropriate to use a rate as low as the value of $\approx 10^{-9} - 10^{-8}$ given for SNP loci by Nielsen [26]. The reason is that only SNPs that are polymorphic across the individuals in the data set are chosen for analysis. Therefore, the ascertainment process for the data set insures the existence of at least 1 mutation at the SNP within the time-frame of the coalescence of the study sample at that SNP. Thus, conditional on a SNP being in the data set, its mutation rate over the time since the most recent common ancestor of the variant is substantially increased over the unconditional mutation rate given by Nielsen [26]. The extent of the increase depends on assumptions about the population history, but the conditional mutation rate could be several orders of magnitude larger than the unconditional mutation rate. Specification of a larger mutation rate would be expected to lead to a more conservative analysis. In our analysis of the CF data set, we use a mutation rate of 1×10^{-4} mutations per meiosis per marker. Note that in contrast to SNPs, microsatellites would be less affected by this selection effect. The mutation rate of a microsatellite is typically sufficiently high that its conditional mutation rate, over the time period since the most recent common ancestor of the variant, conditional on it being polymorphic in the study sample is very close to its unconditional mutation rate.

In expression (1), the model for background LD enters the likelihood through $P_{null}(\mathbf{h})$, which gives the frequency of haplotype $\mathbf{h}$ in the control population. In principle, one could think of leaving the control haplotype frequencies unconstrained (beyond the requirement that they sum to 1) when estimating them from data. However, with m loci, the number of parameters is $2^m - 1$ for SNP data (with many more for microsatellite data), and the size of the control sample available to estimate these parameters is typically small. Our approach is to constrain the control haplotype frequency distribution to be Markov (η), *i.e.* for a given lag η, we require $P_{null}\{\mathbf{h}(t) = i_t | \mathbf{h}(t-s) = i_{t-s}, \ldots, \mathbf{h}(t-1) = i_{t-1}\} = P_{null}\{\mathbf{h}(t) = i_t | \mathbf{h}(t-\eta) = i_{t-\eta}, \ldots, \mathbf{h}(t-1) = i_{t-1}\}$ for all s such that $\eta \leqslant s \leqslant t + l_e$, and all choices $i_{t-s}, \ldots, i_t$ for alleles, where $\mathbf{h}(t)$ is the allele at locus t in the haplotype $\mathbf{h}$. Such a Markov model can be useful as a simple tool for capturing the local dependence structure among loci on haplotypes randomly selected from a control population. McPeek and Strahs [22] implement the case $\eta = 1$ when complete haplotype data are available. Here, we implement the cases $\eta = 1$ and $\eta = 2$, when either haplotype or genotype data are available, by means of a HMM.

HMM for haplotype data, with Markov(η) model for background LD

For computational efficiency in calculating and maximizing the likelihood, we reformulate, as a HMM [1], the probability model of sub-section "DHS method for haplotype data". The HMM we give here differs from the one given in McPeek and Strahs [22]. Although the underlying likelihood is the same, the HMM given here has computational advantages over the previous version. These advantages are related to the extension of the HMM to allow background LD to be modeled by a Markov model with lag $\eta > 1$, which is given at the end of this sub-section. In section "Extension of HMM to genotype data", we extend this HMM to the case when only genotype data are available.

Suppose that, as in the previous subsection, the putative location of the variant of interest is labeled marker 0, and the markers are numbered with consecutive positive integers increasing in the direction of the centromere and consecutive negative integers decreasing in the direction away from the centromere. We define a discrete-time Markov chain $\{Q_l, 0 \leqslant l \leqslant l_{re}\}$, where l indexes loci on the centromeric side of the variant. The state space of $\{Q_l\}$ is $\{A, N\}$, where A stands for "ancestral" and N stands for "non-ancestral." We define the event $\{Q_l = A\}$ to occur when the entire segment between locus l and the variant of interest (locus 0) is inherited, unbroken by crossovers, from the ancestral haplotype. Note that $\{Q_l = A\}$ holds in this case even if one or more mutations have occurred at locus l (or elsewhere in the segment) in the time since the ancestor. We define $\{Q_l = N\} = \{Q_l = A\}^c$. The initial distribution of $\{Q_l\}$ is $P\{Q_0 = N\} = p = 1 - P\{Q_0 = A\}$. The transition probability matrix for $\{Q_l\}$ is given in Table 1a. In fact, we find it convenient to reverse the conditioning of Q. That is, we define the initial distribution to be $P\{Q_{l_{re}} = A\} = (1-p)e^{-\tau d_{0,l_{re}}} = 1 - P\{Q_{l_{re}} = N\}$, and we use the one-step transition probability matrix $P\{Q_l | Q_{l+1}\}$ given in Table 1b.

The resulting process $\{Q_l\}$ has the same distribution as before. The computational convenience of the reverse conditioning is related to the modeling of background LD and is explained below after the observation distribution is introduced. We can define a mirror-image Markov chain for the loci on the distal side of the variant, with the two chains conditionally independent given Q_0.

Table 1a: Transition probability matrix $P(Q_l|Q_{l-1})$

Current State	Probability that Next State Entered Is	
	A	N
A	$e^{-\tau d_{l,l+1}}$	$1-e^{-\tau d_{l,l+1}}$
N	0	1

Table 1b: Transition probability matrix $P(Q_l|Q_{l+1})$

Current State	Probability that Next State Entered Is	
	A	N
A	1	0
N	$\frac{(1-p)(e^{-\tau d_{l+1,0}}-e^{-\tau d_{l,0}})}{1-(1-p)e^{-\tau d_{l,0}}}$	$\frac{1-(1-p)e^{-\tau d_{l+1,0}}}{1-(1-p)e^{-\tau d_{l,0}}}$

Consider the observation sequence $\{O_l, 0 \leqslant l \leqslant l_{re}\}$ associated with the Markov chain $\{Q_l, 0 \leqslant l \leqslant l_{re}\}$, where O_l is the observed allele at locus l. Our formulation of the distribution of O_l conditional on Q_l depends on our model for background LD as well as on our model for mutations. For simplicity of exposition, we first assume background linkage equilibrium. In that case,

$$P\{O_l \mid Q_l\} = \begin{cases} m(l, \tau, \mathbf{h}_{anc}(l), O_l) & \text{if } Q_l = A \\ f_l(O_l) & \text{if } Q_l = N, \end{cases} \tag{2}$$

where $f_l(\alpha)$ is the frequency of allele α at locus l in the controls. We can allow the observed allele at locus l to be missing by setting $P\{O_l \mid Q_l\} = 1$ when O_l is missing. This will yield the appropriate likelihood calculation for the case when the event that O_l is missing is independent of Q_l.

We now relax the assumption of background linkage equilibrium. Assuming a Markov(1) model for background LD, the observation distribution for $0 \leqslant l < l_{re}$ is given by

$$P\{O_l \mid Q_l, O_{l+1}\} = \begin{cases} m(l, \tau, \mathbf{h}_{anc}(l), O_l) & \text{if } Q_l = A \\ f_{l,l+1}(O_l \mid O_{l+1}) & \text{if } Q_l = N, \end{cases} \tag{3}$$

where $f_{l,l+1}(\alpha|\beta)$ is the conditional frequency, in the controls, of allele α at locus l given allele β at locus $l+1$. If O_l is missing, we set $P\{O_l \mid Q_l, O_{l+1}\} = 1$, and if O_{l+1}

is missing, we set $P\{O_l \mid Q_l, O_{l+1}\}$ equal to expression (2). $P\{O_{l_{re}} | Q_{l_{re}}\}$ remains the same as in expression (2). Under our model, when $0 \leqslant l < l_{re}$, $P\{O_l \mid Q_l, Q_{l+1}, O_{l+1}\} = P\{O_l \mid Q_l, O_{l+1}\}$, which does not depend on Q_{l+1}. Note that if we had conditioned in the other direction, $P\{O_l \mid Q_l, Q_{l-1}, O_{l-1}\}$ would depend on Q_{l-1}, and this is the reason for our choice of the direction of conditioning. This is particularly useful for implementing a Markov model of lag $\eta > 1$ for the background LD. In that case, the observation distribution for $0 \leqslant l \leqslant l_{re} - \eta$ becomes

$$P\{O_l \mid Q_l, O_{l+1}, \ldots, O_{l+\eta}\} = \begin{cases} m(l, \tau, \mathbf{h}_{anc}(l), O_l) & \text{if } Q_l = A \\ f_{l,\ldots,l+\eta}(O_l \mid O_{l+1}, \ldots, O_{l+\eta}) & \text{if } Q_l = N, \end{cases} \quad (4)$$

and $P\{O_l \mid Q_l, Q_{l+1}, \ldots, Q_{l+\eta}, O_{l+1}, \ldots, O_{l+\eta}\} = P\{O_l \mid Q_l, O_{l+1}, \ldots, O_{l+\eta}\}$ does not depend on $Q_{l+1}, \ldots, Q_{l+\eta}$. This allows us to extend the model for background LD to Markov of lag $\eta > 1$ without increasing the size of the state space of the hidden Markov chain.

The joint process $\{Q_l, O_l\}$ was so far defined for $0 \leqslant l \leqslant l_{re}$. There is a corresponding mirror-image process defined on $-l_{le} \leqslant l \leqslant 0$. When background linkage equilibrium is assumed, these two processes are conditionally independent given $\{Q_0, O_0\}$. When background LD is modeled by a Markov model of lag $\eta > 0$, the two processes are conditionally independent given $\{Q_0, O_k, \ldots, O_{k+\eta-1}\}$, for any choice of k with $-\eta + 1 \leqslant k \leqslant \eta - 1$. In practice, however, we generally take the position of the variant ($l = 0$) to be in between markers, rather than at a marker, so that O_0 is always missing (this is discussed further in Appendix A). We have developed extensions to the Baum algorithms for likelihood calculation and maximization that are applicable to our model, as outlined in Appendix A.

Extension of HMM to genotype data

In practice, unambiguously-determined haplotype data are often unavailable. Instead, genotype data, in which phase is unknown, are commonly available. We describe an extension of the HMM of the previous sub-section to this case, which allows computationally efficient analysis of data sets involving genotype data on many loci.

Consider the model for multilocus genotype data from a single individual. To simplify the exposition, we first assume background linkage equilibrium. In that case, we consider the Markov chain $\{R_l^G, 0 \leqslant l \leqslant l_{re}\} = \{(Q_l^M, Q_l^P), 0 \leqslant l \leqslant l_{re}\}$, where $\{Q_l^M, 0 \leqslant l \leqslant l_{re}\}$ is the Markov chain of the previous subsection defined for the individual's maternally-inherited haplotype, while $\{Q_l^P, 0 \leqslant l \leqslant l_{re}\}$ is the Markov chain of the previous subsection defined for the individual's paternally-inherited haplotype, with $\{Q_l^M\}$ independent of $\{Q_l^P\}$. The state space of $\{R_l\}$ is $\{A,N\}^2$, the transition probabilities of $\{R_l\}$ are the products of the transition probabilities for the independent chains $\{Q_l^M\}$ and $\{Q_l^P\}$, and the initial distribution of $\{R_l\}$ is similarly obtained from the initial distributions of $\{Q_l^M\}$ and $\{Q_l^P\}$. The observation O_l^G is the genotype data

for the individual at locus l. The two possible phases for the genotype are *a priori* equally likely, so the observation distribution takes a simple form, given in Appendix B. As before, there is a corresponding mirror image process to $\{Q_l, O_l\}$ that is defined on $-l_{le} \leqslant l \leqslant 0$, with the two processes conditionally independent given $\{Q_0, O_0\}$.

We now extend the method to allow the background LD to be modeled by a Markov chain with lag $\eta \geq 1$. In order to implement a Markov model of lag η, we need to retain, at each locus l, the information of the phase of the genotype at l with respect to the genotypes at $l+1, \ldots, l+\eta$. When the genotype at locus l is recorded in a computer file, an arbitrary order of the two alleles at locus l is chosen, and we introduce the random variable Φ_l which represents the arbitrary order of the alleles in the recorded genotype. We assume that $\Phi_l = (M,P)$ or (P,M) with chance 1/2 each, where $\{\Phi_l = (M,P)\}$ denotes the event that the first allele listed in the file is the maternal allele and the second allele listed is the paternal allele, and vice versa for $\{\Phi_l = (P,M)\}$. For each l we define $(Q_l^1, Q_l^2) = (Q_l^M, Q_l^P)$ if $\Phi_l = (M,P)$ and $(Q_l^1, Q_l^2) = (Q_l^P, Q_l^M)$ if $\Phi_l = (P,M)$. That is, Q_l^i is the ancestral state corresponding to the ith recorded allele in the genotype, $i = 1,2$. Furthermore, we define $I_{l,l+1}$ to be the indicator of the event $\{\Phi_l = \Phi_{l+1}\}$, that is, the indicator of the event that the genotypes at loci l and $l+1$ happen to be recorded so that their phase with respect to one another is correct. We define the hidden Markov chain $\{R_l^G, 0 \leqslant l \leqslant l_{re}\}$ by $R_{l_{le}}^G = (Q_{l_{re}}^1, Q_{l_{re}}^2) \in \{A,N\}^2$ and $R_l^G = (Q_l^1, Q_l^2, I_{l,l+1})$ for $0 \leqslant l < l_{re}$, except that when $Q_l^1 = Q_l^2 = A$, the information of $I_{l,l+1}$ is not needed, so we collapse the two states $(A,A,1)$ and $(A,A,0)$ into a single state. The state space for $\{R_l^G, 0 \leqslant l < l_{re}\}$ is thus $\{(A,A)\} \cup [\{(A,N),(N,A),(N,N)\} \times \{0,1\}]$. The initial distribution and transition probabilities for this chain are easily determined, assuming that $\{Q_l^M\}$ and $\{Q_l^P\}$ are independent copies of the Markov chain in the previous sub-section, with the Φ_l i.i.d as given above. The observation distribution is given in Appendix B.

Note that in order to accommodate a Markov model of lag 1 for the background LD, we have increased the size of the state space from 4 to 7. Interestingly, we are able to accommodate a Markov model of lag 2 without any change in the state space. This is because, in the calculation of the observation distribution by the Baum algorithm, both R_l^G and R_{l+1}^G are available to condition on, as well as O_{l+1}^G and O_{l+2}^G, so there is no need to store extra information. As in sub-section "HMM for haplotype data, with Markov(η) model for background LD," we extend to our model the Baum algorithms for likelihood calculation and maximization (see Appendix A).

Choice of η for modeling background LD

For modeling background LD, we have defined a nested sequence of Markov models indexed by the lag η. The question arises as to how η should be chosen in practice. The usual trade-offs apply. In our case, if η is too small, background LD may be erroneously identified as LD with a trait-associated variant, while if η is too large, overfitting will quickly become a problem, as the number of parameters increases exponentially with η.

We view the choice of η as a problem of model selection. We consider two criteria, the Akaike information criterion (AIC) (Akaike 1972) and the Bayesian information criterion (BIC) (Schwartz 1978). Comparable formulations of these criteria are $AIC = -2L + 2k$ and $BIC = -2L + k\log n$, where L is the log-likelihood and k is the number of parameters. For AIC and BIC, the model that minimizes the criterion would be selected. The likelihood component L will always increase with η; both procedures include a penalty for the number of parameters, which offers some protection against overfitting.

In addition to these generic model selection techniques, we also perform an informal diagnostic, suggested by Paul Van Eerdewegh (personal communication), which is more specific to our method. To perform this diagnostic, which we call "mapping in controls", we plug the control haplotypes/genotypes into the mapping program in place of the affecteds' haplotypes/genotypes. We use the same control haplotypes to fit the Markov(η) model for background LD. To assess the results of mapping in controls, we generate the resulting log profile likelihood plot for the location of the variant. Because the same data are used both to estimate the parameters of the model for background LD and for mapping, if the model for background LD is adequate, the procedure should "recognize" these data as fitting the model. The existence of a pronounced peak in the resulting plot would suggest the presence of LD in the controls that is not adequately modeled by the Markov (η) model. If this peak coincides with the peak in the affecteds, then this suggests that the peak in the affecteds may be spurious or at least higher than is warranted. To remedy this, we would try increasing the lag η to capture more of the background LD.

3 Results

Importance of modeling background LD

We demonstrate the importance of modeling background LD in the CF data set of Kerem *et al.* [15]. The data set includes 94 haplotypes from affected individuals and 92 haplotypes from normal individuals. Pairs of haplotypes in individuals are not identified. Each haplotype consists of 23 biallelic markers within a 2-Mb region covering the gene. All physical distances are converted to genetic distances by use of the equivalence 1Mb $\approx$ 1 cM. Note that if the mutation rate were assumed to be 0, then the DHS results would be invariant under a rescaling of distance, *i.e.* 1Mb $\approx k$ cM for any k. When the mutation rate is low, the DHS results would be expected to be robust to deviation of k from 1. We further note that experiments (results not shown) have demonstrated that the method is relatively robust to misspecifications of genetic distances while errors in map order of the marker loci may have a more serious effect.

To demonstrate the importance of modeling background LD in the CF data set of Kerem *et al.* [15], we first perform mapping in the controls. Figure 1 gives the profile log-likelihood curves for the control haplotypes for LE ($\eta = 0$), $\eta = 1$ and $\eta = 2$. For

both $\eta = 0$ and $\eta = 1$, the plot for mapping in the controls is sharply peaked, suggesting that background LD that is present in the controls could be driving some of the mapping results in the affecteds. In contrast, the plot for $\eta = 2$ is completely flat, suggesting that there is little unmodeled background LD detected in the controls, but leaving open the possibility that the Markov(2) model could be overfitting the data. Table 2 gives the results of the AIC and BIC model selection procedures, where we have added constants to AIC and BIC so that each has value 0 for $\eta = 0$. The results indicate that the model with $\eta = 2$ is strongly preferred over $\eta = 0$ and $\eta = 1$ by both the AIC and BIC criteria. This leaves open the possibility that a value of $\eta > 2$ may be optimal according to the AIC and/or BIC criteria. However, the results of our diagnostic in Figure 1 suggest that additional unmodeled background LD, if present, is having little effect on the procedure.

Table 2: Model Selection for background LD in CF data set of Kerem *et al.* [15]

η	log-lik.	# param.	AIC	BIC
0	-1208.325	23	0	0
1	-739.951	45	-892.748	-837.269
2	-628.621	87	-1031.408	-870.013

Figure 2 shows the results of LD mapping when each of the three models for background LD is used. In the case of $\eta = 0$ (linkage equilibrium), the resulting profile log-likelihood curve for the affecteds resembles the profile log-likelihood for the controls given in Figure 1, suggesting that the mapping results in this case may be misleading due to unmodeled background LD. In fact, neither the 95% CI assuming independence of recombinational histories nor that assuming the conditional-coalescent model contain the true location of the variant. The resemblance between the curves for cases and controls is less strong for the case of $\eta = 1$ and non-existent for the case of $\eta = 2$. In both of these cases, the CIs cover the true location of the variant. In this data set, LD around the $\Delta 508$ mutation in the affecteds is very strong relative to the background LD, but if the LD signal were weaker, as might be expected in many data sets, the model for background LD would presumably become even more critical.

Mapping results for genotype vs. haplotype data

The cystic fibrosis (CF) data set of Kerem *et al.* [15] has almost complete haplotype information for affecteds and controls. By randomly combining haplotypes into genotypes and then throwing away the phase information, we can compare the results of LD mapping based on haplotype data with the results when only genotype data are available. Figure 3 gives the profile likelihood curve for one random pairing of affecteds' haplotypes to form genotypes for 47 individuals, where we treat phase as unknown. Here, a Markov model with lag $\eta = 2$ is used to capture background LD in the analysis. For comparison, figure 2 gives the same plot assuming haplotype data are available.

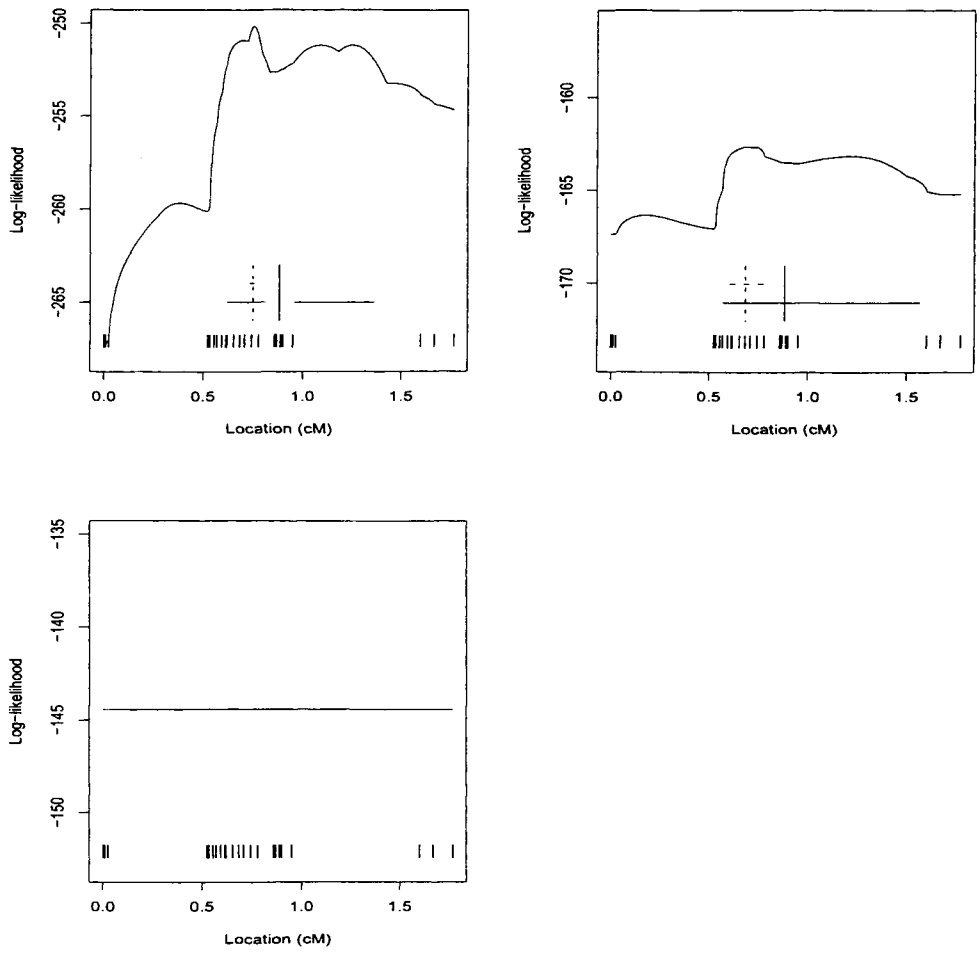


Figure 1: **Results of mapping-in-controls diagnostic** for DHS analysis of CF data set of Kerem *et al.* [15], where background LD is modeled as Markov(η) with $\eta = 0$ (upper left), $\eta = 1$ (upper right), $\eta = 2$ (lower left). Curve gives log profile likelihood vs. (putative) location x of variant, where x is expressed as distance from D21S1885. The dotted vertical line is the estimated variant location in controls, the unbroken vertical line is the true variant location, the dotted horizontal line is the 95% CI when independence of recombinational histories is assumed, and the unbroken horizontal line is the 95% CI when a conditional-coalescent model is assumed. The assumed mutation rate is 10^{-4} mutations per marker per meiosis. The hash marks give the locations of the biallelic markers. (Because the curve is flat for $\eta = 2$, we omit the CIs, both of which cover the entire region.)

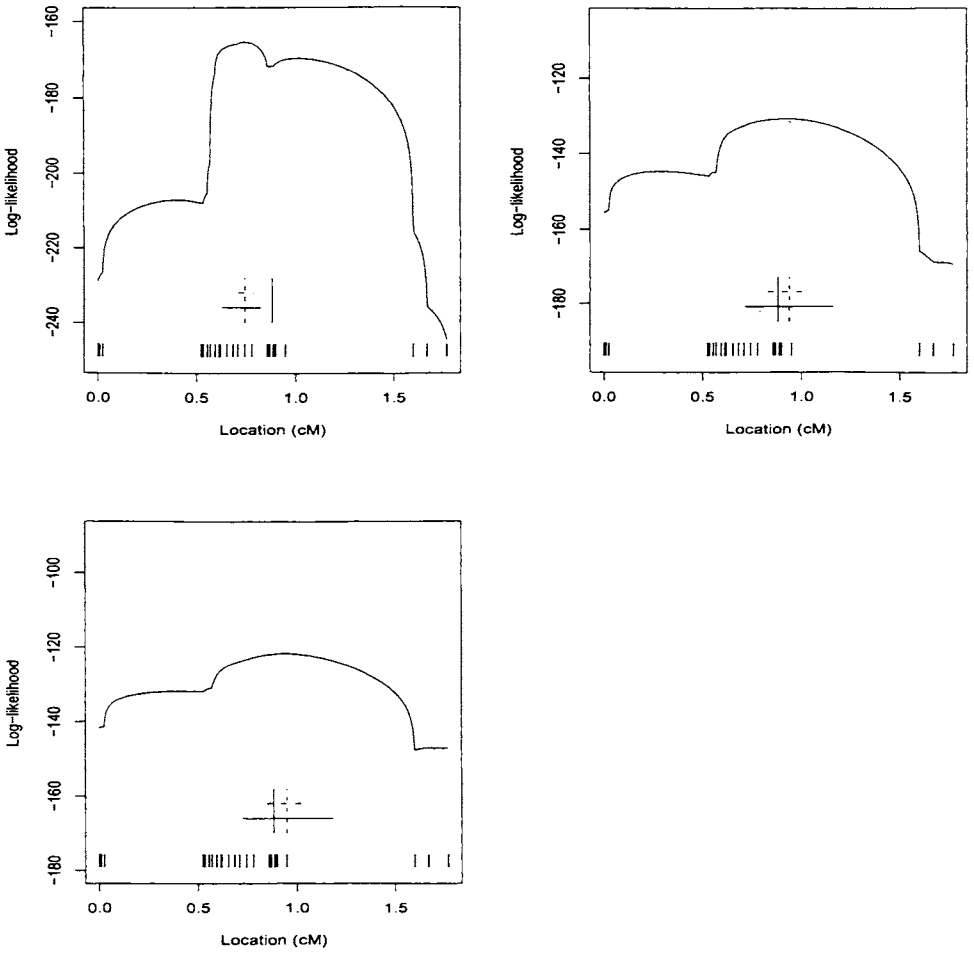


Figure 2: **Results of LD mapping for CF haplotype data** by the DHS method, where background LD is modeled as Markov(η) with $\eta = 0$ (upper left), $\eta = 1$ (upper right), $\eta = 2$ (lower left). Curve gives log profile likelihood vs. (putative) location x of variant, where x is expressed as distance from D21S1885. The dotted vertical line is the estimated variant location based on affecteds' haplotypes. The unbroken vertical line, unbroken and dotted horizontal lines, and the hash marks have the same meaning as in Figure 1.

The curves are nearly identical (after rescaling of the vertical axis), and the CIs for genotype data are only slightly wider, reflecting a slight decrease in information about the location of the variant due to the missing phase. This is expected because CF has a recessive mode of inheritance and low heterogeneity. This lack of heterogeneity means that many loci are homozygous and little information about phase is lost.

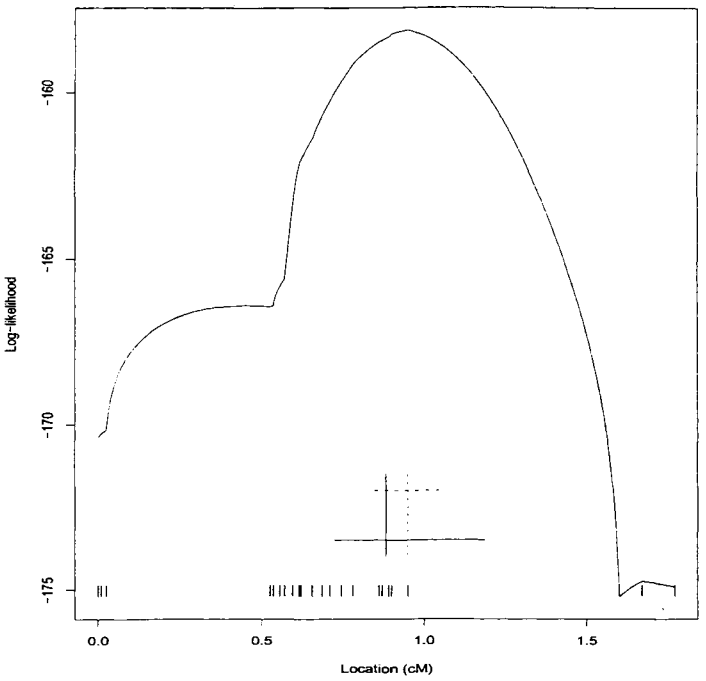


Figure 3: **Results of LD mapping for CF genotype data based on a random pairing of affecteds' haplotypes** Curve gives log profile likelihood vs. (putative) location x of variant, where x is expressed as distance from D21S1885, and where background LD is modeled as $\eta = 2$. The dotted vertical line is estimated location based on affecteds' unphased genotype data. The unbroken vertical line, unbroken and dotted horizontal lines, and the hash marks have the same meaning as in Figure 1.

To add more heterogeneity, we sample 47 affecteds' haplotypes without replacement and pair each with a randomly chosen control haplotype. This mimics the case of a rare dominant trait. We then assume we have only (unphased) genotype information for these 47 pseudo-individuals. The resulting log-likelihood curve, and the log-likelihood curve for the same data assuming haplotypes are available, are given in figure 4. In this case, the difference in the CIs between the cases when only genotype data are available and when haplotype data are available is somewhat more noticeable.

For this simulated example, the assumption of a multiplicative model for the mode of inheritance does not hold, but, at least in this case, the DHS method appears to be relatively robust to the deviation from that assumption.

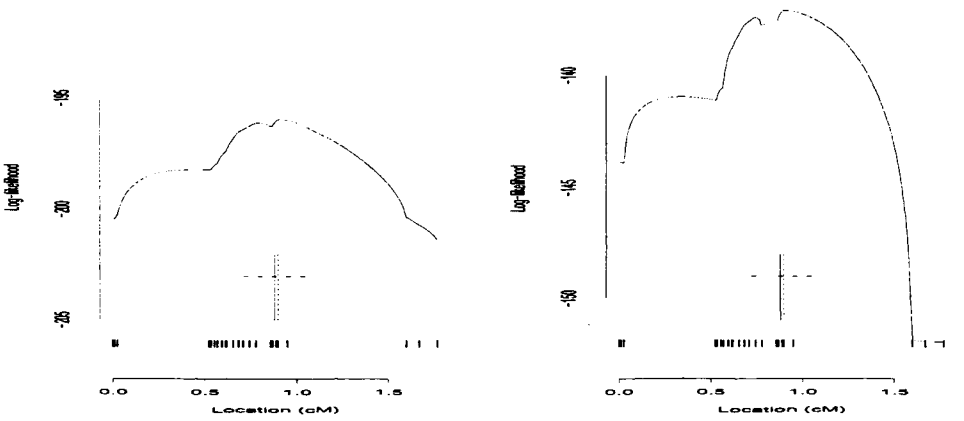


Figure 4: **Results of LD mapping for CF genotype data with added heterogeneity** (left) with results assuming haplotype data with added heterogeneity shown (right) for comparison. In each case, the curve gives log profile likelihood vs. (putative) location x of variant, where x is expressed as distance from D21S1885, and where background LD is modeled as $\eta = 2$. In the top plot, the dotted vertical line is estimated location based on unphased genotype data, while in the bottom plot, the dotted vertical line is estimated location based on haplotype data. The unbroken vertical line, unbroken and dotted horizontal lines, and the hash marks have the same meaning as in Figure 1. The details of the added heterogeneity are given in the text.

4 Discussion

In LD mapping, the presence of unmodeled background LD can have potentially serious consequences. LD that is common to both cases and controls may be mistaken for LD in cases that is due to the presence of a trait-associated variant. This is particularly likely to happen when markers are densely-spaced. For instance, in the CF data set of Kerem *et al.* [15], when background LD is inadequately modeled ($\eta = 0$ or 1) and the mapping-in-controls diagnostic is performed, the highest peak based on the DHS analysis corresponds to the most densely-genotyped region. That background LD would tend to be stronger among closely-spaced markers is to be expected under a model in which LD is broken up by recombination, leading to a sharp drop-off in LD with distance. As a result, if background LD is not appropriately modeled when performing localization, then the tendency will be for the estimated location of the

variant to fall in the most densely-genotyped region or in a region of very low marker information adjacent to the most densely-genotyped region. In addition to the effects of marker spacing, the degree of polymorphism of the markers is also a factor. The lag η of the Markov model may need to be greater with less polymorphic loci. Thus, for example, larger η may be required to adequately model background LD in SNP data than in microsatellite data of the same marker density.

The conclusions regarding importance of adequately modeling background LD would be expected to hold not only for the DHS method, but for other methods with a similar case-control approach, including those of Service *et al.* [29], Morris *et al.* [25], Liu *et al.* [18], and Morris *et al.* [24]. Our results address the effects of background LD on the localization problem. For the problem of detecting association with a variant, when such methods are applied, the presence of unmodeled background LD could be expected to increase the chance of false positive detection of association.

We extend the DHS methodology of McPeek and Strahs [22], who allow for haplotype data and model background LD using a Markov model with lag 1. For haplotype data, we develop a computationally efficient method that allows a Markov model for background LD with lag η. We also extend the method to allow use of (unphased) genotype data, which are much more commonly available than haplotype data. In practice, the order of the Markov model for background LD is limited by the size of the sample of controls needed to estimate the frequency parameters and by the increasing computational demands of higher order models, especially for genotype data. We have implemented, in free software, the DHS method for both haplotype and genotype data, with background LD modeled as Markov(η) where $\eta \leq 2$. These methods are implemented, for both haplotype and genotype data, using an efficient HMM. The genotype-data DHS HMM incorporates uncertainty about phase into the likelihood and allows the method to operate on data sets with a large numbers of marker loci. In addition to the CF data set which includes 23 markers, we have applied the methodology to data sets with (unphased) genotype data on 80+ markers (data not shown).

We demonstrate the importance of modeling background LD using the CF data set. For that data set, our results indicate that when background LD is assumed absent ($\eta = 0$) or is modeled by a Markov ($\eta = 1$) model, additional unmodeled background LD present in the controls could be driving some of the mapping results in the affecteds. The model selection criteria AIC and BIC both prefer the Markov model with $\eta = 2$ to those with $\eta = 1$ or $\eta = 0$. Based on the mapping-in-controls diagnostic, there is no detectable unmodeled background LD in the controls when the model with $\eta = 2$ is used. When mapping (with affecteds) is performed, the 95% CI for location does not cover the true variant when $\eta = 0$ is used, while the CIs do cover the true location in the cases $\eta = 1$ and $\eta = 2$. In the CF example, there is little heterogeneity among the affecteds, so the model for background LD plays less of a role in the analysis than it would in a situation of greater heterogeneity. Thus, we might reasonably expect the effects of background LD to be more important in other data sets. In practical situations, one could apply the AIC and BIC model selection criteria to compare models

of background LD, and one could apply the mapping-in-controls diagnostic to assess the adequacy of the chosen model.

We have extended our methods to allow for unphased genotype data as well as for haplotype data. While in some cases, genotype data on close family members may provide considerable haplotype information, our extension to unphased genotype data may be particularly useful when it is difficult to obtain genetic data from close relatives, as may happen when studying diseases with a late age of onset, such as Type 2 diabetes. In addition to allowing for haplotype or unphased genotype data, the DHS method can also be extended to allow for trio data [33].

There have been some interesting recent results regarding the possible nature of background linkage disequilibrium [5, 14]. These results suggest that high-resolution haplotype structure, at least in certain regions of the human genome, takes a relative simple form. This consists of disjoint haplotype blocks (of tens to hundreds of kb), where within each block there is very strong LD with only a few (*e.g.* ~2-7) commonly-occurring haplotypes. Between the blocks are regions over which there is lower LD (possibly representing recombination hotspots, at least in some cases [13]). Many questions remain about the extrapolation of these observations to the human genome as a whole and to various human populations, from the select regions, populations, and data sets that have so far been studied. There is currently some interest in a large-scale effort to explore this hypothesis and to take advantage of it for LD mapping (*e.g.* see `http://www.genome.gov/page.cfm?pageID=10001676`). The Markov models considered by [5] are extensions of the models we consider here. In order to characterize this block structure, if it exists, a tremendous amount of data would need to be collected and an enormous number of parameters estimated (including start and end points of blocks, common haplotypes in blocks and their frequencies, associations between common haplotypes in different blocks, and also characteristics of the regions of low LD between blocks). Were such information available, these more detailed models for background LD could be incorporated in a natural way in the DHS model. Furthermore, the DHS likelihood for a single haplotype could itself be modified to incorporate a model of block structure for fine-resolution haplotypes.

Another interesting extension would be to combine the DHS method with a method such as the structured association method of Pritchard *et al.* [27], which uses genotypes at unlinked markers to infer population substructure which is then used to test association at the locus of interest. The information of population substructure could presumably also be used for the localization problem with multilocus data. Alternatively, an idea similar to that of genomic control [8] might be adaptable to the localization problem.

Electronic-Database Information

Software for mapping with haplotype or genotype data and $\eta = 0, 1, 2$ is freely available at `http://galton.uchicago.edu/~mcpeek/software/dhsmap`.

Acknowledgments

This work is supported by National Institutes of Health grants DK55889 and HG01645. We are grateful to Nancy Cox and Jian Zhang for helpful discussions, Ken Wilder for his assistance with the software, and Paul Van Eerdewegh for his suggestion of the mapping-in-controls diagnostic.

Dedication

We dedicate this paper to Terry Speed on the occasion of his 60th birthday, with gratitude for his help, support, and encouragement.

Appendix A: Likelihood calculation and maximization

We have developed extensions to the Baum algorithms for likelihood calculation and maximization that are applicable to our model. We first define $\gamma_l(i)$, for a given sampled haplotype, as the probability that $Q_l = i$ at locus l, conditional on the observed haplotype and the parameter values (all of the following probabilities are conditional on the parameter values), *i.e.*,

$$\gamma_l(i) = P\{Q_l = i \mid \mathbf{O}\},$$

which by the definition of conditional probability is $P\{Q_l = i, \mathbf{O}\}/P\{\mathbf{O}\}$. The numerator is computed as the product of two complementary recursively generated variables, a "forward variable" α and a "backward variable" β. For l, $-l_{le} \leqslant l < 0$,

$$\begin{aligned} P\{Q_l = i, \mathbf{O}\} &= P\{O_{-l_{le}}, O_{-l_{le}+1}, \ldots, O_l, Q_l = i\} \times \\ &\quad P\{O_{l+1}, \ldots, O_{l_{re}-1}, O_{l_{re}} \mid O_{-l_{le}}, \ldots, O_l, Q_l = i\} \\ &= P\{O_{-l_{le}}, O_{-l_{le}+1}, \ldots, O_l, Q_l = i\} \times \\ &\quad P\{O_{l+1}, \ldots, O_{l_{re}-1}, O_{l_{re}} \mid O_l, \ldots, O_{l-\eta+1}, Q_l = i\} \\ &= \alpha_l(i)\beta_l(i), \end{aligned}$$

where

$$\alpha_l(i) = P\{O_{-l_{le}}, O_{-l_{le}+1}, \ldots, O_l, Q_l = i\}$$

and

$$\beta_l(i) = P\{O_{l+1}, \ldots, O_{l_{re}} \mid O_l, O_{l-\eta+1}, Q_l = i\}.$$

The definitions for the forward and backward variables on the centromeric side of locus 0 are mirror images of the previous case, *i.e.*, for l, $0 < l \leqslant l_{re}$,

$$\alpha_l(i) = P\{O_{-l_{le}}, O_{-l_{le}+1}, \ldots, O_{l-1} \mid O_l, \ldots, O_{l+\eta-1}, Q_l = i\}$$

and

$$\beta_l(i) = P\{O_l, \ldots, O_{l_{re}-1}, O_{l_{re}}, Q_l = i\}.$$

We note that the left and right sides of the chain are dependent, conditional on Q_0, only for $Q_0 = N$. In this case, the likelihood of the haplotype is just $P_{null}(\mathbf{h}_{obs})$. For $\eta = 2$,

$$\begin{aligned} P_{null}(\mathbf{h}_{obs}) \;=\; & f_{-l_{le},-l_{le}+1}(\mathbf{h}(-l_{le}), \mathbf{h}(-l_{le}+1)) \times \\ & \left(\prod_{l=-l_{le}, l\neq -2,-1,0}^{l_{re}-2} f_{l,l+1,l+2}(\mathbf{h}(l+2) \mid \mathbf{h}(l), \mathbf{h}(l+1)) \right) \times \\ & f_{-2,-1,1}(\mathbf{h}(1) \mid \mathbf{h}(-2), \mathbf{h}(-1)) \times \\ & f_{-1,1,2}(\mathbf{h}(2) \mid \mathbf{h}(-1), \mathbf{h}(1)) \end{aligned}$$

We note that we "skip over" locus 0 in this product, *e.g.*, we include (recall that O_0 is missing) $P\{O_1 \mid O_0, O_{-1}, \ldots, O_{-\eta}\}$ rather than $P\{O_1 \mid O_0, O_{-1}, \ldots, O_{-\eta-1}\}$, as a Markov model of order η generally implies. Were we to include the latter, the likelihood given $Q_0 = N$ would depend on the marker interval within which the variant is assumed to lie, although, according to our model, the haplotype is drawn from the normal population. Thus, we use

$$\begin{aligned} & P\{Q_l = i, \mathbf{O}\} = \\ & P\{O_{-l_{le}}, O_{-l_{le}+1}, \ldots, O_{-1} \mid O_0, O_1, \ldots, O_\eta, Q_0 = i\} P\{O_0, \ldots, O_{l_{re}-1}, O_{l_{re}}, Q_0 = i\} \end{aligned}$$

to compute $\gamma_0(i)$.

Where $\mathbf{h}$ indexes the sampled haplotypes, $c_l^* = \sum_{\mathbf{h}} \gamma_{l,\mathbf{h}}(A)$ and $b_l^* = \sum_{\mathbf{h}} \gamma_{l,\mathbf{h}}(A) \times 1_{\mathbf{h}(l)\neq \mathbf{h}_{anc}(l)}$, then $(c^*_{-l_{le}}, b^*_{-l_{le}}, \ldots, c^*_{-1}, b^*_{-1}, c^*_0, c^*_1, b^*_1, \ldots, c^*_{l_{re}}, b^*_{l_{re}})$ is the conditional expectation of the complete data sufficient statistic for $(1/\tau, p)$ given the data and current model. The model parameters are then re-estimated by maximizing the complete data log-likelihood, substituting this statistic for the complete data sufficient statistic.

The extension to genotype data is straightforward. The primary differences are (1) the state space of the Markov chain is larger and (2) c_l^* is formed by summing over the genotypes, rather than the haplotypes.

Appendix B: Genotype HMM Observation Distributions

Assuming linkage equilibrium,

$$\begin{aligned} P\{O_l^G = (\alpha, \beta) \mid R_l^G\} = 1/2 P\{O_l^M = \alpha \mid Q_l^M\} P\{O_l^P = \beta \mid Q_l^P\} \\ + 1/2 P\{O_l^M = \beta \mid Q_l^M\} P\{O_l^P = \alpha \mid Q_l^P\} \end{aligned}$$

for the case when both $\alpha \neq \beta$ and $Q_l^M \neq Q_l^P$, and

$$P\{O_l^G = (\alpha,\beta) \mid R_l^G\} = P\{O_l^M = \alpha \mid Q_l^M\} P\{O_l^P = \beta \mid Q_l^P\}$$

if either $\alpha = \beta$ or $Q_l^M = Q_l^P$, where O_l^M is defined to be the allele at locus l on the maternally-inherited haplotype, O_l^P is defined to be the allele at locus l on the paternally-inherited haplotype, and $P\{O_l^M \mid Q_l^M\}$ and $P\{O_l^P \mid Q_l^P\}$ are given by expression (2).

Let $O_l^G = (O_l^1, O_l^2)$ be the alleles of the genotype at locus l, given in order Φ_l. For a Markov model of lag 1, the observation distribution is given by

$$\begin{aligned} P\{O_l^G \mid R_l^G, O_{l+1}^G\} = {} & P(O_l^1|Q_l^1, O_{l+1}^1) P(O_l^2|Q_l^2, O_{l+1}^2) I_{l,l+1} \\ & + P(O_l^1|Q_l^1, O_{l+1}^2) P(O_l^2|Q_l^2, O_{l+1}^1)(1 - I_{l,l+1}) \end{aligned}$$

for $0 \leqslant l < l_{re}$, where $P(O_l|Q_l, O_{l+1})$ is given by expression (3). When $Q_l^1 = Q_l^2 = A$, this equation reduces to $P(O_l^G|Q_l^1 = Q_l^2 = A, O_{l+1}^G) = P(O_l^1|Q_l^1 = A) P(O_l^2|Q_l^2 = A)$. For a Markov model of lag 2, the observation distribution is given by

$$\begin{aligned} P\{O_l^G \mid R_l^G, R_{l+1}^G, O_{l+1}^G, O_{l+2}^G\} = {} & P(O_l^1|Q_l^1, O_{l+1}^1, O_{l+2}^1) P(O_l^2|Q_l^2, O_{l+1}^2, O_{l+2}^2) I_{l,l+1} I_{l+1,l+2} \\ & + P(O_l^1|Q_l^1, O_{l+1}^1, O_{l+2}^2) P(O_l^2|Q_l^2, O_{l+1}^2, O_{l+2}^1) I_{l,l+1} (1 - I_{l+1,l+2}) \\ & + P(O_l^1|Q_l^1, O_{l+1}^2, O_{l+2}^2) P(O_l^2|Q_l^2, O_{l+1}^1, O_{l+2}^1)(1 - I_{l,l+1}) I_{l+1,l+2} \\ & + P(O_l^1|Q_l^1, O_{l+1}^2, O_{l+2}^1) P(O_l^2|Q_l^2, O_{l+1}^1, O_{l+2}^2)(1 - I_{l,l+1})(1 - I_{l+1,l+2}) \end{aligned}$$

for $0 \leqslant l < l_{re} - 1$, where $P(O_l|Q_l, O_{l+1}, O_{l+2})$ is as given by expression (4). When $Q_l^1 = Q_l^2 = A$, this equation reduces to

$$P(O_l^G|Q_l^1 = Q_l^2 = A, R_{l+1}^G, O_{l+1}^G, O_{l+2}^G) = P(O_l^1|Q_l^1 = A) P(O_l^2|Q_l^2 = A).$$

Andrew L. Strahs, Department of Biostatistics, Harvard School of Public Health, Boston,
astrahs@hsph.harvard.edu

Mary Sara McPeek, Department of Statistics, University of Chicago, Chicago,
mcpeek@galton.uchicago.edu

References

[1] L. E. Baum. An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities*, 3:1–8, 1972.

[2] C. Bourgain, E. Genin, H. Quesneville, and F. Clerget-Darpoux. Search for multifactorial disease susceptibility genes in founder populations. *Annals of Human Genetics*, 64:255–265, 2000.

[3] D. Clayton and H. Jones. Transmission/disequilibrium tests for extended marker haplotypes. *American Journal of Human Genetics*, 65:1161–1169, 1999.

[4] A. Collins and N. E. Morton. Mapping a disease locus by allelic association. *American Journal of Human Genetics*, 95:1741–1745, 1998.

[5] M. J. Daly, J. D. Rioux, S. F. Schaffner, T. J. Hudson, and E. S. Lander. High-resolution haplotype structure in the human genome. *Nature Genetics*, 29:229–232, 2001.

[6] A. de la Chappelle and F. A. Wright. Linkage disequilibrium mapping in isolated populations: the example of Finland revisited. *Proceedings of the National Academy of Sciences, USA*, 95:12416–12423, 1998.

[7] B. Devlin, N. Risch, and K. Roeder. Disequilibrium mapping: composite likelihood for pairwise disequilibrium. *Genomics*, 36:1–16, 1996.

[8] B. Devlin and K. Roeder. Genomic control for association studies. *Biometrics*, 55:997–1004, 1999.

[9] L. Excoffier and M. Slatkin. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular Biology and Evolution*, 12:921–927, 1995.

[10] J. Hästbacka, A. de la Chapelle, I. Kaitila, P. Sistonen, A. Weaver, and E. Lander. Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. *Nature Genetics*, 2:204–211, 1992.

[11] J. Hästbacka, A. de la Chapelle, M. M. Mahanti, G. Clines, M. P. Reeve-Daly, M. Daly, B. A. Hamilton, K. Kusumi, B. Trivedi, and A. Weaver. The diastrophic dysplasia gene encodes a novel sulfate transporter: positional cloning by fine-structure linkage disequilibrium mapping. *Cell*, 78:1073–1087, 1994.

[12] M. Hawley and K. Kidd. Haplo: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. *Journal of Heredity*, 86:409–411, 1995.

[13] A. J. Jefffreys, L. Kauppi, and R. Neumann. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genetics*, 29:217–222, 2001.

[14] G. C. L. Johnson, L. Esposito, B. J. Barratt, A. N. Smeith, J. Heward, G. Di Genova, H. Ueda, H. J. Cordell, I. A. Eaves, F. Dudbridge, R. C. J. Twells, F. Payne, W. Hughes, S. Nutland, H. Stevens, P. Carr, E. Tuomilehto-Wolf, J. Tuomilehto,

S. C. L. Gough, D. G. Clayton, and J. A. Todd. Haplotype tagging for the identification of common disease genes. *Nature Genetics*, 29:233–237, 2001.

[15] B. Kerem, J. M. Rommens, J. A. Buchanan, D. Markiewicz, T. K. Cox, A. Chakravarti, M. Buchwald, and L.-C. Tsui. Identification of the cystic fibrosis gene: genetic analysis. *Science*, 245:1073–1080, 1989.

[16] J. C. Lam, K. Roeder, and B. Devlin. Haplotype fine mapping by evolutionary trees. *American Journal of Human Genetics*, 66:659–673, 2000.

[17] L. Lazzeroni. Linkage disequilibrium and gene mapping: an empirical least-squares approach. *American Journal of Human Genetics*, 62:159–170, 1998.

[18] J. S. Liu, C. Sabatti, J. Teng, J. B. Keats, and N. Risch. Bayesian analysis of haplotypes for linkage disequilibrium mapping. *Genome Research*, 11:1716–1724, 2001.

[19] J. C. Long, R. C. Williams, and M. Urbanek. An E-M algorithm and testing strategy for multiple-locus haplotypes. *American Journal of Human Genetics*, 56:799–810, 1995.

[20] C. J. MacLean, R. B. Martin, P. C. Sham, H. Wang, R. E. Straub, and J. S. Kendler. The trimmed-haplotype test for linkage disequilibrium. *American Journal of Human Genetics*, 66:1062–1075, 2000.

[21] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, 1989.

[22] M. S. McPeek and A. L. Strahs. Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. *American Journal of Human Genetics*, 65:858–875, 1999.

[23] A. P. Morris and J. C. Whittaker. Fine scale association mapping of disease loci using simplex families. *Annals of Human Genetics*, 64:223–237, 2000.

[24] A. P. Morris, J. C. Whittaker, and D. J. Balding. Fine-scale mapping of disease loci via shattered coalescent modeling of genealogies. *American Journal of Human Genetics*, 70:686–707, 2000.

[25] A. P. Morris, J. C. Whittaker, and D. J. Balding. Bayesian fine-scale mapping of disease loci, by Hidden Markov Models. *American Journal of Human Genetics*, 67:155–169, 2002.

[26] R. Nielsen. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics*, 154:931–942, 2000.

[27] J. K. Pritchard, M Stephens, N. A. Rosenberg, and P. Donnelly. Association mapping in structured populations. *American Journal of Human Genetics*, 67:170–181, 2000.

[28] B. Rannala and J. Reeve. High-resolution multipoint linkage-disequilibrium mapping in the context of a human genome sequence. *American Journal of Human Genetics*, 69:159–178, 2001.

[29] S. Service, D. Temple Lang, N. Freimer, and L. Sandkuijl. Linkage-disequilibrium mapping of disease genes by reconstruction of ancestral haplotypes in founder populations. *American Journal of Human Genetics*, 64:1728–1738, 1999.

[30] M. Stephens, N. J. Smith, and P. Donnelly. A new statistical method for haplotype reconstruction. *American Journal of Human Genetics*, 68:978–989, 2001.

[31] J. D. Terwilliger. Fine-scale genetic mapping based on linkage disequilibrium: theory and applications. *American Journal of Human Genetics*, 56:777–787, 1995.

[32] M. Xiong and S.-W. Guo. Fine-scale genetic mapping based on linkage disequilibrium: theory and applications. *American Journal of Human Genetics*, 60:1513–1531, 1997.

[33] J. Zhang. *Linkage disequilibrium mapping by the decay of haplotype sharing in a founder population*. PhD thesis, University of Chicago, 2001.

[34] S. Zhang and H. Zhao. Linkage disequilibrium mapping in populations of variable size using the decay of haplotype sharing and a stepwise-mutation model. *Genetic Epidemiology*, 19(Suppl 1):S99–S105, 2000.

[35] S. Zhang and H. Zhao. Linkage disequilibrium mapping with genotype data. *Genetic Epidemiology*, 22:66–77, 2002.

[36] H. Zhao, S. Zhang, K. R. Merikangas, M. Trixler, D. B. Weldenauer, F. Sun, and K. K. Kidd. Transmission/disequilibrium tests using multiple tightly linked markers. *American Journal of Human Genetics*, 67:936–946, 2000.

Some Considerations for the Design of Microarray Experiments

John H. Maindonald, Yvonne E. Pittelkow and Susan R. Wilson

Abstract

Issues relevant for the design of gene expression experiments using spotted cDNA microarrays and gene chip microarrays are overviewed. Emphasis is placed on the uses of replication, and on the importance of identifying major sources of variation.

Keywords: microarrays; oligonucleotide; design of experiments; variability; replication; gene expression

1 Introduction

Microarrays are new and evolving technologies that enable large numbers of genes, up to the order of tens of thousands, to be evaluated simultaneously. Our aim is to give a brief overview of principles of experimental design, and to comment on their application to microarray experiments. A major theme is that, for purposes of design, the different sources of variation in gene expression are not well understood.

The objective of a microarray experiment might be to investigate genes which are differentially up or down regulated in cells between, say, a control group and cells which have undergone some treatment, or between cells of animals of different genetic background (*e.g.*, control mice compared to knockout mice) or between cells in healthy tissue and diseased tissues, or between cells at different time points (*e.g.*, developmental biology). Many studies search for genes that have similar expression profiles, often in an attempt to determine genes involved in biological pathways, or in development, or genes involved in regulatory functions. The focus would then be on the analysis of dependency structure. Time course experiments may investigate how the pattern of expression or relative expression changes over the cycle of cell division, or following administration of a drug. Finally, interest may be in estimation of gene expression levels.

The primary goal of the experiment should be clear, as this gives focus to the investigation, desirable even if a major part of the analysis will be a general search for interesting patterns of expression. Many experiments have multiple aims; these must be prioritized. Both in its scale and in the processes that are under investigation, the

biology has a large element of novelty, with implications for statistical design and analysis. Vingron [52], commenting on the "big science" issues that such large-scale technologies raise, draws attention to "a major upcoming challenge for the bioinformatics community to adopt a more statistical way of thinking and to interact more closely with statisticians." Bioinformaticians need to educate themselves in statistics. "Not so much with the goal of mastering all of statistics but with the goal of sufficiently educating ourselves in order to pull in the statisticians."

Our focus here is on design issues for comparative studies for two types of array platform – two-channel cDNA spotted microarrays [17, 20, 24], and high density oligonucleotide microarray chips produced by Affymetrix [1] for expression analysis, which we refer to as gene chip microarrays. For both types of array, DNA sequences are laid out in a grid on a solid substrate. Occasionally we refer to the spotted microarrays as *slides*, recognising however that glass is just one of several possible substrates, and we refer to Affymetrix oligonucleotide microarrays as *chips*. Much of our discussion of spotted cDNA microarrays applies also to oligonucleotide spotted microarrays (distinct from Affymetrix oligonucleotide arrays, which are produced by photolithography rather than spotting), which we do not explicitly discuss. We note that gene chip microarrays can in principle, with suitable calibration, yield *absolute* expression measures. Each individual spotted microarray slide is by contrast used to yield *relative* expression measures, for example between a treatment and a reference, or between one treatment and another. We note also that, perhaps inevitably for technology that is rapidly changing and developing, there is no single established nomenclature that distinguishes clearly between the different types of arrays. A feature that distinguishes microarray experiments from more conventional experiments described in the biostatistical literature is the very large number of parallel measurements on typically only a few cases. Summary measurements are typically provided for each of a large number of genes or of Expressed Sequence Tags (ESTs), which are partial gene sequences. The small number of cases is, in part, a function of the (initial) high costs of the microarrays, especially chips, limitation of available sample, and the (apparent) failure to involve scientists with statistical training in the early stages of the development of microarrays.

The processing of microarray data raises a variety of statistical, mathematical and computational issues, see for example [12, 19, 45, 47, 49]; some of these are alluded to in passing.

The remainder of the paper is organized as follows: Section 2 gives examples of experiments, Section 3 considers outcome measures, Section 4 notes experimental design principles and discusses their application to microarray experiments, Section 5 considers sources of variation, Section 6 discusses the design of microarray slides and chips, and Section 7 summarizes the discussion.

2 Examples of Experiments

2.1 Spotted Microarrays

In a typical spotted microarray experiment, samples from a treatment and from a reference are combined in equal proportions and hybridized to cDNA probes that have been spotted on a slide. A key question is whether the comparisons that are of interest will be made directly or indirectly. In an indirect comparison, each treatment that is of interest is compared with a reference sample, and the responses of the treatments relative to this reference sample are then compared. In a direct comparison, treatments are directly compared with each other.

For example, Callow *et al.* [6] used the indirect comparison approach to search for genes that were differentially expressed between liver tissue from apolipoprotein apoAI-knockout (test) mice and liver tissue from C57B1/6 (control) mice. Each of 8 test mice was compared with the reference sample, and each of 8 control mice was also compared with the reference sample. For a reference sample, material from the same eight control mice was pooled.

For each of the 16 mice, cDNA, labeled to reflect the source of the mRNA, was prepared by reverse transcription of mRNA. The experiment we describe used Cy5 ("red") and Cy3 ("green") dyes, with Cy5 for individual mice and Cy3 for the reference. The cDNA from each mouse was combined with the cDNA from the reference sample and hybridized to a slide. This experiment resulted in 8 comparisons between control mice and reference, and 8 comparisons between test mice and reference.

Preparation of a spotted microarray slide involves choosing and fixing a large number of spots on a slide, with each spot containing a number of strands of DNA or cDNA that are intended to uniquely hybridize, or bind, to the corresponding gene in the labeled cDNA sample. In this experiment around 6000 spots, one or two per gene, were laid down (spotted) on each of 16 microarray slides (one per "treatment"). After separate labeling, the mixed sample was hybridized to the slide in specially humidified chambers. Laser-induced fluorescence imaging was then used to detect dye intensities. This gave two images of the slide, one for the treatment (test or control) and one for the reference. Image analysis software, together with some post-processing, was then used to derive a background-corrected relative intensity measure for each spot.

Results, for each spot on each of the 16 slides, were expressed as the logarithm of a ratio of the intensity value for each mouse to the intensity value for the pooled reference. Two-sample t-tests, with an adjustment for the large number of comparisons made, were then used to compare the log-ratios from the test mice and the control mice. The study identified eight spots, corresponding to four genes, that were under-expressed in test mice relative to controls.

2.2 Gene chip expression microarrays

In a typical gene chip microarray experiment, prepared cRNA sample is hybridized to the probes on a chip. The chip is then scanned to obtain fluorescence intensity readings of stains incorporated during the laboratory procedures. Image processing software is then used to compute intensity values for each probe.

In contrast to typical spotted microarray experiments, only one sample is hybridized to a chip, allowing, in principle, the estimation of absolute expression values. Because of the high cost of these chips, efficient use is important.

The main characteristics of gene chip microarrays are:

1. Thousands of short oligonucleotide probes (commonly 25-mer, *i.e.*, 25 bases in length) are synthesized *in situ* on a glass substrate, using photolithographic techniques. Multiple paired sets of probes (commonly 11, 16 or 20) are used for each gene or EST. The probe sequences are chosen according to specific criteria described in Lockhart *et al.* [35].

2. One probe in a pair has the exact sequence from the gene or EST, while in the other member of the pair the middle base is changed to its complement. The mismatched probes (*MM*) provide a probe-specific control or nonspecific hybridisation control. The collection of perfect match (*PM*) probes and mismatched probes (*MM*) corresponding to one gene or EST makes up a probe set.

3. User control over the choice and layout of probes requires the construction of custom arrays, whose cost is beyond the resources of many laboratories.

We note that probes are not chosen at random, nor are they independent, although some analyses make this assumption.

In an experiment described by Efron *et al.* [14], the aim was to study transcriptional responses to ionising radiation in the context that some cancer patients have severe life-threatening reactions to radiation treatment. It is important to understand the genetic basis of this sensitivity so that patients with high rates of sensitivity can be identified before being allocated treatment. The design was a factorial experiment with two levels each of two factors, namely (i) RNA was taken from two wild-type human lymphoblastoid cell lines; (ii) the growing state was either irradiated or unirradiated; in addition RNA samples were labeled and divided into two identical aliquots for independent hybridizations. Each microarray provided expression estimates for 6810 genes/ESTs.

Another type of gene chip microarray experiment is described by Golub *et al.* [23]. Their aims were essentially class prediction (assigning tumours to known classes) and class discovery (identifying new cancer classes). They analysed leukemia data of 38 bone marrow samples obtained at time of diagnosis: 27 acute lymphoblastic leukemia (ALL) and 11 acute myeloid leukemia (AML).

3 Issues Concerning Outcome Measures

As noted, spotted microarrays typically yield two intensity measurements for each spot, which are combined into a single ratio or logratio. Gene chip microarrays yield one intensity measurement for each probe. The information from each probe set is generally combined into a single expression index for the probe set. The outcome measure is, in either case, essentially multivariate.

Evidence for the form of the link between expression summary measures and mRNA concentration (or number of molecules) is sparse; however see [8, 25, 28, 32] for gene chip microarrays. When an antibody amplification step is employed, the link is more tenuous, due to nonlinearity in its action. It is important to note that even with replicate slides or chips that use different subsamples from the same sample, and where laboratory procedures have been carried out as similarly as possible, the scanned images can show considerable differences. The normalization or scaling techniques that attempt to make intensity measures comparable between slides or chips are different for the two technologies; see [28] for chips, [55] for slides.

Saturation effects, *i.e.* intensity readings close to or above the upper detection limit of the scanner, are an extreme form of nonlinearity. At high mRNA concentration or high laser power, all intensity measurements may be inaccurate due to saturation. Where one of two estimates being compared is affected by saturation, the estimated difference is attenuated. If both are affected by saturation, the difference will be meaningless [26]. Due to the large number of genes or probes, each with a potentially different saturation level, global avoidance of all such regions may not be feasible, and detection strategies are required.

For both technologies, negative controls (*i.e.* spots or probe sets that should never show a signal) or positive controls (*i.e.* should always show a signal), can be useful checks.

3.1 Spotted Microarrays

Each slide may be used either for a comparison between treatment and reference, or for a comparison between two treatments. In either case, there is one intensity ratio or log-ratio for each spot.

There are typically separate background corrections for the red and the green signals. Both foreground and background signals will differ, depending on the scanner settings and on the image analysis software used [54]. Important considerations are the identification of the spot boundary, the choice of the region used to estimate background and the form of the background adjustment. Negative intensity estimates that can result from background subtraction are a nuisance for later data processing, and should be avoided.

Ramdas *et al.* [41] noted that signal quenching associated with excessive dye concentrations led to nonlinearity in signal intensities. Spot size and morphology can affect intensity measurements. Thus, the routine use of the intensity ratio or logarithm of the

intensity ratio as the comparative expression measure is open to question. If, for example, the intensity measurements were changing additively, then differences could be used. On the other hand, if the intensity measurements were changing proportionately then differences in the log values would be used. Currently this is the scale that is widely chosen. If there are three (or more) treatments, then an experiment that has all pairwise comparisons allows us in principle to check that the chosen scale is appropriate. It is prudent to check, to the extent that this is possible, that measurements are in a range where response is linear.

3.2 Gene chip microarrays

In statistical terms, the data from each chip is a single multivariate response vector, with complex dependencies inherent from the biology and the technology. As mentioned earlier, generally a summary measure or estimate of expression is computed from the multiple probes in each probe set, following suitable background estimation and chip normalization (calibration). A number of different summary measures or expression indices are in use. Some are based on differences between the probe intensity (*PM*) and its nonspecific hybridization (*MM*) control; examples include the Affymetrix trimmed average difference (AvDiff, [1]), the model-based expression indices of Li and Wong [33], and the average median filtered differences of Alon *et al.* [2]. Since as many as a third of the *MM* control probes can have intensity readings higher than their paired *PM* probe, truncation, filtering or transformation are often used to accommodate the negative values of $PM - MM$ differences. Some measures do not use the nonspecific hybridization control probes except to calculate a background estimate [28, 34, 39]. Other possibilities include the log of the ratio of the *PM* probe to *MM* probe [1, 32, 39], the robust multi-array average (RMA) approach [28], and empirical Bayes estimation [14]. Other summary measures are also found in the biological literature (*e.g.* [21]).

4 Experimental Design

This section is organized as follows: An introductory subsection discusses aims and principles of experimental design, then bias and replication are discussed in more detail; 4.1 discusses pooling, which is an issue for both types of array; finally 4.2 discusses special issues for spotted microarrays, including the choice between direct and indirect comparison, and dye bias. There are many excellent texts and papers that discuss general principles of experimental design, including [5, 9, 10, 15, 42, 36]. Here we discuss these in the context of microarrays.

Design questions relevant to the aim of the experiment that should be clear before proceeding include:

1. What are the "treatments"?
2. What are the experimental units?

3. What are the experimental measurements?
4. What is measured, and what do the measurements mean?
5. What comparisons are of interest? (Note that interactions are a form of comparison.)

For microarray experiments, "treatments" refer not only to defined procedures, for example treatment by a drug, but also to qualitatively different units, such as tissues from healthy and unhealthy organs, or tissues from wild type model organisms and genetically modified organisms.

For example, in the Callow *et al.* [6] experiment the comparison was between test (knockout) mice and control mice. In the Efron *et al.* [14] experiment, the main interest was in the comparison between irradiated and unirradiated cells, allowing for a possible difference in effect between cell lines, *i.e.*, for a possible interaction between the irradiation effect and cell line.

Cox and Reid [10, p. 4] define an experimental unit as the "smallest subdivision of the experimental material such that any two different experimental units might receive different treatments". The sample may be from a single organism, or it may be a pooled sample of material from several organisms.

In the Callow *et al.* [6] experiment, it is convenient to regard the separate red and green labeled samples that are mixed and hybridized onto a slide as a pair of experimental units, yielding separate intensity information that will (usually), for analysis, be combined into a single log intensity ratio. In Efron *et al.* [14], the experimental units are, strictly, the four separate mRNA samples, each of which is repeated.

A broad over-riding aim of experimental design is to use resources in the manner that will best achieve the intended purpose and produce conclusions that are widely valid (*i.e.*, that are not restricted to too specific a set of conditions). However, this needs to be balanced against the need for simplicity and robustness of design. We begin with a list of broad aims and principles of statistical experimental design, using experiments with spotted microarrays for illustrative purposes, followed by further discussion of some of the issues. Later, we consider special issues for the design of spotted microarray experiments.

Broadly, the aims are to find designs that:

1. *Allow generalization* of results to the relevant wider population;
2. *Avoid bias*, or systematic error;
3. *Minimize the effects of random error*, for a given cost;
4. Allow an *assessment of the accuracy of estimates* of effects that are of interest;
5. *Are robust*, in the sense that they will still give useful results even if there are occasional failures in the experimental protocol, or if some assumptions that motivated the design prove to be false.

Basic devices that are available to achieve these aims are:

1. *Controlling for all "fixed" effects* for which this is possible. For example, the expression of genes in some tissues will be different depending on whether the tissue is from a male or female;

2. *Blocking*, or local control, to allow an accurate assessment of effects under varying experimental conditions. In two-channel spotted microarray experiments, each pair of samples is a block. In general, it is desirable to match the treatment and control samples as closely as possible;

3. *Randomisation* of treatment allocations with respect to factors that cannot be controlled. For example, in a two channel spotted microarray experiment, it is inherently desirable to randomise the allocation of dyes to treatments, in such a way that each treatment occurs equally often with each dye;

4. *Replication* of experimental units, at least to an extent that an estimate of accuracy is possible. In principle, replication may be further increased to achieve a pre-specified accuracy. Additionally, by reducing the opportunity for one unsatisfactory replicate to damage results, replication makes experiments more robust;

5. *The use of repeats*, *e.g.*, repeated spots, within experimental units, where this makes a useful contribution to reducing variability between experimental units. As with replication of experimental units, this has the additional effect that experiments are more robust;

6. Giving first priority in use of experimental resources to *controlling the effects that have the largest implications* for results. For example, once appropriate forms of correction have been applied, the dye effect may, for the present spotted microarray technology, be inconsequential; *i.e.*, any remaining bias from this source may be dwarfed by other sources of variability.

Avoiding Bias

The best way to deal with bias is to modify instrumentation or experimental procedures to avoid it. Where a bias is associated with instrumentation, it may be possible to find an analytical adjustment that verifiably removes or reduces the bias. If neither of these approaches is completely successful, and the necessary information is available, one of devices 1–3 above can be used.

A major difficulty in discussing methods of avoiding bias in microarray experiments is that there is insufficient systematic information available about the biases involved. At present, the exception for spotted microarrays is the bias arising from differences between the dyes used to label the different samples [13]. There is some evidence of day effects, *i.e.* changes in response from one day to another, for both types of microarray. Concerning other sources of bias, until appropriate experiments are performed it might

be prudent to make the laboratory situations as uniform as possible during the course of an experiment and to randomise treatment allocation over any potential sources of bias that are not otherwise controlled.

Replication

A discussion of replication and decisions on the optimal level of replication are intimately linked with understanding the sources of error, which we address in a later section. In the context of replication, it is useful to consider a hierarchy of corresponding variation, as in Yang and Speed [56], with the following levels:

1. Separate slides/chips to (separately) obtain measurements on samples from distinct biological sources – biological replicates;

2. Separate slides/chips to probe each of several replicate preparations of RNA from the same biological source (sometimes, and rather misleadingly, also referred to as biological replicates);

3. Technical replicates that use distinct slides/chips to obtain measurements on different target samples of RNA from the same preparation;

4. For spotted microarrays, replicate spots on the slide.

Biological replication is essential when the intention is to make claims about a broader population of patients, plants or animals. Since biological organisms can vary substantially, such replication would be necessary even if the measurement device gave exactly reproducible results when repeated on an individual. Note in this context the broad distinction between technical reproducibility and biological reproducibility. Note also that in the above hierarchy, variation at any lower level contributes to variation at all higher levels.

Since the reasons for replication are not transparent to all, we repeat them here in the microarray context: (i) to allow generalization to the wider biological population (and replication at the biological level is essential for this); (ii) to provide information that will make it possible to do a better experiment next time; (iii) to reduce variation (and increased replication at the biological level will certainly do this, but may be an unnecessarily expensive method if a similar improvement could be achieved by increased replication further down the hierarchy); (iv) to allow identification of major sources of variability, in the hope that something might be done about some of them (and in this context we might want to consider crossed, *i.e.* nonhierarchical, sources of variation); (v) to allow identification of outliers, at levels where that may be important; (vi) to make experiments more robust.

The calculation of the number of replicates required to be able to detect a difference of a given size (power calculations) is challenging in microarray experiments, not only because the newness of the field means that even rough guides to variance estimates for

given probe sequences are unknown but also because estimates will change between probe sequences.

Above, we distinguish "technical replicates" from biological replicates. When replication is used to reduce variance (because analysis can be based on the mean or other summary measure) it is important that the replicates be as independent as possible. For example, using different sample preparation hybridized to chips/slides is probably preferable here to using duplicate chips/slide but the same mRNA sample.

At least for spotted microarrays, a further level of replication is possible, namely replicate spots on the same slide, as recommended in Tseng *et al.* [51]. However, the placement of these duplicate spots needs to be carefully considered to avoid potential systematic bias; see Yang and Speed [56]. Removal of one apparently contaminated spot may enable remaining spots to be used in further analysis [51].

For gene chip microarrays, limited available sample material and the relatively high cost of chips often limit the number of biological or technical replicates. While noting that there are no firm standards on the number of replicates required in a microarray chip experiment, Novak *et al.* [40] mention that they commonly design their initial experiments to include three replicates for each biological state, including control. Li and Wong [33] recommend 10 replicates for estimating standard errors used for detecting outliers in gene chip microarray studies. Glynne at at [22] recommend between two and five replicates.

The value of replication in a spotted microarray experiment was shown by Lee *et al.* [31] who, limiting their attention to the red signal, carried out an experiment in which 32 out of 288 genes were expected to be strongly expressed, while the remaining genes should not have been expressed. They used a mixture model to identify genes that were expressed. Although the assumptions required for their analysis can be questioned, their qualitative conclusion holds, in particular that results from individual replicates are unreliable, and of unknown accuracy. With two replicates, there is some indication of the extent of irreproducibility; however, Lee *et al.* recommend doing at least three replicates.

In general, and depending on the tissue, experiments with human tissue are likely to require more extensive replication than experiments with tissue from highly inbred strains of laboratory animals.

Multiple independent estimates of treatment effects

Designs that allow multiple independent estimates of treatment effects may allow reduced replication, or even no replication. For example, for spotted microarrays consider the "all possible pairs" experimental design with three treatments A, B and C. There are two estimates of the contrast between A and B: one that is obtained directly by comparing B with A, and the other that is obtained by subtracting the A versus C effect from the B versus C effect. Thus, if each pairwise comparison is made only once, there is one degree of freedom that can be used for the estimation of "noise"; we prefer this

term to the commonly used term "error". If the design has two replicates of each of the three two-way comparisons, there are four degrees of freedom for estimation of noise.

With four or more treatments, there are several alternatives to designs in which all comparisons are with a reference. The design that has each of the six possible comparisons between four treatments has three degrees of freedom for estimation of noise for evaluating each treatment comparison. An alternative is the loop design [30] that compares A with B, B with C, C with D, and D with A. This design has one degree of freedom for estimation of noise. The comparisons that must be made indirectly, between A and C and between B and D, are on average less precise than the comparisons that can be made directly. Where there are many treatments, some comparisons in a loop design will involve many links, with a consequent loss of precision. Modification of loop designs to add comparisons that avoid many connecting links is therefore desirable.

Considerations that will affect the choice between the different designs include: the number of slides that are required; the precision of the comparisons that are of chief interest; the amount of available mRNA, for treatments and where relevant for the reference; the robustness of the design; and the ease of carrying out the analysis.

Factorial designs

Following the structuring of comparisons in terms of main effects and interactions of factors, it may be possible to incorporate into the noise term high order interactions that are not statistically significant, thus increasing the available degrees of freedom for estimating the relevant noise variance. This should be considered at the design stage, although often it is left to the analysis stage.

For example, Efron *et al.* [14] used an initial exploratory analysis to satisfy themselves that the effect of radiation was similar for both levels of cell line, for both aliquots. Hence, they felt able to assume that the three interactions involving irradiation were zero, giving three degrees of freedom for estimating the relevant noise variance. This does, however, ignore the implications for variance structure of the nesting that arises from the way that aliquots were formed in this experiment, namely by splitting samples in two.

For a general discussion of factorial design issues, see Cox [9, pp.94–96] and Cox and Reid [10, pp.99–101].

4.1 Pooling – an issue for both technologies

If there is insufficient RNA from the tissues under investigation from one individual, then it is common practice to prepare RNA from, say, several individuals from a pure (inbred) line, kept as far as possible in a common environment. Other reasons for pooling include provision of adequate quantities of a standard that can be maintained consistently over time, and to "reduce" variation. An alternative to pooling is amplification. Depending on how it is done, however, amplification can bias abundance

relationships [4, 29]. At the same time amplification can, for spotted microarrays, lead to results that are more consistent between slides.

A concern is that pooling might increase or modify potential masking effects that may arise from the hybridization of RNA to itself or to other strands of RNA. Self-hybridization is an aspect of secondary structure as described in Zuker [57]. Consistently with comments in Yang and Speed [56], we have been unable to find direct experimental evidence on this point. If masking is not a serious problem and pooling is indeed a form of averaging, then it should be used wherever possible, for treatments as well as for any control. Replication will then require the use of replicate pooled samples, with different individuals used for the different pooled samples. Or is pooling perhaps more problematic for treatment samples than for reference samples, *e.g.*, for knockout or transgenic organisms? There is a clear demand for better knowledge of effects at this level.

For gene chip microarray experiments, Novak *et al.*. [40] suggested that pooling to reduce biological variation is of limited value. On the other hand, Bakay *et al.* [3] concluded that pooling is of value. Such conflicting claims are due, in part, to the different methods used to examine variability, but the issue is clearly unresolved.

4.2 Some special issues for spotted microarrays

The issues that we discuss here are special to spotted microarrays because each slide gives comparative information – either between two treatments, or between a treatment and a reference.

The design used by Callow *et al.* [6], described above, is analogous to the conventional completely randomised design. Note that the use of a common reference sample creates a correlation between the two sets of comparisons with the reference. Additionally, for this experiment one of the comparisons is between the reference and individual mouse samples that are correlated with the reference. An alternative is a design in which each slide gives a direct comparison between a test mouse and a control mouse. Such a direct comparison will, with 8 slides, be more precise than the indirect comparison that used 16 slides, while requiring less mRNA from each control mouse and the same amount of mRNA from each test mouse. Often, though not in the Callow *et al.*. experiment, the comparison with reference will have intrinsic interest. The choice is then between the design that has all pairwise comparisons, and the design that has only the comparisons between treatments and reference.

We have noted that a direct paired comparison of the two treatments should be more precise than the indirect comparison (see also Dudoit *et al.* [13]; Yang and Speed [56]; Kerr and Churchill [30]). Applying such a design to the Callow *et al.* experiment, each slide compares a test mouse with a control mouse. A consequence of the correlations alluded to above is that, as demonstrated in [50], the improvement in precision is not as great as a naive analysis might suggest. Paired comparison designs are a simple type of block design, with each pair of samples (mice) that are compared forming a block. Readers who are familiar with classical experimental design will recognise

this as a "paired comparison" experiment, though now with many such comparisons made using a single slide. Fisher [15] discusses such experiments. They are the subject of David's [11] book; see also Cox [9]. These designs have been widely used in food tasting and other sensory evaluation experiments [18]. They are a special case of more general balanced incomplete block designs. For technical details, see Yang and Speed [56] who also discuss and compare many different experimental designs.

The precision of the comparisons that are of interest is not the only consideration. Depending on the experimental context and aim, the experiment in which all comparisons are with a baseline has the following merits: assuming that dye bias affects all comparisons with the reference equally, though perhaps differently for different probe sequences, the swapping of dyes is unnecessary; the comparison between treatments and reference may have an intrinsic interest of its own; limitations in the amount of available mRNA, for one or all of the treatments, may require the use of a design that compares treatments with a reference [56]; use of a reference that is common over different experiments allows treatment effect comparisons across those experiments.

Dye bias

It is now well known that the dye bias varies nonlinearly with the average intensity of the signals [13]. The loess correction, which is one of several corrections that Dudoit *et al.* [13] discuss, seems to work well, but like other such corrections can at best ensure that the bias over all spots is on average reduced to zero. It is in principle possible that the strength of the binding may vary with the sequence of bases to which the dye binds, thus leading to variation between different differentially expressed genes. A cautious approach therefore requires the routine use of dye flips, *i.e.*, each dye occurs equally often with each treatment. This allows an analysis that averages out any bias that remains after the correction.

5 Sources of Variation

The following scheme, adapted from Cox and Reid [10, p. 10], gives a framework for discussion of sources of variation in microarray experiments. Inevitably, it cannot capture the complex ways in which sources of variation may interact:

1. Intrinsic or baseline noise (or "error"), *i.e.*, variation that is inherent in the subjects of the experiment
    (a) Errors associated with the biological, genetic/environmental sources (*e.g.* SNP or different animals or cultures)
    (b) Errors associated with hybridization process (which may be probe dependent);

2. Intermediate noise, *i.e.*, variation associated with the process that leads from treatment to response
   (a) Laboratory (RNA extraction, amplification and labeling)
   (b) Biological sample sources (tissue, homogeneity, contamination);
3. Measurement error, *i.e.*, error associated with the instrumentation
   (a) Chip/slide manufacture (including for spotted microarrays the size and shape of spots)
   (b) Scanning
   (c) Algorithms, including the image processing and scaling procedure used
   (d) Defects arising in the manufacturing process, or in the subsequent handling of slides or chips.

References addressing these sources of variation include [25, 28, 32, 37, 38, 39, 40, 46, 56].

A hierarchy of levels of variation can be envisaged, as detailed in Yang and Speed [56], and might be formalized in a multi-level model, with components of variance attached to each level of the hierarchy. Such models provide a useful framework for thinking about sources of noise, and in addition have a role in the examination of the effects of individual genes. They allow us, *e.g.*, to compare the improvement in precision that arises from the use of multiple spots for the one probe sequence with the improvement from increased technical or biological replication, a point that is demonstrated in the next section. We note that from its beginning, the analysis of variance has been multi-level; see Speed [48]. Many of the models that Fisher [15] analysed had multiple levels of variation.

From a design perspective, we require an estimate of technical variability because we wish to know the contribution that it makes to the variability of biological measurements. Where technical variability is a substantial component, it will be necessary to break it down further, so that we can identify the major sources of noise and take whatever steps are possible to reduce their effect. For a variety of biological and technical measurement reasons, the relative contributions of different noise sources may vary between probe sequences.

Note that:

1. There are several different components of the experimental procedure. If one of these components is, relative to the others, a major component of the variation, attempts should be made to identify it;
2. Comparisons made within individuals, *e.g.*, a cell line from an individual versus a knockout cell line created from the same individual, can be more precise than when the sample and the knockout sample are from different individuals. Experimental procedure becomes more than ever important for controlling the variation that remains;

3. If interest is in getting an accurate estimate of variation, for purposes of generalizing (*e.g.*, to mice generally of a particular strain), then the demand is for repeat results from several individuals, *i.e.*, for genuine biological replication. Then although the standard errors of treatment comparisons can be estimated, it will not be possible to distinguish between variation that arises from experimental procedure and the effects of variation between individuals. The distinction between these two sources of variation may be useful in deciding whether effort on the improvement of laboratory procedure is justified.

6 The Design of Microarray Chips and Slides

There are two aspects of microarray experiment design – the design of the array/chip, and the allocation of the mRNA samples to the array/chip. Because the fabrication of a custom gene chip is expensive, most users accept one of a set of standard gene chip microarray designs. By contrast, users of spotted microarrays do often design their own slides. They then face important issues that include the choice of genes (or ESTs), the number of repeats of each probe sequence, and the relative positioning of repeats. In addition, each gene may be represented by more than one probe sequence. A major advantage of fabricated oligonucleotide sequences, for spotted arrays as well as for chips, is in the opportunities that they offer for selecting and testing probe sequences. This is an important ongoing research area, which is however beyond the scope of this paper; we refer the reader to Rouillard *et al.* [44].

The remaining discussion will comment on the number and possible prioritization of genes represented on the slide or chip, and the use of repeats. Our comments have direct relevance to cDNA microarray slides, where there is ordinarily one probe sequence, perhaps repeated, for each gene or EST, but the principles are general.

Many probes, or few probes

It is tempting to include as many probes for genes as possible on a slide. However, as the number of different genes represented on the slide increases, so also does the potential for false positives when, say, analysing a comparative experiment. To avoid this situation, the criteria for establishing differential expression becomes more stringent for statistical tests as the number of tests are increased. For example a t critical value that equals 2.1 for a single t-test (for a single gene) may, depending on the adjustment used and on the choice of reference distribution, increase to 4.5 when there are 5000 such tests.

An attractive design option can be to divide probes into two groups – a smaller "likely" group, and a much larger "possible" group. Statistical comparisons can then be done separately for the two groups, with a much less stringent criterion for establishing differential expression used for probes in the smaller group. The highest priority for the use of repeated spots will be given to the smaller group of genes chosen for careful

scrutiny. Such a classification of genes into two groups builds in prior knowledge, with implications for the subsequent statistical inference.

Repeated spots

What is the effect on precision from repeating probes multiple times on a single slide, by comparison with repeating slides?

Writing m_b for the between array mean square, and m_w for the within array mean square, and with k spots per probe sequence, and assuming a simple form of multi-level model where the between spots (within array) component of variance is σ^2, while the between array component of variance is σ_b^2, it follows that:

$$\begin{aligned} \mathrm{E}[m_b] &= k\sigma_b^2 + \sigma^2 \\ \mathrm{E}[m_w] &= \sigma^2. \end{aligned}$$

Thus $\mathrm{E}[m_b]/\mathrm{E}[m_w]$ equals 1 if $\sigma_b^2 = 0$, and is otherwise greater than one.

The variance of the mean $\bar{x}$ over all k spots on each of n slides is

$$\mathrm{var}[\bar{x}] = \frac{\sigma_b^2}{n} + \frac{\sigma^2}{kn}.$$

If $\sigma_b^2 = 0$, then $\mathrm{var}[\bar{x}] = \frac{\sigma^2}{kn}$, and the repeating of spots is just as effective, for increasing precision, as the repeating of slides.

The Callow *et al.* [6] data are interesting in this connection. Out of 5544 non-blank spots, 175 were duplicates of the same probe sequence, while 6 were triplicates. For each of these probe sequences, we can thus use an analysis of variance calculation to determine both a within array (between spot) mean square, and a between array mean square.

Individual sample ratios are too inaccurate and variable, ranging from 0.11 to 11.2, to give useful indications for experimental design. We can however use a quantile-quantile plot (Figure 1) to study the pattern of change of the ratio over many different genes, and assess the extent to which these ratios behave like independent ratios from an F-distribution with 14 and 16 d.f.

The smallest 156 values are consistent with the assumption that the ratios follow the theoretical F-distribution corresponding to $\sigma_b^2 = 0$, independently between probe sequences. Included among these 156 probe sequences are the only two out of the 181 that were identified as differentially expressed.

Thus for these duplicated or replicated data, for the majority of probe sequences, increasing the number of spots on a slide gives the same improvement in precision as increasing the number of slides by the same factor. There is no way to know whether the same would be true for the probe sequences that were not repeated. Data from a less homogeneous biological population, *e.g.*, tissues from distinct human sources, are inherently likely to show stronger evidence of biological variation. In some types of

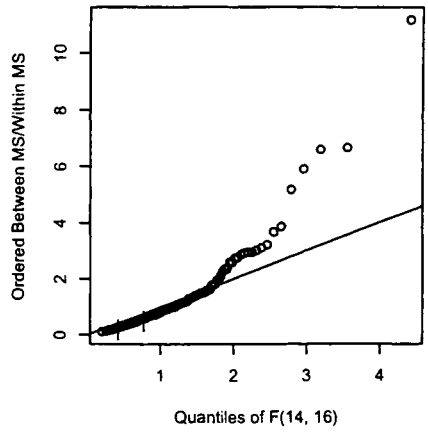


Figure 1: Quantile-quantile plot that compares ordered ratios of between to within slide mean squares, for 181 probe sequences that appear more than once, to quantiles of the F-distribution with 14 and 16 d.f. The line $y = x$ is superimposed on the plot. The two points that correspond to genes identified as differentially expressed are marked with a vertical bar (|).

study, for some probe sequences, increasing the number of spots per gene may be a highly effective way to improve precision.

Note that in more traditional applications of multi-level models, the relevant variances are rarely known with sufficient accuracy that they give a secure basis for use in setting priorities in the future use of experimental resources. For microarray experiments, the combining of information across large numbers of probe sequences can provide such a secure basis. This is an area that requires further investigation.

Published information on mean square ratios such as just given, for a range of different experimental conditions and probe sequence sets, would greatly assist the design of future experiments.

Some special issues for gene chip microarrays

Important chip design issues that require further investigation include the following:

1. There is some evidence that the use of mismatch probes in expression indices reduces precision; see [28, 39]. Further research is required on the optimal assessment of nonspecific hybridization and background.

2. Some probes appear consistently unresponsive, arguing for their removal or replacement.

3. The inclusion of control probe sets can assist quality control and calibration.

In designing experiments, consideration should be given to the inclusion of "spikes" of known concentration in the sample, to allow for more accurate normalization between chips.

7 Discussion

While statistical methodology is now seen as an important part of microarray experiments, its penetration into this area remains, in many respects, superficial. This is especially true for experimental design. Effort at the design phase of a microarray experiment will often save considerable effort and frustration at the analysis stage; see Yang and Speed [56] for further discussion. Good experimentation can be seen as a sequential learning process in that what has been learned from one experiment can contribute to the design of the next experiment.

This paper outlines many of the issues that require consideration when designing a microarray experiment. There has been emphasis on replication and sources of error because of their pivotal role in analysis and subsequently inference. For example, in a comparative experiment researchers should consider that an observed difference is 'real' only if it is greater than what could be expected by chance. The estimate of the size of that difference is a function of all the noise that has contributed to the difference, and is obtained from replicates. Too often, the need for replication has been overlooked in microarray experiments. Yet recall Fisher's [16] comment over seventy years ago concerning plant experimentation:

> No one would now dream of testing the response to a treatment by comparing two plots, one treated and the other untreated.

It is unusual when measuring with, say, a tape measure, to make replicate measurements on the same object. The accuracy of the instrument is commonly high relative to the variability of the object that is measured. Hopefully, technological improvements will lead to arrays with correspondingly high levels of technical reproducibility. In the meantime, there are large potential gains that may come from a better understanding both of the technology and of quantitative aspects of gene expression. Experiments that will assist in an understanding of the technical characteristics of this methodology and the sources of variation and bias should be a priority.

Combining information from the different platforms and laboratories also is important (see, for example, Glynne *et al.* [22]). As yet, we are not aware of studies that directly investigate the extent to which results from a microarray experiment can be reproduced by other workers in other laboratories. If, however, results from some microarray studies point in one direction and some in another, it may be necessary to undertake a statistical overview analysis, or meta-analysis, such as is done in clinical medicine (see for example, Chalmers and Altman [7]). In a related context, Ionnidis *et al.* [27] examined the extent to which genetic association studies stand up when repeated by other researchers, and found that results from the first study often suggest

a stronger effect than is found in later studies, and show poor correlation with subsequent research on the same association. This observation may be in part a manifestation of the so-called "file drawer problem" [43], that positive results are more likely to be published than negative results. Epistatic effects such as are discussed in Wilson [53] provide another likely explanation.

The challenges that arise from the massively parallel measurement of gene expression are new. At the analysis stage, what choice of designs will ease the task of interpreting and summarizing the potentially huge number of individual results? This is clearly an area for further research. Meanwhile, we recommend the use of designs that are both reasonably robust against unexpected behavior, and that are also capable of revealing effects that have not been anticipated.

Acknowledgements

We thank Dr. Aude Fahrer, Dr. Matthew Wakefield and the anonymous referee for helpful comments, which led to several improvements.

John H. Maindonald, Centre for Bioinformation Science, Mathematical Sciences Institute and John Curtin School of Medical Research, Australian National University, Canberra ACT 0200, Australia, john.maindonald@anu.edu.au

Yvonne Pittelkow, Centre for Bioinformation Science, Mathematical Sciences Institute and John Curtin School of Medical Research, Australian National University, Canberra ACT 0200, Australia, yvonne.pittelkow@anu.edu.au

Susan Wilson, Centre for Mathematics and its Applications, Mathematical Sciences Institute and Centre for Bioinformation Science, Mathematical Sciences Institute and John Curtin School of Medical Research, Australian National University, Canberra ACT 0200, Australia, sue.wilson@anu.edu.au

References

[1] Affymetrix. *Affymetrix Microarray Suite User Guide, Version 4 edition.* Affymetrix, Santa Clara, CA.

[2] U. Alon, N. Barkai, D. A. Notterman, K. Gish, D. Mack S. Ybarra, and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by microarray chips. *Proceedings of the National Academy of Sciences, USA*, 96:6745–6750, 1999.

[3] M. Bakay, Y.-W. Chen, R. Borup, P. Zhao, K. Nagaraju, and E. P. Hoffman. Sources of variability and effect of experimental approach on expression profiling data interpretation. *BMC Bioinformatics*, 3:4, 2002.

[4] L. R. Baugh, A. A. Hill, E. L. Brown, and C. P. Hunter. Quantitative analysis of mRNA amplification by *in vitro* transcription. *Nucleic Acids Research*, 29(5):e29, 2001.

[5] G. Box, W. Hunter, and S. Hunter. *Statistics for Experimenters*. Wiley, New York, 1978.

[6] M. J. Callow, S. Dudoit, E. L. Gong, and E. M. Rubin. Microarray expression profiling identifies genes with altered expression in HDL-deficient mice. *Genome Research*, 10:2022–2029, 2000.

[7] I. Chalmers and D. G. Altman. *Systematic Reviews*. BMJ Publishing Group, London, 1995.

[8] E. Chudin, R. Walker, A. Kosaka, S. X. Wu, D. Rabert, T. K. Chang, and D. E. Kreder. Assessment of the relationship between signal intensities and transcript concentration for Affymetrix GeneChip® arrays. *Genome Biology*, 3(1):research0005, 2001.

[9] D. R. Cox. *Planning of Experiments*. Wiley, New York, 1958.

[10] D. R. Cox and N. Reid. *Theory of the Design of Experiments*. Chapman and Hall, London, 2000.

[11] H. A. David. *The Method of Paired Comparisons*. Oxford University Press, New York, 1988.

[12] S. Dudoit, Y. H. Yang, and B. Bolstad. Using R for the analysis of DNA microarray data. *R News*, 2:24–32, 2002. `http://cran.R-project.org/doc/Rnews`.

[13] S. Dudoit, Y. H. Yang, M. J. Callow, and T. P. Speed. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, 12:111–140, 2002.

[14] B. Efron, R. Tibshirani, J. D. Storey, and V. Tusher. Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96:1151–1160, 2001.

[15] R. A. Fisher. *The Design of Experiments*. Oliver and Boyd, Edinburgh, 1935; 7th edition 1960.

[16] R. A. Fisher and J. Wishart. The arrangement of field experiments and the statistical reduction of the results. *Imperial Bureau of Soil Science (London). Technical Communication*, 10:1–23, 1930.

[17] S. H. Friend and R. B. Stoughton. The magic of microarrays. *Scientific American*, 286:34–39, 2002.

[18] M. D. Gacula and J. Singh. *Statistical Methods in Food and Consumer Research*. Academic Press, Orlando, FL, 1984.

[19] R. Gentleman and V. Carey. Bioconductor. Open source bioinformatics using R. *R News*, 2:11–17, 2002. `http://cran.R-project.org/doc/Rnews`.

[20] G. Gibson and S. V. Muse. *A Primer of Genome Science*. Sinauer Associates, Madison, WI, 2001.

[21] R. J. Glynne, S. Akkaraju, J. I. Healy, J. Rayner, C. C. Goodnow, and D. H. Mack. How self-tolerance and the immuno-suppressive drug FK506 prevent B-cell mitogenesis. *Nature*, 403:672–676, 2000.

[22] R. J. Glynne, G. Ghandour, and C. C. Goodnow. Genomic-scale expression analysis of lymphocyte growth, tolerance and malignancy. *Current Opinion in Immunology*, 12:210–214, 2000.

[23] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

[24] P. Hegde, R. Qi, K. Abernathy, C. Gay, S. Dharap, R. Gaspard, J. Earle-Hughes, E. Snesrud, N. Lee, and J. Quackenbush. A concise guide to cDNA microarray analysis. *Biotechniques*, 29:548–562, 2000.

[25] A. A. Hill, E. L. Brown, M. Z. Whitley, G. Tucker-Kellogg, C. P. Hunter, and D. K. Slonim. Evaluation of normalization procedures for oligonucleotide microarray data based on spiked cRNA controls. *Genome Biology*, 2(12):research0055, 2001.

[26] L.-L. Hsaio, R. V. Jensen, T. Yoshida, K. E. Clark, J. E. Blumenstock, and S. R. Gullans. Short technical report: Correcting for signal saturation errors in the analysis of microarray data. *Biotechniques*, 32:330–336, 2002.

[27] J. P. A. Ioannidis, E. E. Ntzani, T. A. Trikalinos, and D. G. Contopoulos-Ioannidis. Replication validity of genetic association studies. *Nature Genetics*, 29:306–309, 2001.

[28] R. A. Irizzary, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 2002. In press.

[29] N. N. Iscove, M. Barbara, M. Gu, M. Gibson, C. Modi, and N. Winegarden. Representation is faithfully preserved in global cDNA amplified exponentially from sub-picogram quantities of mRNA. *Nature Biotechnology*, 20:940–943, 2002.

[30] M. K. Kerr and G. A. Churchill. Experimental design for gene expression microarrays. *Biostatistics*, 2:183–201, 2001.

[31] M.-L. T. Lee, F. C. Kuo, G. A. Whitmore, and J. Sklar. Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proceedings of the National Academy of Sciences, USA*, 97:9834–9839, 2000.

[32] W. T. Lemon, J. T. Palatini, R. Krahe, and F. A. Wright. Theoretical and experimental comparisons of gene expression estimators for oligonucleotide arrays. *Bioinformatics*, 18:1470–1476, 2002.

[33] C. Li and W. Wong. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of the National Academy of Sciences, USA*, 98:31–36, 2001.

[34] C. Li and W. Wong. Model-based analysis of oligonucleotide arrays: Model validation, design issues and standard error application. *Genome Biology*, 2:1–11, 2001.

[35] D. J. Lockhart, H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E. L. Brown. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 14:1675–1680, 1996.

[36] J. H. Maindonald. Statistical design, analysis and presentation issues. *New Zealand Journal of Agricultural Research*, 35:121–141, 1992.

[37] J. Mar and S. Grimmond. A review of image analysis software for spotted microarrays. Technical Report, 2002.

[38] J. C. Mills and J. I. Gordon. A new approach for filtering noise from high-density oligonucleotide microarray datasets. *Nucleic Acids Research*, 29(15):e72–2, 2001.

[39] F. Naef, D. A. Lim, N. Patil, and M. O. Magnasco. From features to expression: High density oligonucleotide microarray analysis revisited. LANL e-print physics/0102010. To appear in the Proceedings of the DIMACS Workshop on Analysis of Gene Expression Data, 2001.

[40] J. P. Novak, R. Sladek, and T. J. Hudson. Characterization of variability in large-scale gene expression data: Implications for study design. *Genomics*, 79:104–113, 2002.

[41] L. Ramdas, K. R. Coombes, K. Baggerly, L. Abruzzo, W. E. Highsmith, T. Krogmann, S. R. Hamilton, and W. Zhang. Sources of nonlinearity in cDNA microarray expression measurements. *Genome Biology*, 2(11):research0047, 2001.

[42] G. K. Robinson. *Practical Strategies for Experimenting*. Wiley, New York, 2000.

[43] R. Rosenthal. The "file drawer problem" and tolerance for null results. *Psychological Bulletin*, 86:638–641, 1979.

[44] J.-M. Rouillard, C. J. Herbert, and M. Zuker. Oligarray: genome-scale oligonucleotide design for microarrays. *Bioinformatics*, 18:486–487, 2001.

[45] G. Sawitzki. Quality control and early diagnostics for cDNA microarrays. *R News*, 2:6–9, 2002. `http://cran.R-project.org/doc/Rnews`.

[46] E. Schadt, C. Li, C. Su, and W. H. Wong. Analyzing high-density oligonucleotide gene expression array data. *Journal of Cellular Biochemistry*, 80:192–202, 2000.

[47] G. K. Smyth, Y. H. Yang, and T. P. Speed. Statistical issues in cDNA microarray data analysis. In *Functional Genomics: Methods and Protocols*. Humana Press, 2002.

[48] T. P. Speed. What is an analysis of variance? *Annals of Statistics*, 15:885–910, 1987.

[49] T. P. Speed. *Statistical Analysis of Gene Expression Microarray Data*. CRC Press, 2002. In press.

[50] T. P. Speed and Y. H. Yang. Direct versus indirect designs for cDNA microarray experiments. Technical Report # 616, 2002.

[51] G. C. Tseng, M.-K. Oh, L. Rohlin, J.-C. Liao, and W.-H. Wong. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Research*, 29:2549–2557, 2001.

[52] M. Vingron. Editorial. Bioinformatics needs to adopt statistical thinking. *Bioinformatics*, 17:389–390, 2001.

[53] S. R. Wilson. Epistasis. In *Encyclopedia of the Human Genome*. Macmillan, 2003. In press.

[54] Y. H. Yang, M. J. Buckley, S. Dudoit, and T. P. Speed. Comparison of methods for image analysis on cDNA microarray data. *Journal of Computational and Statistical Graphics*, 11:108–136, 2002.

[55] Y. H. Yang, S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, and T. P. Speed. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, 30(4):e15, 2002.

[56] Y. H. Yang and T. P. Speed. Design issues for cDNA microarray expression experiments. *Nature Reviews*, 3:579–588, 2002.

[57] M. Zuker. Calculating nucleic acid secondary structure. *Current opinion in structural biology*, 10:303–310, 2000.

Measures of Gene Expression for Affymetrix High Density Oligonucleotide Arrays

Rafael A. Irizarry

Abstract

High density oligonucleotide expression array technology is widely used in many areas of biomedical research for quantitative and highly parallel measurements of gene expression. In Affymetrix GeneChip array technology, each gene is typically represented by a set of 11-20 pairs of oligonucleotides, separately referred to as probes, arrayed on a silicon chip. After chip measurements are preprocessed, a fluorescence intensity value for each probe is obtained. A necessary step for defining a measure of expression (*ME*) is to summarize the probe intensities for a given gene. In this paper, we review the ideas that motivate a summary statistic, referred to as the robust multi-array average (*RMA*), that improves the default Affymetrix approach and provides substantial benefits to users of the GeneChip technology.

Keywords: Affymetrix GeneChip arrays; background correction; gene expression; normalization; summary measure

1 Introduction

High density oligonucleotide expression array technology is widely used in many areas of biomedical research for quantitative and highly parallel measurements of gene expression. Affymetrix GeneChip arrays use oligonucleotides of length 25 base pairs to probe genes. In this technology, each gene is typically represented by a set of 11-20 pairs of oligonucleotides, separately referred to as *probes*, arrayed on a silicon chip. Details of this array technology are described by [1] and [10]. Briefly, though, RNA samples are prepared according to a specific protocol. A fluorescently labeled RNA sample is hybridized to probes on the chip. After some processing steps, the array is scanned with a laser. This scan produces an image that is analyzed to produce an intensity value for each probe (see [9] for more details). These intensities quantify the extent of the hybridization between the labeled target sample and the oligonucleotide probe. A final step to obtain a measure of gene expression (*ME*) is to summarize the intensities for a given gene in order to quantify the amount of corresponding mRNA species in the sample. The intensities obtained for each probe are denoted by PM_{ijn} and MM_{ijn}, $i = 1,\ldots,I$, $j = 1,\ldots,J_n$, and $n = 1,\ldots,N$, with i representing different RNA samples, j representing the probe pair number (this number is related to the physical position of

the oligonucleotide in the gene), and n representing the different genes. The number of genes N usually ranges from 8,000 to 20,000, the number of arrays I is usually small but may be as large as a few hundred, and the number of probe pairs within each gene J_n usually ranges from 11 to 20. Throughout the text, indices are suppressed when there is no ambiguity.

Several researchers have found problems with the *ME* provided by the first version of the Affymetrix system [1] and have suggested alternatives, the most cited example being that of Li and Wong [7]. In their most recent version, Affymetrix provides an alternative as well [2]. There are papers in the literature that compare *ME* by assessing variances, see [8] for an example. Typically, *ME* are obtained from arrays hybridized to RNA aliquots (technical replicates). Throughout the text, we denote the *ME* obtained for a given gene by E_i, with $i = 1,\ldots,I$ representing arrays. When there are replicate arrays, we define the sample variance as $\hat{\sigma}^2 = \sum_{i=1}^{I}(E_i - \bar{E})^2$ with $\bar{E}$ representing the average. *ME* that, in general, have smaller $\hat{\sigma}$ are considered better. However, without an accompanying assessment of the ability to detect signal (which can be thought of as assessing bias), this could produce misleading results. For example, a *ME* that is always $E_k = 0$ cannot be considered appropriate because of its small variance.

Irizarry *et al.* [6] carried out a comparison study of *ME* using two data sets: (*i*) part of the data from an extensive spike-in study conducted by GeneLogic and the Genetics Institute involving about 95 HGU95A human GeneChip arrays, and (*ii*) part of a dilution study conducted by GeneLogic involving 75 HGU95A GeneChip arrays. Four *ME* are compared: (*i*) the Affymetrix commercial software MicroArray Suite MAS 4.0 default (AvDiff) (*ii*) their updated software MAS 5.0 default, (*iii*) the Li and Wong [7] multiplicative model-based *ME*, and (*iv*) a summary based on a log-scale additive model, referred to as the log-scale robust multi-array average (*RMA*). This study seems to be the first to compare *ME* and also to check the reliability of the technology with data for which both bias and variance can be assessed. They find that in general the technology works well, and also that *RMA* outperforms the other three *ME*. In this paper, we give a brief overview of these findings, propose a statistical framework for data using these arrays, and demonstrate with an example why *RMA* works better.

2 Methods

2.1 Background Correction

Several processes can affect the intensities read from each probe. Apart from the specific hybridization directly related to the quantity to be measured, there is also background (or optical noise), nonspecific hybridization, and cross-hybridization. The Affymetrix strategy for extracting the signal of interest from the observed *PM* (perfect match) intensity is to subtract the corresponding *MM* (mismatch) probe intensity. In MAS 4.0, an *ME* for a gene is formed by considering the average difference (AvDiff) of the *PM* and *MM* in the probe set. More precisely, an *ME* for a gene is formed by

defining

$$\text{AvDiff} = N_A^{-1} \sum_{j \in A} (PM_j - MM_j) \tag{1}$$

with A the subset of probes for which $d_j = PM_j - MM_j$ are within 3 SDs away from the average of $d_{(2)}, .., d_{(J-1)}$, where $d_{(j)}$ is the j^{th} smallest difference. N_A represents the number of probes in A. The MAS ME and the ME from the Li and Wong reduced model, discussed in more detail in Section 2.4, are also based on $PM - MM$. Dividing instead of subtracting, *i.e.* using PM/MM, has also been suggested.

The rationale for using the $PM - MM$ quantities is that they correct the effects that bias the PM quantities. Another measure offered in MAS 4.0 software is an average based on the log of ratios PM/MM. There may be biological or physical motivation for considering differences (or ratios). We believe, though, that it is important to corroborate such assumptions empirically.

Figure 1 shows intensities of the PM, MM, PM/MM and $PM - MM$ values for each of the 20 probes representing the BioB-5 probe set in a set of 12 arrays. BioB-5 has been spiked-in on the 12 different arrays at concentrations of 0.5, 0.75, 1, 1.5, 2, 3, 5, 12.5, 25, 50, 75, and 150 picoMolar. All arrays had a common background cRNA from an acute myeloid leukemia (AML) tumor cell line. All plots in Figure 1 are on the log scale except for 1c. The low values of the $PM - MM$ are plotted on a linear scale because there are several negative values (in fact about 1/3 of the non-spiked in probes have $PM - MM < 0$). The 20 different probe pairs are represented with different symbols and colors. As expected, the PM values are growing in proportion to the concentration. Notice also that the lines representing the 20 probes are close to being parallel showing that there is a strong additive (in the log scale) probe-specific effect. The fact, seen in Figure 1b, that the additive probe-specific effect is also detected by the MM provides motivation for subtracting these values from the PM. However, in Figures 1c and 1d the parallel lines are still seen in $PM - MM$, demonstrating that subtracting is not enough to remove the probe effect. The lack of parallel lines in Figure 1e shows that dividing by MM removes, to some degree, the probe effect. However, since the MM also grow with concentration, and therefore detect signal as well as non-specific binding, results in an attenuated signal. Notice in particular that using PM/MM would make concentrations of 25 and 150, a six-fold difference, indistinguishable. The $PM - MM$ demonstrate some attenuation for the high concentration spike-ins but clearly not as much as PM/MM. Since subtracting probe-specific MM adds noise with no obvious gains in signal detection, and because PM/MM results in a biased signal, [6] propose background correction approaches which are different from subtracting or dividing by MM. We now give a brief review.

The horizontal lines in Figure 1 represent the median intensity obtained from an array for which no spike-in for BioB-5 was added. The dashed lines represent the first and third quartiles. For the lower concentrations, it is hard to distinguish the measured intensities from this median value. Notice also that the signal is attenuated for the lower concentrations. A possible explanation is that background correction is needed.

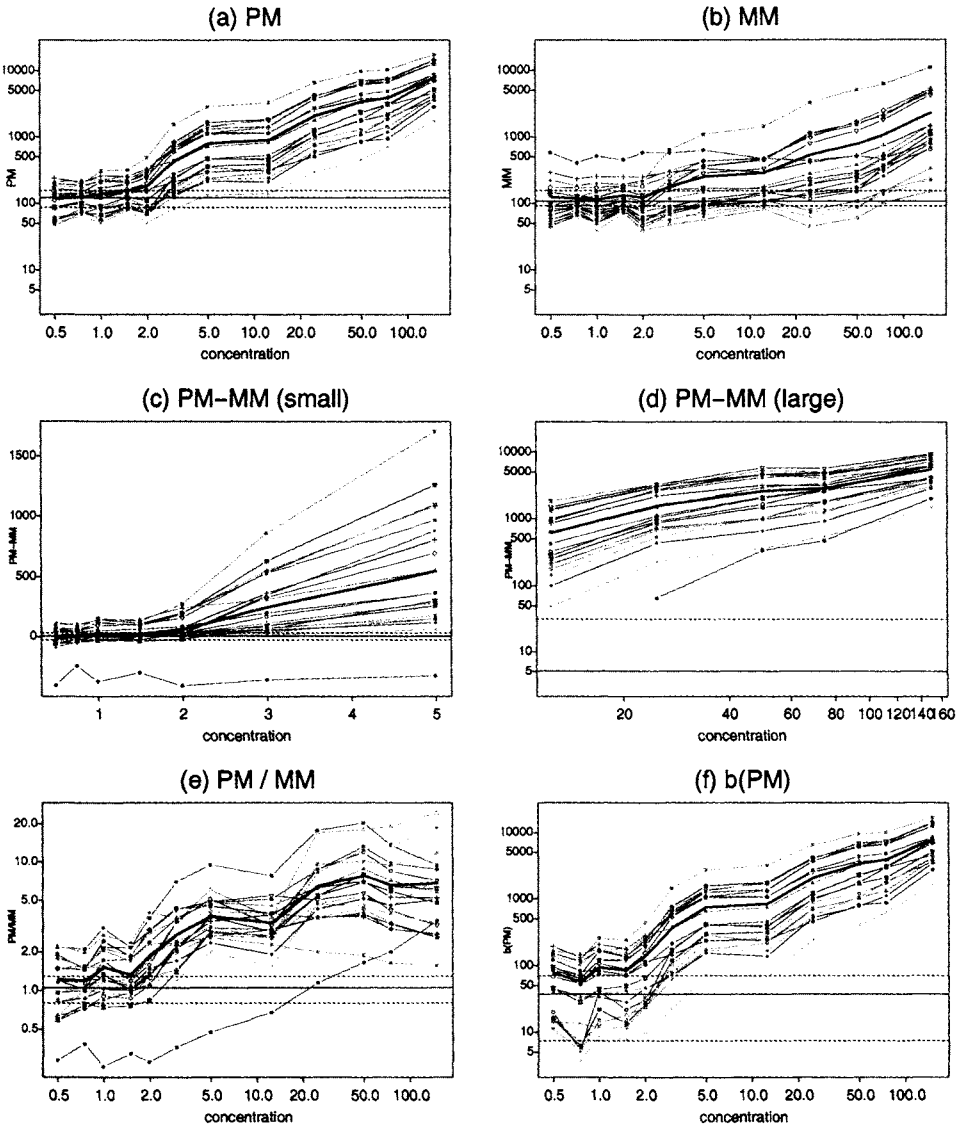


Figure 1: *PM*, *MM*, *PM*/*MM*, $PM - MM$, and $b(PM)$ intensities, for each of the 20 probes representing BioB-5 in 12 arrays where the probe set has been spiked-in, plotted against concentration. Except for 1(c), axes are on the log scale. Different probes are represented by the different colors and symbols. The horizontal line represents the median of the 20 BioB-5 probes from an array where no spike-in was added. The dashed lines are at the 25th and 75th quantiles.

To see this, consider a hypothetical case with two arrays where the signals of a probe set is twice as big in one of the arrays, but an additive signal of 100 units occurs due

to non-specific binding and/or background noise in both arrays. In this case, the observed difference in the signals would be about $\log_2(100+2s) - \log_2(100+s)$ instead of $\log_2(2s) - \log_2(s) = 1$. For small values of s, the incorrect difference would instead be close to 0.

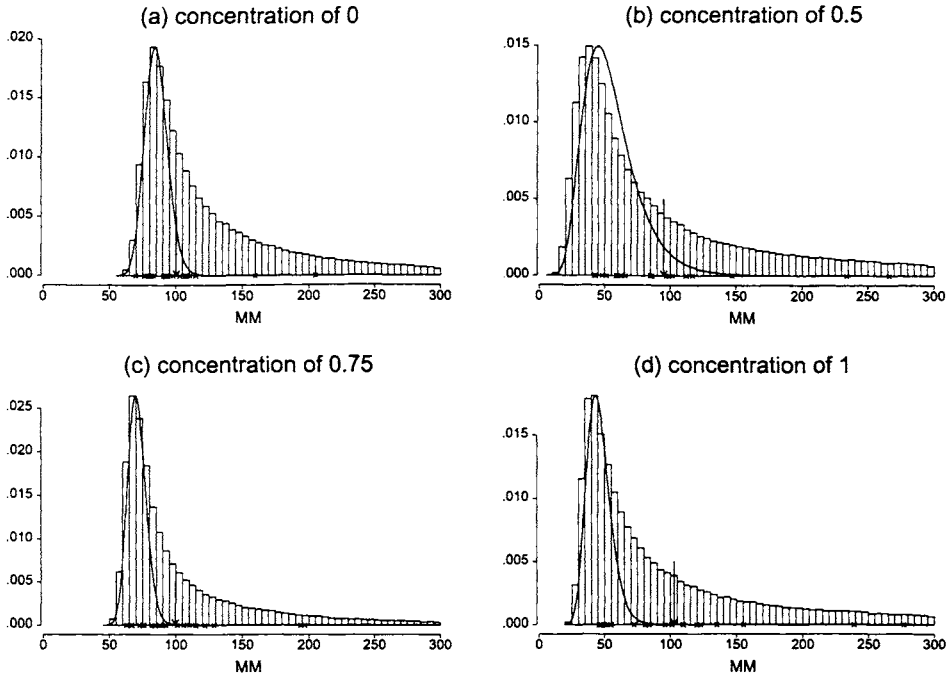


Figure 2: Histograms (density scale) of $\log_2(MM)$ for an array in which no probe set was spiked along with the 3 arrays in which BioB-5 was spiked-in at concentrations of 0.5, 0.75, and 1 picoMolar. The observed *PM* values for the 20 probes associated with BioB-5 are marked with crosses and the average with an arrow.

Figure 2 shows histograms of *MM* for an array in which no probe set was spiked, along with the 3 arrays in which BioB-5 was spiked-in at concentrations of 0.5, 0.75, and 1 picoMolar. The observed *PM* values for the 20 probes associated with BioB-5 are marked with crosses and the average with an arrow. All the average *PM* values are close to 100. Thus, based solely on the average, a difference would be hard to detect. Figures 2 and 3 suggest that the *MM* to the left of the mode of the histogram are similar to the left half of a normal distribution. This suggests that the *MM* are a mixture of (*i*) probes for which an intensity is read due to non-specific binding and background noise and (*ii*) probes detecting transcript signal (cross-hybridization) just like the *PM*. The distance of the average *PM* from the average background noise does in fact increase with concentration. This suggests that background correction of the data is necessary. As noted earlier, $PM - MM$ is not a solution we recommend.

The approach suggested by [6] is to use a global, instead of probe specific, back-

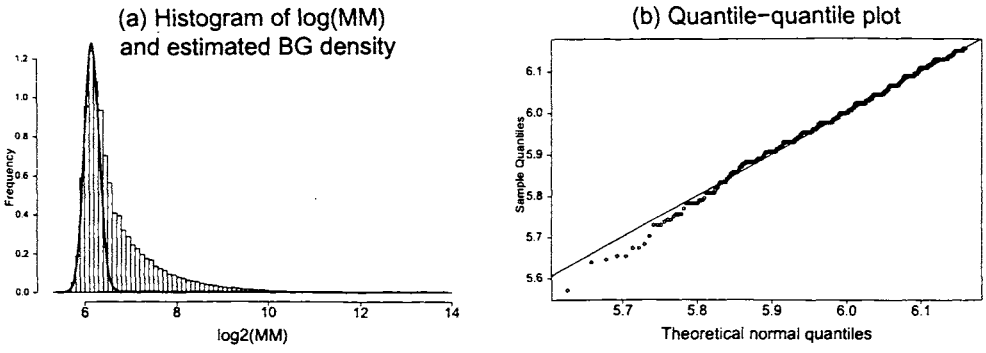


Figure 3: (a) Histogram of $\log_2(MM)$ for spike-in concentration 12.5 picoMolar array in the varying concentration series. (b) QQ plot of the MM left of the mode of histogram (a) compared to a log-normal distribution with mean and SD estimated from the data.

ground correction. We assume that the observed intensity for each PM probe is the sum of a specific binding component and a background component (which may include non-specific binding). Denote these by $PM = S + B$. Because we are interested in S, we use $b(PM) = \mathrm{E}[S|PM]$. We refer to b as a background correcting transformation. Irizarry *et al.* [6] assume B is normally distributed and that S follows an exponential distribution. This assumption is convenient because in this case there is a closed-form solution to $\mathrm{E}[S|PM]$. The solution depends on the mean and variance of the normal distribution and the rate of the exponential distribution. These parameters can be estimated from the PM and MM probe level data. Figure 1f shows the background-corrected PM for the BioB-5 probes. After background transformation, the low concentration values can be distinguished from the values obtained for the array with no spike-in (represented by the horizontal line). In addition, the fact that the slope is larger for the low concentrations in Figure 1f than in Figure 1a demonstrates that the signal is less attenuated for low intensities. However, the intensity values for PM and $b(PM)$ do not grow as a straight line (in the log scale). Further improvements may be obtained with array normalization.

2.2 Normalization

In many of the applications of high density oligonucleotide arrays, the goal is to learn how RNA populations differ in expression in response to genetic and environmental differences. For example, large expression of a particular gene or genes may cause an illness resulting in variation between diseased and normal tissue. Observed expression levels also include variation introduced during sample preparation and array manufacture and processing. Unless arrays are appropriately normalized, comparisons of data from different arrays can lead to misleading results. One approach is *quantile normalization* [6], which forces the empirical distributions of probe intensities from all arrays

to be equal. The approach works well in practice, see [3] for details.

2.3 Statistical Models

Figure 1f demonstrates that the background corrected probe intensities follow an additive model in the log scale. Irizarry *et al.* [6] propose the following model for each probe set

$$\log_2\{b(PM_{ij})\} = \mu_i + \alpha_j + \varepsilon_{ij}, \quad i = 1,\ldots,I, \quad j = 1,\ldots,J, \tag{2}$$

with μ_i representing the log scale *ME* for array i, α_j a probe affinity effect, and ε_{ij} representing an independent identically distributed error term with mean 0. For identifiability of the parameters, we assume that $\sum_j \alpha_j = 0$ for all n. This assumption is equivalent to saying that Affymetrix technology has chosen probes with expected intensities that on average are representative of the associated gene expression.

Under model (2), an unbiased estimate of μ_i, the log scale *ME* for each array, can be obtained using the average

$$\hat{\mu}_i = J^{-1} \sum_{j=1}^{J} \log_2\{b(PM_{ij})\}. \tag{3}$$

Model (2) lends itself to various practical extensions. For example, to compare two populations of RNA species for which there are technical replicates assumed to have the same expected RNA expression, we can write

$$\log_2\{b(PM_{ij}^a)\} = \mu_i^a + \alpha_j + \varepsilon_{ij}^a, \quad i = 1,\ldots,I, \quad j = 1,\ldots,J, \quad a = 1,2.$$

Here i denotes replicate and a the population. The natural estimate of μ^a would be based on I times more data than (3). If instead of technical replicates there were biological replicates, a term Z_{ij}, representing a random effect, could be added to the model.

Li and Wong [7] demonstrate that estimation procedures that remove outliers reduce the variance of *ME* estimates. Model (2) can be easily extended to a context that motivates robust estimates of μ. We refer to the *ME* obtained from estimating μ in model (2) using a robust method, such as the median polish approach used by [4] or robust linear regression, as *RMA* (robust multi-array average).

2.4 Measures of Expression

Figure 4 shows a standard deviation versus average probe intensities scatter-plot from a random sample of *PM* and *MM* obtained from five replicate arrays. Figure 4a shows that the SD increases from roughly 50 to 5000, a factor of 100 fold, as the average increases on its entire range. Figure 4b shows that after a log transformation of the intensities there is only a 1.5 fold increase. This makes the log scale a more natural scale for operations such as averaging. Apparently Affymetrix has also noticed this and,

unlike the MAS 4.0 *ME* AvDiff, their MAS 5.0 *ME* is based on a log scale average. Specifically, for each probe set the MAS 5.0 signal (measure) is defined as

$$\text{signal} = \exp\{\text{Tukey Biweight}(\log(PM_j - CT_j))\}$$

with $CT_j = MM_j$ if $PM_j > MM_j$; if $PM_j < MM_j$, then CT_j is a quantity derived from the *MM* that is never bigger than its *PM* pair. See [5] for more details.

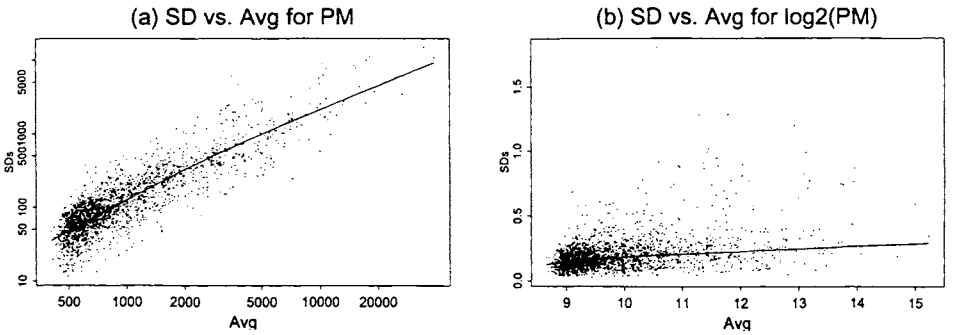


Figure 4: Standard deviations (SDs) plotted against averages from 5 MGU74A mouse arrays for a random sample of 2000 defective probe sets for (a) *PM* and (b) $\log 2(PM)$. The curves are loess fits.

Li and Wong [7] propose using the following model to obtain *ME*:

$$PM_{ij} - MM_{ij} = \theta_i \phi_j + \varepsilon_{ij}, \tag{4}$$

with ϕ_j representing the probe-specific affinities and independent identically distributed mean 0 normally distributed errors ε_{ij}. For each probe set, an *ME* is defined as the maximum likelihood estimate of θ_i, $i = 1,\ldots,I$ obtained from fitting the multiplicative model. The estimation procedure includes rules for outlier removal. For computational speed, Li and Wong [7] use an iterative procedure that leads to estimates of the form

$$\hat{\theta}_i = \frac{\sum_{j=1}^{J}(PM_{ij} - MM_{ij})\hat{\phi}_j}{\sum_{j=1}^{J}\hat{\phi}_j}, \tag{5}$$

which is basically a weighted version of (1), although their algorithm does remove outliers. This means that probes that are in general high will have a larger influence on $\hat{\theta}_i$. If in fact (2) is a better approximation than (4), then (5) leads to an expression measure with larger variance than *RMA* (see [6]).

3 Results and Discussion

There is no gold standard to compare and test summaries of probe level data. For this reason, data from spike-in experiments have been used to assess the technology

and to motivate normalization procedures. In a similar way, [6] used data from spike-in and dilution experiments to assess the MAS 4.0, MAS 5.0, Li and Wong [7], and *RMA* expression measures. These data are especially useful because here there is an expected result. Irizarry *et al.* [6] demonstrate through examples that *RMA* provides more precise estimates of expression, as well as better specificity and sensitivity for detection of differential expression, than the other three measures. In this section, we give some specific examples that demonstrate why *RMA* performs better.

Figure 5 shows MVA plots: log ratios (or log fold changes) $M_n = \log(E_{1n}/E_{2n})$ versus average expression $A_n = \log(\sqrt{E_{1n}E_{2n}}) = (\log E_{1n} + \log E_{2n})/2$ for *ME* E_{1n} and E_{2n} for all genes, n=1,...,N on two arrays. The arrays being compared here are part of the spike-in experiment described in [6]. We show MVA plots for *ME* obtained using MAS 5.0, Li and Wong [7], and *RMA*. To be able to fit the Li and Wong model and to use a median polish for *RMA*, we compute *ME* using all 33 arrays that were part of the experiment. Because MAS 5.0 is an improved version of MAS 4.0 [2, 6], MAS 4.0 is not shown in Figure 5. The two arrays have 11 control genes spiked-in at different concentrations, but for illustrative purposes we show only DapX-M, which has been spiked in at concentrations of 2 picoMolar and 1 picoMolar on the two arrays respectively. The log ratio for DapX-M should be about 1, corresponding to a fold change of about 2. All other genes represented in the MVA plots should have log ratios of 0 (fold changes of 1, or equal expression) because the samples hybridized to the arrays represent the same biological assay. In the figures, genes having bigger observed fold changes than DapX-M (false positives) are represented with big dots. Only *RMA* has no false positives here. All measures result in an observed log fold change for DapX-M of over 2, which is quite different from 1. Error associated with adding the spike-in to the hybridization sample may account for this difference.

The barplots in Figure 5 show the *PM* and *MM* values for DapX-M and for two other genes that produce false results. One had a large fold change (false positive) estimated from the Li and Wong model (4), the other had a large fold change estimated from MAS 5.0. The barplots show why subtracting the *MM* can cause problems. Notice in particular the 11th probe in DapX-M, where the *MM* are several times higher than the *PM*. They also demonstrate why giving large weight to probes with high values can produce misleading results. For example, probe 13 in the set 33007_at, which is not called an outlier by the Li and Wong algorithm, will have a large weight. Numerical results obtained from these genes are given in Table 1. Table 1 shows that different results can be obtained by using the different *ME*. The values shown in the barplot for probe set 33658_at suggest that there is no fold change occurring for that gene. However, the MAS 5.0 *ME* gives a log ratio of 1/40. The variance added by subtracting the *MM* values causes MAS 5.0 to incorrectly assign a large fold change to this gene.

A possible explanation for why *RMA* outperforms the Li and Wong model is that model (2) fits the data better than (4). The following example supports this explanation. The method of Li and Wong provides not only an estimate of θ_i but a nominal SE for this estimate, denoted here with $\hat{\sigma}_i$. Under (2), one can obtain a naive nominal estimate

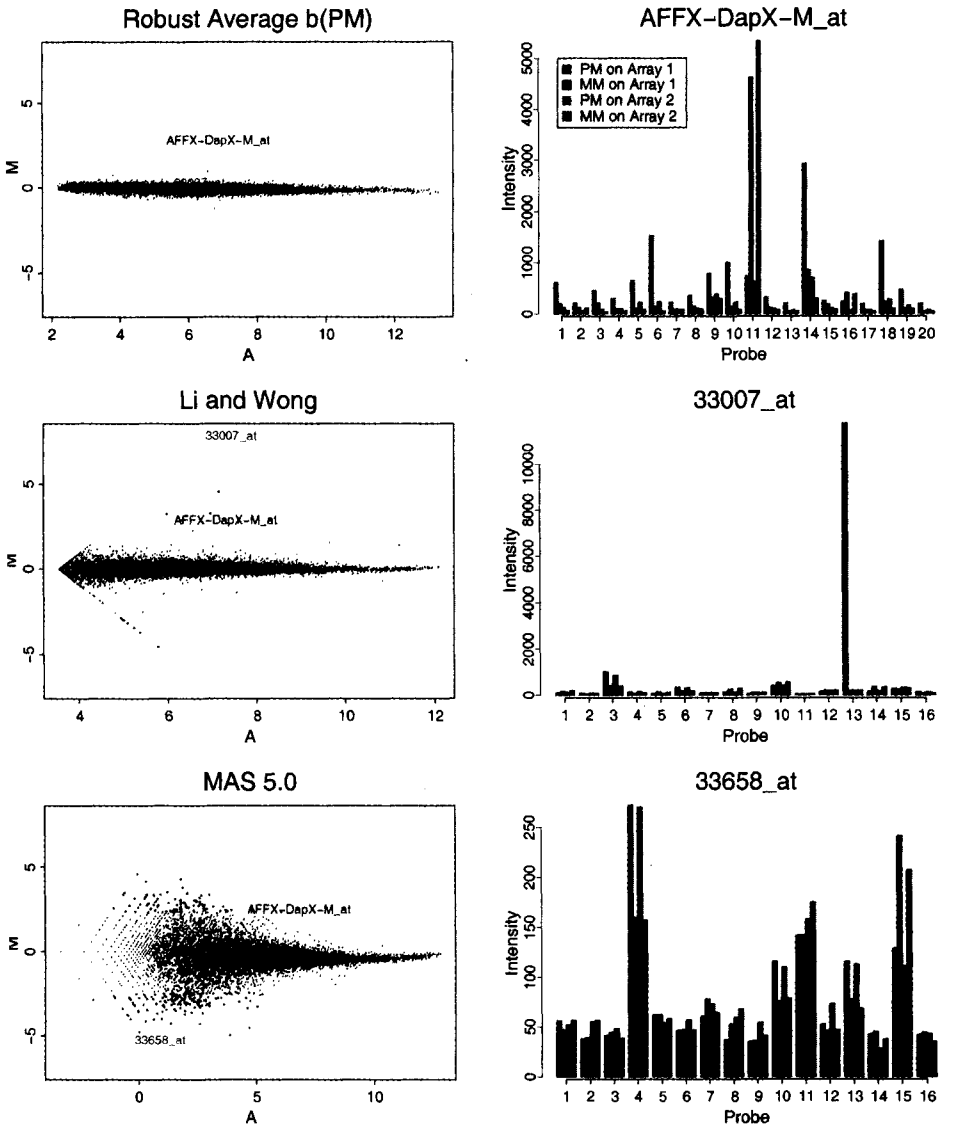


Figure 5: MVA plots indicating the position of the DapX-M which was spiked in at a concentration of 2:1. Barplot for the three genes highlighted in the MVA-plots.

for the SE of $\hat{\mu}$ using an analysis of variance approach. Because there are five replicates, one can also obtain an observed SE of any estimate by simply considering SD_i. If the model is close to the actual mechanism giving rise to the data, the nominal and observed SE should agree. Figure 6 plots the log ratio of nominal to observed variance versus expression measure. These show that in general, the observed and nominal standard errors are closer when using (2) instead of (4).

Table 1: *ME* obtained using *RMA*, the Li and Wong model, and MAS 5.0 for three different genes shown in Figure 5. Only DapX-M should be found to have true fold change.

Gene	*ME*	Array 1	Array 2	Obs. log_2 ratio	Obs. fold change
DapX-M	*RMA*	296.0	46.1	2.7	6.4
DapX-M	LiWong	414.1	61.1	2.8	6.8
DapX-M	MAS 5.0	256.9	49.8	2.4	5.2
33007_at	*RMA*	85.4	74.6	0.1	1.1
33007_at	LiWong	2595.6	11.7	7.8	222.0
33007_at	MAS 5.0	8.4	3.9	1.1	2.2
33658_at	*RMA*	10.0	9.9	0.0	1.0
33658_at	LiWong	25.3	22.7	0.1	1.1
33658_at	MAS 5.0	0.3	12.0	-5.4	1/40

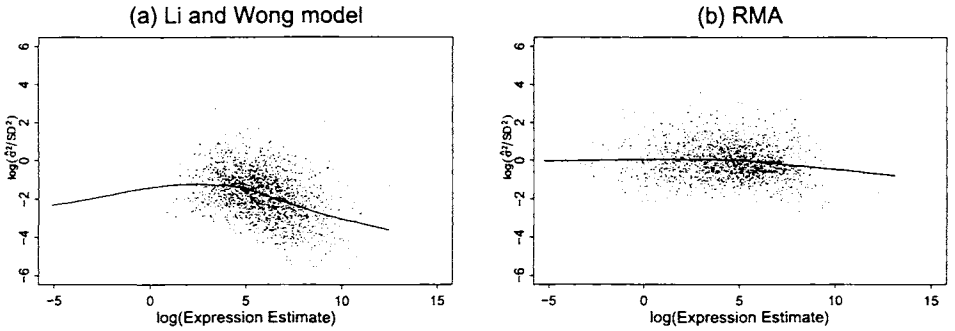


Figure 6: (a) $\log(\hat{\sigma}^2/SD^2)$ plotted against log expression of the Li and Wong *ME*; (b) $\log(\hat{\sigma}^2/SD^2)$ plotted against *RMA*

Irizarry *et al.* [6] developed *RMA*, a summary of Affymetrix GeneChip probe level data, that provides a measure of gene expression, which gives an improved measure compared to other standard measures. The above serves as a specific example demonstrating why *RMA* works better.

4 Acknowledgments

I would like to thank all of the people that have been part of this work, namely: Kristen J. Antonellis, Magnus Åstrand, Yasmin D. Beazer-Barclay, Ben Bolstad, Francois Collin, Leslie Cope, Laurent Gautier, Bridget Hobbs, and Uwe Scherf. I would also like to thank Darlene Goldstein for helpful comments. And a special thanks to Terry Speed. The availability of the dilution experiment, which Terry Speed helped to design, allowed careful assessment and comparison of the different expression measures.

These data are especially useful because we can define outcomes for which there is an expected result.

Rafael A. Irizarry, Department of Biostatistics, Johns Hopkins University, rafa@jhu.edu

References

[1] Affymetrix. *Affymetrix Microarray Suite User Guide.* Affymetrix, Santa Clara, CA, version 4 edition, 1999.

[2] Affymetrix. *Affymetrix Microarray Suite User Guide.* Affymetrix, Santa Clara, CA, version 5 edition, 2001.

[3] B.M. Bolstad, R.A. Irizarry, M. Åstrand, and T.P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics.* In press.

[4] D. Holder, R. F. Raubertas, V. B. Pikounis, V. Svetnik, and K. Soper. Statistical analysis of high density oligonucleotide arrays: a SAFER approach. In *Proceedings of the ASA Annual Meeting, Atlanta, GA 2001*, 2001.

[5] E. Hubbell. Estimating signal with next generation Affymetrix software. In *GeneLogic Workshop on Low Level Analysis of Affymetrix GeneChip® data*, 2001. http://www.stat.berkeley.edu/users/terry/zarray/Affy/GL_Workshop/genelogic2001.html.

[6] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics.* In press.

[7] C. Li and W. H. Wong. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of the National Academy of Science U S A*, 98:31–36, 2001.

[8] C. Li and W. H. Wong. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biology*, 2:1–11, 2001.

[9] R. Lipshutz, S. Fodor, T. Gingeras, and D. Lockhart. High density synthetic oligonucleotide arrays. *Supplement to Nature Genetics*, 21:20–24, 1999.

[10] D. J. Lockhart, H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E. L. Brown. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 14:1675–1680, 1996.

Normalization for Two-color cDNA Microarray Data

*Yee Hwa Yang and Natalie P. Thorne**

Abstract

There are many sources of systematic variation in microarray experiments which affect the measured gene expression levels. Normalization is the term used to describe the process of removing such variation. Two-color cDNA microarray experiments are comparative in nature; therefore, commonly used normalization methods focus on adjusting the value of log-intensity ratios between the red and the green channels. This paper reviews some normalization procedures required to ensure that observed differences across spots both within and between slides are reliably measured. In addition, the paper investigates the possibility of obtaining meaningful single-channel information from two-color microarray experiments after careful single-channel normalization.

Keywords: cDNA microarray; normalization; dye bias; robust smoother; single-channel normalization

1 Introduction

Microarray experiments measure the expression of thousands of genes simultaneously and generate large and complex multivariate datasets. One of the challenges imposed by the enormous growth in this area of biology is the development of computational and statistical tools for processing such datasets. Pre-processing steps such as image analysis and normalization are important aspects of microarray experiments, since they can have a potentially large impact on subsequent data analyses such as clustering or the identification of differentially expressed genes. This paper is concerned with the normalization of two-color cDNA microarray data and examines various procedures applicable to different types of datasets. Normalization is essential to extract reliable measures of the fluorescence intensities and to ensure that the observed differences in intensity indeed reflect differential gene expression and not artefactual bias inherent to the experiment.

We begin in Section 2 with a brief introduction to the biology and technology of cDNA microarrays. This is followed by a discussion in Section 3 on the motivation behind the two main types of normalization procedures: two-channel and single-channel.

*Both authors contributed equally to this work.

Sections 4 and 5 review a number of two-channel and single-channel normalization methods respectively. In particular, Section 5 investigates the possibility of getting useful information from the normalization and analysis of single-channel data from cDNA microarrays. Finally, Section 6 discusses the implications for assessing these different normalization procedures and outlines some of the open questions that remain on this topic.

2 Background on DNA microarrays

DNA microarrays are part of a new class of biotechnologies which allow the monitoring of expression levels for thousands of genes simultaneously. Applications of microarrays range from the study of gene expression in yeast under different environmental stress conditions [6, 10, 13, 22] to the comparison of gene expression profiles for tumors from cancer patients [2, 3, 9, 12, 18, 19]. In addition to the enormous scientific potential of microarrays to help in understanding gene regulation and gene interactions, microarrays are being used increasingly in pharmaceutical and clinical research. Our focus here is on complementary DNA (cDNA) microarrays, where thousands of distinct DNA sequences representing different genes are printed in a high-density array on a glass microscope slide using a robotic arrayer. The relative abundance of each of these genes in two RNA samples may be estimated by fluorescently labeling the two samples, mixing them in equal amounts, and hybridizing the mixture to the sequences on the glass slide. More fully, the two samples of messenger RNA (mRNA) from cells (known as *target*) are reverse-transcribed into cDNA, and labeled using differently fluorescing dyes (usually the red fluorescent dye Cyanine 5 and the green fluorescent dye Cyanine 3). The mixture then reacts with the arrayed cDNA sequences (known as *probes* following the definitions adopted in "*The Chipping Forecast*", a January 1999 supplement to Nature Genetics). This chemical reaction, known as competitive hybridization, results in complementary DNA sequences from the targets and the probes base-pairing with one another. The slides are scanned at wavelengths appropriate for the two dyes, giving fluorescence measurements for each dye for each spot on the array. The underlying assumption in microarray analysis is that these red and green fluorescence intensities for a typical spot represent the amount of mRNA (gene expression) from the corresponding gene in the respective samples. We refer the reader to Schena [21] for a more detailed introduction to the biology and technology of cDNA microarrays.

3 Normalization

Microarray experiments are performed to investigate relationships between different biological samples based on their genes expression. A general approach is to identify genes with relative differential expression between different target samples. The rela-

tive expression from each array is usually measured as the ratio of the red and green fluorescence intensities for each spot. This ratio represents the relative abundance of the corresponding DNA probe in the two mRNA samples. Although these ratios, or fold-changes, provide an intuitive measure of relative expression, they have the disadvantage of treating up- and down- regulated genes differently. Using a log (base 2) scale for intensity is preferred for a number of reasons, including: variation of log-ratios is less dependent on absolute magnitude, and taking the log of the ratio evens out the highly skewed distribution, providing a more realistic sense of variation. For the rest of this review, we base our discussion on log-ratios and log-intensities.

In general, before performing statistical analysis, it is necessary to identify and adjust for artefactual systematic variation in intensities between samples on the same slide and also between slides; that is, variation which cannot be attributed to true biological differences between mRNA samples. This process is known as *normalization.* We define normalization methods based on adjusting the log-ratios as *two-channel* normalization. The need for normalization can be seen most clearly in Figure 1, which shows a plot of a *self–self hybridization.* Here, two identical mRNA samples are labeled with different dyes and hybridized to the same slide. The data are represented by an M versus A plot, or MA plot, where the log-ratios are given by $M = \log_2(R/G)$ and average log-intensity by $A = \log_2 \sqrt{RG}$. Because there is no true differential expression in a self-self hybridization, one would expect the red and green intensities to be equal. However, we observe from Figure 1 that the red intensities tend to be lower than the green intensities. This systematic variation may be a consequence of different labeling efficiencies and scanning properties of the Cy3 and Cy5 dyes; different scanning parameters, such as PMT (photo multiplier tube) settings; print-tip, spatial, or PCR plate effects. Furthermore, the imbalance in the red and green intensities is usually not constant across the spots within and between arrays, and can vary according to overall spot intensity A, location on the array, plate origin, and possibly other variables. Section 4 describes procedures for two-channel normalization.

The advantage of relying on the *log-ratio* for measuring relative gene expression within two samples on the same slide rather than considering *log-intensity* values for individual channels is because log-ratios are considered to be more stable than the absolute intensities across slides. Absolute log-intensities are often confounded by spot-spot variation inherent to printed microarrays. This is demonstrated in Figure 2, where we show the spatial plots of an experiment comparing stages E15 and E18 of the olfactory epithelium (OE) in embryonic mice. Panels (a) and (b) show spatial plots of log-intensities from the red channel (Cyanine 5) and green channel (Cyanine 3) respectively. Panel (c) shows the same spatial plot of log-ratios. We observe reproducible spatial effects of the single channels within a slide that are effectively canceled out by the log ratios. This demonstrates the stability of log-ratios in general compared to log-intensities, and provides a clear warning that analysis of single-channel data should proceed with great care.

The main disadvantage of an analysis based solely on log-ratios is that it constrains

researchers to comparative investigations. At times the nature of the research problem requires *single-channel* analysis, for example, when the aim is to identify genes that are expressed in a certain sample, or perhaps at particular time points in a time series experiment. In this case, the quantity of interest is a separate log-intensity measurement for each channel. Compared to log-ratios, separate log-intensities are usually less stable in cDNA microarrays.

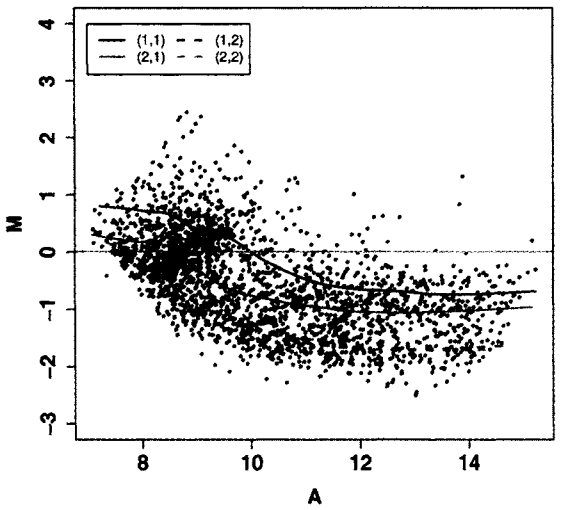


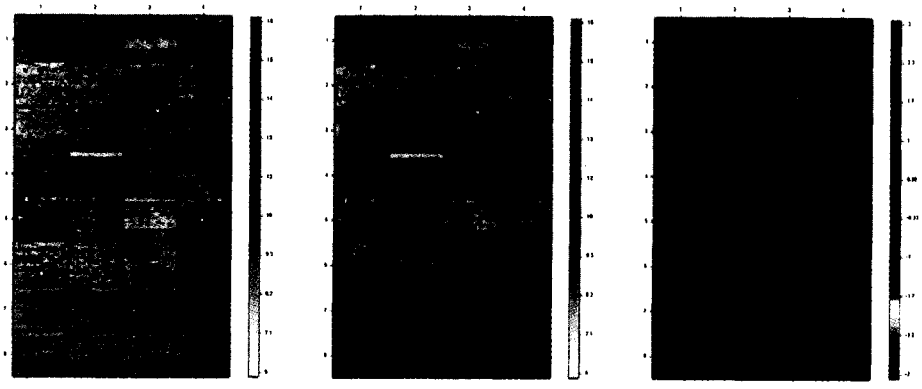
Figure 1: Self-self hybridization illustrating systematic variation. Colored lines indicate the `loess` fit for each of 4 print-tips used to spot the array.

Given the breadth and nature of systematic variation observed in log-ratios, there is an inevitable step-up in complexity of biases for single-channel data. Therefore, the problem of normalization to make the channels from multiple arrays comparable is a more challenging one. Section 5 presents some procedures for single-channel normalization and a discussion on the assessment of single-channel normalization methods.

4 Two-channel normalization

The process of two-channel normalization can be separated into two main components: *location* and *scale*. In general, methods for location and scale normalization adjust the center and spread, respectively, of the distribution of log-ratios. The normalized intensity log-ratios M_{norm} are generally given by

$$M_{norm} = \frac{M - l}{s},$$



(a) (b) (c)

Figure 2: Illustration of the spatial effects that exist in the log-intensities from single-channels that are not observed in the log-ratios. Shown are spatial plots from a single slide in the OE dataset. (a) Spatial plot of red channel. (b) Spatial plot of green channel. (c) Spatial plot of log-ratios.

where l and s denote the location and scale normalization values respectively.

Location normalization

The location value l can be obtained by a wide range of methods. The most commonly used method is *global normalization* with l equal to a constant c and $s = 1$; that is, log-ratios are corrected by subtracting a constant c with $M_{norm} = M - c$. Common choices for this constant c are the median or the mean of the log-intensity ratios (M) for a specified set of genes assumed not to be differentially expressed. There are also many other estimation methods for the constant c. For example, Chen *et al.* [5] propose an iterative method based on ratio statistics for estimating normalization constants. In another approach, Kerr *et al.* [16] and Wolfinger *et al.* [23] propose an ANOVA model for the single channels and perform normalization by including a dye main effect and treatment and array interaction terms in the model. This is followed by adjusting every gene on the array by the same fitted value obtained from model. Figure 3(a) shows an *MA* plot of a mutant *swirl* versus *wild type* comparison of zebrafish prior to normalization. The goal of the *swirl* experiment is to identify genes with altered expression in the mutant compared to *wild type* zebrafish. In this instance, the vast majority of genes on the microarray should show no difference in expression level. This figure depicts a clear dye bias which appears to be dependent on spot intensity. All global methods which subtract the same constant c from every log-ratio on the array do not correct such intensity-dependent biases.

It follows that *location normalization* methods which account for such biases are

often necessary. The intensity-dependent bias is noticeable in an *MA* plot (Figure 3(a)) as a distinct curve in the scatter plot varying with spot intensity. The log-ratios can be normalized by $M_{norm} = M - c(A)$, where $c(A)$ is a function of average spot intensity A. Several intensity-dependent methods have been proposed for location normalization. In Yang *et al.* [24, 25], estimates of $c(A)$ are made using the local scatter plot smoother function `loess` [7, 8] within the software package R. Kepler *et al.* [15] propose a similar approach using a different local regression method. Finkelstein *et al.* [11] present an iterative linear normalization, also known as a robust linear regression, which can be viewed as a constrained version of the robust locally–weighted intensity-dependent normalization.

Figure 3(b) shows boxplots of log-ratios stratified by print-tip groups after intensity-dependent normalization. This figure shows that after intensity-dependent normalization, other systematic biases still remain. We can generalize further to account for other bias by fitting different intensity-dependent curves to different regions of the array: $M_{norm} = M - c_i(A)$, where i indexes different regions of the array. For example, Yang *et al.* [24] to use i to index print-tip groups. Often, systematic differences result from such differences between the print-tips as slight variations in length or in the size of the tip opening, or variable tip deformation after many hours of printing. In addition, because each tip prints DNA spots on different areas of the slide, print-tip groups are proxies for spatial effects on the slide. Figure 3(c) shows an *MA* plot after print-tip group `loess` normalization.

Scale normalization: within and between slides

The effect of location normalization is to center log-ratios around zero by accounting for intensity- and spatially-dependent bias. In addition, it is important to consider *scale normalization*, since large scale differences between multiple slides can lead some slides giving undue weight to an average of log-ratios across slides. One common method of scale normalization is to divide each intensity by the total of the intensities on the slide, so that all slides then have the same total intensity. Yang *et al.* [24] instead propose a robust estimate of scale, such as the *median absolute deviation (MAD)*, for both within-slide and multiple-slide (across slide) scale adjustment. Yang *et al.* [24] also discuss that the need for scale normalization is often determined empirically, as there is a trade-off between the gains achieved by scale normalization and the possible increase in variability introduced by this additional step. In cases where scale differences appear fairly small, it may thus be preferable to perform only a location normalization.

Comparing different methods

We can compare different within-slide normalization methods by examining their effects on the location and scale of the normalized log-ratios M_{norm}. Figure 3(d) shows density plots of the log-ratios for different normalization methods. Without normaliza-

tion (black curve), the log-ratios are centered around -0.5 indicating a bias toward the green (Cy3) dye. A global median normalization (red curve) shifts the center of the log-ratio distribution to zero but does not affect the spread. The dependence of the log-ratio M on the overall intensity A is also still present. Both the intensity-dependent (green curve) and within print-tip group (blue curve) location normalization methods reduce the spread of the log-ratios compared to a global normalization. It is important to note that these approaches implicitly assume that relatively few genes are differentially expressed, or there is no systematic relationship between differential gene expression and intensity or location of the spots on the slide.

Control genes

For most of the methods described, the set of genes to use for the normalization must be decided. In general, the set of genes most appropriate for normalization depends on the nature of the experiment, the amount of observed variation in gene expression, and possibly also on the normalization method applied to the data. Frequently, biological comparisons made on microarrays are of a very specific nature, and differences in gene expression are only detected in a small proportion of genes. In these experiments, it is usual to use most of the genes on the array. Instead of using all genes for normalization, one may use a selected subset of constantly expressed genes. These include the traditional "housekeeping genes", spiked controls, genomic DNA, Microarray Sample Pooled (MSP) titration series [24] and rank-invariant genes. Further details on the effect of different sets of control genes on normalization procedures are provided in [24].

5 Single-channel normalization

Single-channel normalization aims to remove systematic intensity bias, that is, intensity not due to real gene expression, from the red (Cy5) and green (Cy3) channels separately, both within and between arrays. This normalization allows comparisons of absolute intensities between arrays.

Jin *et al.* [14] performed a factorial experiment on age, sex and genotype (two levels for each factor) of *Drosophila melanogaster* flies, where age was the only factor compared within slides. The main effects for the remaining factors were estimable only via single-channel analysis, not by analysis of the log-ratios. Notably, a different experimental design would have enabled all main effects and interactions to be estimated from log-ratios while still maintaining a reasonable level of replication for each comparison type. Here we draw attention to the fact that complex multi-factor designs may not facilitate the estimation of all contrasts of interest from log-ratios alone. In such cases, it may be desirable to recover information from single-channel analysis. Indeed, future complex microarray experiments may be specifically designed to incorporate both log-ratio and log-intensity single-channel analysis methods. In time series experiments, absolute intensity estimates at each time could tell us which genes are

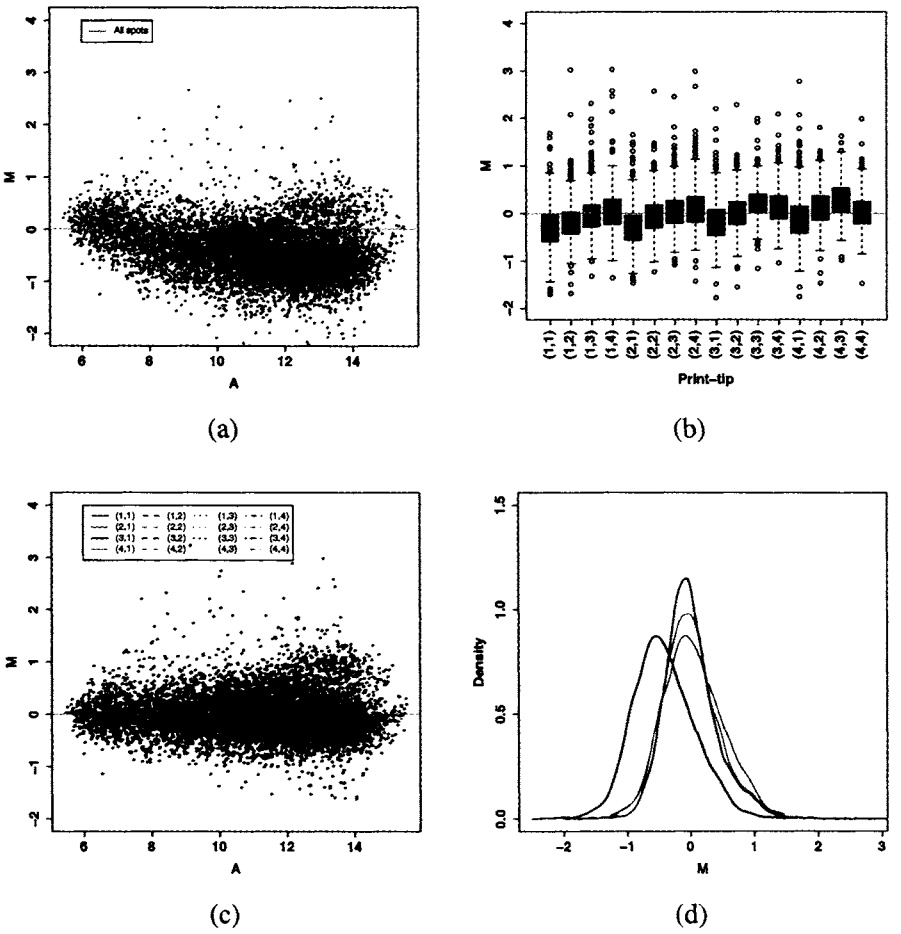


Figure 3: Illustration of two-channel normalization using the *swirl* dataset. (a) *MA* plot before normalization; the green curve corresponds to the `loess` fit for the entire dataset. (b) Boxplots, stratified by print-tip group, of log-ratios after intensity-dependent (`loess`) normalization, but before within print-tip group normalization. (c) *MA* plot after within print-tip group normalization. (d) Density plots of the log-ratios for different normalization procedures. The solid black curve represents the density of the log-ratios without normalization. The red, green, and blue curves represent the densities after global median normalization, intensity-dependent location normalization, and within print-tip group location normalization, respectively.

expressed or not at any given time, or allow estimation of between array single-channel comparisons of time points.

Analysis methods that use ANOVA to model the log-intensities rather than the log-ratios have been investigated by Kerr *et al.* [16] and Wolfinger *et al.* [23]. As mentioned in Section 4, these ANOVA models essentially perform constant global normalization and are therefore inadequate for correcting the nonlinear and spatial systematic variation observed, *e.g.* in Figure 2. Analysis methods that model single-channel intensities have been proposed for "one-color" technologies such as nylon filter arrays and Affymetrix GeneChip. Unlike the cDNA arrays, these technologies generate only a single channel of absolute expression data from each array. Various methods have been proposed [1, 4, 17, 20] to normalize multiple Affymetrix arrays. In this section, we look at extending some of these methods for single-channel normalization of cDNA arrays.

We illustrate the problem of single-channel normalization with a time series dataset examining the olfactory epithelium (OE) of embryonic mice with all possible pairwise dye-swap comparisons of stages E13, E14, till E18. In this paper, we do not explicitly investigate the biological problem of which genes are expressed over time, but rather use the dataset for illustrative purposes only. In addition to the balanced, highly replicated design of this experiment, this dataset is appealing because it contains many controls of different known concentrations. Every print-tip group on every slide includes two different Microarray Sample Pooled (MSP) titration controls of 5 and 6 concentrations respectively [24]. We later outline possible uses for this in assessing single-channel normalization methods.

Single-channel normalization of two-color cDNA microarray experiments can be considered as a two stage process: *within-array* normalization followed by *between-array* (between all channels from multiple arrays) normalization.

In addressing the within-array single-channel normalization problem we see that many parallels can be drawn from the two-channel location normalization approach, such as removing systematic imbalances between the $\log R$ and $\log G$ intensities and correcting for spatial effects within slides. For *dye bias* correction, we can adjust the log-red and log-green intensity by $\log R_p = \log R - \frac{1}{2}c_i(A)$ and $\log G_p = \log G + \frac{1}{2}c_i(A)$ where $c_i(A)$ denotes the normalization adjustment estimated from "print-tip loess" normalization within each slide. We notice that in addition to normalizing spatial effects based on the log-ratios, we must also address spatial effects of the absolute intensity of both channels. This is evident in Figure 2, where we see that even though there is no observable systematic spatial variation in the log-ratios we can still observe reproducible spatial effects of the single-channels. We refer to such arrays as having *systematic spatial variation in intensity within slides*. Efforts are underway to investigate spatial normalization methods which will be robust to extreme local intensity values.

The second stage of single-channel normalization, between-array single-channel normalization, is concerned with comparability of the distributions of log-intensities between arrays. Like the two-channel problem, we wish for the single-channels to have similar scale and location values. At this stage, we do not distinguish which channel is

red and which green, and assume that red-green imbalances were removed by within-array normalization.

For the OE dataset in Figure 4 we see that the distributions of all 60 channels from the 30 arrays are quite varied. The density curves differ in location, variation and shape. Interestingly, the red and green channels within arrays are very close in distribution (data not shown). To adjust for the difference in distribution between channels from multiple arrays, we consider methods developed for *Affymetrix* technology. In particular, we adapt the *quantile normalization* method proposed in Bolstad *et al.* [4]. This method extends the idea of normalizing for equivalent medians or quartiles of the single-channels by requiring *every quantile* across channels be equivalent, and thus forcing each channel to share a common distribution. The distribution is estimated by averaging across channels for each quantile. We refer the reader to Bolstad *et al.* [4] for further details on this method and an algorithm for its implementation. Of particular concern with the use of this method is that replacing quantile values with an average might attenuate log-intensity values, particularly in the tails of the distribution where real expression is potentially affected.

In assessing the performance of these methods, we recommend constructing *MA* plots based on normalized log-intensities to check that dye-biases have been removed. Figure 5 displays *MA* plots for a typical array from the OE dataset showing the effect of different single channel normalization methods. Panel (a) shows the data before any normalization. Between-array quantile normalization (Panel (c)), based on the entire OE dataset, appears to be just as effective at removing intensity dependent dye-bias as the within-array "print-tip loess" single-channel normalization shown in panel (b). We advise using boxplots of the red and green channels to assess red-green imbalances and to check the location and scale of log-intensity distributions after different levels of normalization. It is beneficial to highlight any previously known differentially expressed genes on the *MA* plots to check that they remain distinguishable after normalization.

In the OE dataset, the intensity values of MSP titration controls should remain constant across all 60 channels regardless of what is hybridized. Thus, we can easily determine whether normalization decreases the variability of these control measurements in the single-channels. However, determining bias before and after normalization is more challenging. To measure bias we must be able to compare observed intensities with something known; that is, some truth must be available. The truth regarding absolute intensities for the MSP titration controls is unknown, but there is some knowledge about their relative absolute intensities based on the concentrations of the titration series. We can check (data not shown) that the ratios of intensities between different controls get closer to what we expect. Currently in progress is a variance-bias assessment of the performance of the normalization methods on OE and other similar datasets.

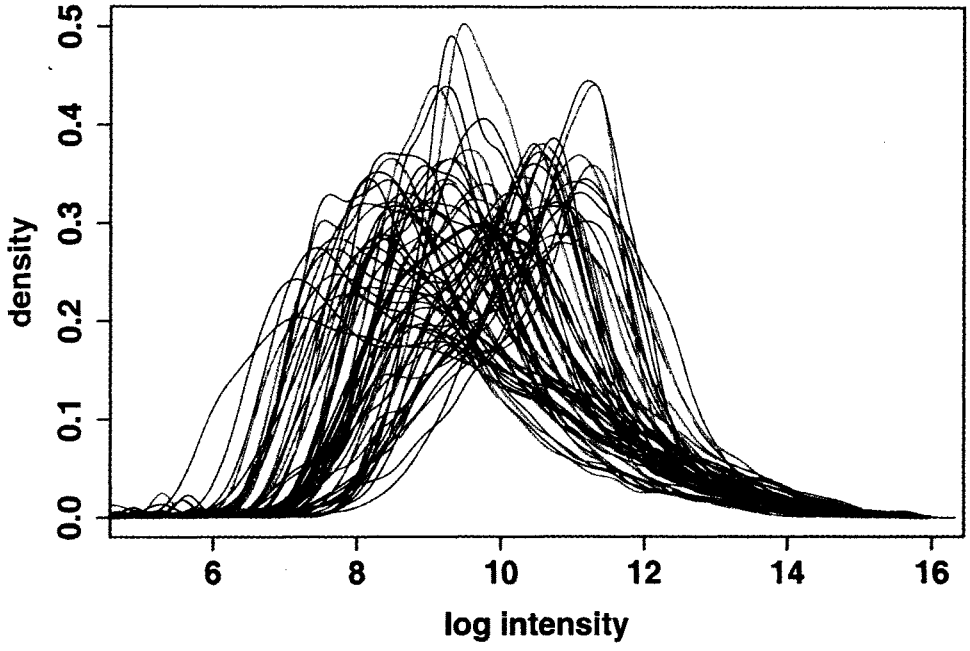


Figure 4: *Single-Channel Quantile Normalization.* Density plots of each of 30 red and green log-transformed single-channels from the OE dataset. The densities of red and green channels within slides are usually very similar. The solid black curve represents the density of all channels after quantile normalization.

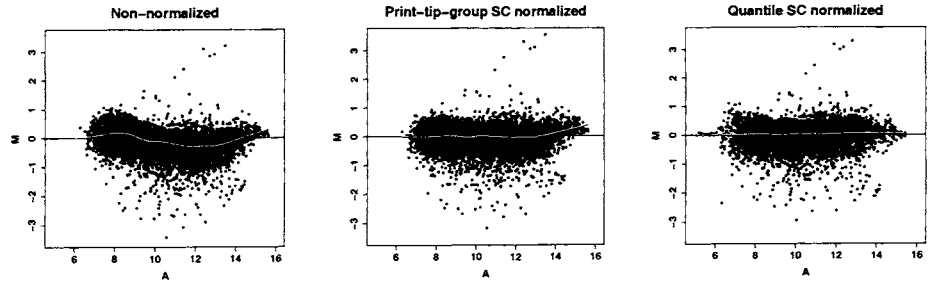


Figure 5: *MA* plots with `loess` curve for a typical array from the OE data (a) before normalization, (b) after single-channel quantile normalization and (c) after "print-tip group" single-channel normalization.

6 Discussion

We have reviewed various normalization approaches applicable to different types of microarray experiments. Most of the normalization work to date is based on two-channel normalization procedures that adjust the log-ratios. However, we have also considered the problem of normalizing single-channels from two-color cDNA microarrays. We have also raised the question of how to assess single-channel normalization is preferable, and what aspects to consider when comparing normalization methods. We neither advocate nor promote the notion of single-channel data analysis in general, but instead suggest that satisfactory normalization of single-channel data is what is lacking for it to be considered a promising option for researchers.

We demonstrate that single-channel analysis is potentially useful in certain circumstances where the nature of the research problem suggests single-channel analysis. Though analyzing microarray data based solely on single-channels is not a new concept [4, 20], limited attention has been given to single-channel normalization of two-color cDNA microarrays. As a place to begin, we have adapted existing procedures from both two-color cDNA and from single-color (*e.g.* Affymetrix) normalizations. The investigation into single-channel normalization raises many other issues of interest, including, in particular, the implications for normalization of log-ratios, for experimental design and analysis and for the replication required for reasonable precision of between array single-channel contrasts.

In any microarray experiment it is important to adjust for the inherent artefactual bias, as well as to understand the assumptions behind any procedure used. In addition, it should be checked that systematic errors are reduced after normalization and that any observed gene expression differences are meaningful (scientific validation). Diagnostic plots such as *MA* plots, spatial plots, density and boxplots can assist in the decision of the level of adjustment needed for both single- and two-channel normalization, and can be used to check that artefacts have been removed by normalization. For example, investigators may decide whether to perform within-slide scale normalization for a dataset by examining boxplots of log-ratios stratified by different print-tip groups.

In general, one should be careful that the gains achieved by further levels of normalization do not introduce a large increase in variability. An important problem that should be addressed is to define formal criteria to assess the effectiveness of various normalization procedures. That is, the issues of bias and variance should be addressed simultaneously. In practice, it is relatively easy to show whether a new normalization method decreases variance. However, it is more challenging to establish that this reduction in variance did not come at the cost of attenuating absolute and relative intensity values (increased bias). To fully address this issue, it is important to obtain a specially constructed dataset with known levels of absolute and differential gene expression, as well as a reasonable number of replications. Examples of such datasets are available for Affymetrix technology `http://qolotus02.genelogic.com/datasets.nsf/` and some initial analyses of these data are available

at `http://www.stat.Berkeley.EDU/users/terry/zarray/Affy/affy_index.html`. In conclusion, until such datasets are available for two-color cDNA microarrays, or until further understanding of the effects of different normalization procedures is gained, it is important to apply normalization algorithms with caution.

Acknowledgments

The authors wish to thank our advisor Terry Speed for all his patience, guidance and encouragement in all aspects of our studies. His wealth of ideas and scientific enthusiasm have made our research both interesting and exciting. We appreciate the discussions, ideas and support given by Gordon Smyth on many aspects of our research, especially on single-channel normalization. Sandrine Dudoit is much appreciated for many discussions on normalization topics, as is Matthew Ritchie. We would like to thank David Kimelman and David Raible at the University of Washington for providing the *swirl* embryos and Katrin Wuennenberg-Stapleton from the Ngai Lab at UC Berkeley for performing the *swirl* microarray experiment. We would like to thank Cynthia Duggan and Carina Howell from the Ngai Lab at UC Berkeley for providing us with the dataset from the OE experiment. Many thanks to our colleagues at the Walter and Eliza Hall Institute who provide a breadth of general and technical support. Finally, we thank the anonymous referees and Darlene Goldstein for comments on an earlier version of this paper.

(Jean) Yee Hwa Yang, Division of Biostatistics, University of California, San Francisco, `jean@biostat.berkeley.edu`

Natalie P. Thorne, Division of Genetics and Bioinformatics, The Walter and Eliza Hall Institute, Australia, `thorne@wehi.edu.au`

References

[1] Affymetrix. *Statistical alogrithms reference guide*, 2001. Technical report.

[2] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson Jr, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt. Different types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–511, 2000.

[3] U. Alon, N. Barkai D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor

and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences, USA*, 96:6745–6750, 1999.

[4] B. M. Bolstad, R. A. Irizzary, M. Åstrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*. In press.

[5] Y. Chen, E. R. Dougherty, and M. L. Bittner. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Journal of Biomedical Optics*, 2:364–374, 1997.

[6] S. Chu, J. DeRisi, M. Eisen, J. Mulholland, D. Botstein, P. O. Brown, and I. Herskowitz. The transcriptional program of sporulation in budding yeast. *Science*, 282:699–705, 1998.

[7] W. S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74:829–836, 1979.

[8] W. S. Cleveland and S. J. Devlin. Locally-weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83:590–610, 1988.

[9] J. DeRisi, L. Penland, P. O. Brown, M. L. Bittner, P. S. Meltzer, M. Ray, Y. Chen, Y. A. Su, and J. M. Trent. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genetics*, 14:457–460, 1996.

[10] J. L. DeRisi, V. R. Iyer, and P. O. Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278:680–685, 1997.

[11] D. B. Finkelstein, J. Gollub, R. Ewing, F. Sterky, S. Somerville, and J. M. Cherry. Iterative linear regression by sector: renormalization of cDNA microarray data and cluster analysis weighted by cross homology. In *CAMDA 2000*, 2000.

[12] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M.L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

[13] T. R. Hughes, M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. A. Bennett, E. Coffey, H. Dai, Y. D. He, M. J. Kidd, A. M. King, M. R. Meyer, D. Slade, P. Y. Lum, S. B. Stepaniants, D. D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, and S. H. Friend. Functional discovery via a compendium of expression profiles. *Cell*, 102:109–126, 2000.

[14] W. Jin, R. M. Riley, R. D. Wolfinger, K. P. White, G. Passador-Gurgel, and G. Gibson. The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. *Nature Genetics*, 29:389–395, 2001.

[15] T. B. Kepler, L. Crosby, and K. T. Morgan. Normalization and analysis of DNA microarray data by self-consistency and local regression. *Genome Biology*, 3(7):research0037.1–12, 2002.

[16] M. K. Kerr, M. Martin, and G. A. Churchill. Analysis of variance for gene expression microarray data. *Journal of Computational Biology*, 7:819–837, 2000.

[17] C. Li and W. H. Wong. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of the National Academy of Sciences, USA*, 98:31–36, January 2001.

[18] C. M. Perou, S. S. Jeffrey, M. van de Rijn, C. A. Rees, M. B. Eisen, D. T. Ross, A. Pergamenschikov, C. F. Williams, S. X. Zhu, J. C. F. Lee, D. Lashkari, D. Shalon, P. O. Brown, and D. Botstein. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proceedings of the National Academy of Sciences, USA*, 96:9212–9217, 1999.

[19] D. T. Ross, U. Scherf, M. B. Eisen, C. M. Perou, P. Spellman, V. Iyer, S. S. Jeffrey, M. Van de Rijn, M. Waltham, A. Pergamenschikov, J. C. F. Lee, D. Lashkari, D. Shalon, T. G. Myers, J. N. Weinstein, D. Botstein, and P. O. Brown. Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics*, 24:227–234, 2000.

[20] E. E. Schadt, C. Li, B. Eliss, and W. H. Wong. Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *Journal of Cellular Biochemistry*, 84:120–125, 2002.

[21] M. Schena, editor. *Microarray Biochip Technology*. Eaton, 2000.

[22] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell*, 9:3273–3297, 1998.

[23] R. D. Wolfinger, G. Gibson, E. D. Wolfinger, L. Bennett, H. Hamadeh, P. Bushel, C. Afshari, and R. S. Paules. Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology*, 8(6):625–638, 2001.

[24] Y. H. Yang, S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, and T. P. Speed. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, 30(4):e15, 2002.

[25] Y. H. Yang, S. Dudoit, P. Luu, and T. P. Speed. Normalization for cDNA microarray data. In M. L. Bittner, Y. Chen, A. N. Dorsel, and E. R. Dougherty, editors,

Microarrays: Optical Technologies and Informatics, volume 4266 of *Proceedings of SPIE*, May 2001.

Classification of Tissue Samples Using Mixture Modeling of Microarray Gene Expression Data

Shili Lin and Roxana Alexandridis

Abstract

Accurate classification of tissue samples is an essential tool in disease diagnosis and treatment. The DNA microarray technology enables disease classification based only on gene expression analysis, without prior biological insights. We present a classification method based on modeling the distribution of the gene expression profile of a test sample as a mixture of distributions, each of which characterizes the levels of gene expression within a class. Class assignment for a test sample is based on the predictive probabilities of class memberships. We believe that this general modeling framework is a flexible scheme for multi-type classification. Since most of the thousands of genes whose expression levels are measured do not contribute to the separation between types of tissue samples, we also explore several measures for gene selection, including T, NPT, BW, NPBW, and a mixture modeling approach based on Markov chain Monte Carlo (MCMC) estimation of parameters. For a classifier based on a gene selection measure, such as the T classifier, the number of genes selected is achieved by cross-validation. The methods are applied to a leukemia dataset; our results are comparable with the best results achieved in a comparative study done by Professor Terry Speed and colleagues.

Keywords: microarray; gene expression; classifier; mixture; EM; MCMC

1 Introduction

DNA microarrays are biotech chips that enable researchers to measure the expression levels of thousands of genes simultaneously; see Schena [15] and The Chipping Forecast [5]. These measurements are obtained by quantifying the hybridization of the mRNA extracted from tissue samples to an array of spotted cDNA (cDNA arrays) or oligonucleotide probes (oligonucleotide arrays) immobilized on the surface of the chip. Details can be found in Schena *et al.* [16] for cDNA arrays and Lockhart *et al.* [9] for oligonucleotide arrays.

After proper image analysis, data processing and normalization (which entails nontrivial efforts, see for example, Dudoit *et al.* [4], Schadt *et al.* [14], Newton *et al.* [11], and Yang *et al.* [17]), a single number, referred to as the level of expression, is obtained for each gene on a microarray.

Statistical methods are needed to address many of the questions for which researchers seek answers from microarray gene expression data, such as (1) identifying genes differentially expressed under two or more conditions, (2) grouping genes with similar expression patterns, (3) finding genes that differentiate one tissue from another, and (4) molecular classification of tissue samples, including class discovery and class prediction. We focus on statistical methods for addressing this last issue.

Accurate classification of tissue samples is an essential tool in disease diagnosis and treatment. DNA microarray technologies enable classification based only on gene expression analysis, without requiring prior biological insight; successful cancer classification by Golub *et al.* [6] provides an excellent example. The idea is to classify a tissue sample into one of K known classes/types, where a sample, also called gene expression profile, is a vector whose components are the levels of gene expressions in a given tissue. Therefore, the problem of classification can be defined as follows: given a set of training samples, *i.e.*, samples whose class memberships are known, and a set of test samples, predict the class assignments of the test samples.

Most of the thousands of genes that make up the gene expression profile of a tissue sample do not contribute to the distinction between classes. Considering such irrelevant genes introduces noise to the classification process, and increases computational hurdles due to the extremely large dimensionality of the data. The combined contribution of many nonsignificant genes could downplay or even cancel the effects of the significant ones [8]. In addition, with a large number of genes whose expression levels are used for classification purposes, the interpretability of the results becomes an issue. When only a few genes are found helpful for separating classes, insight might be gained into the biological significance of these genes, as shown in Golub *et al.* [6].

For binary classification problems, Ben-Dor *et al.* [2] suggest a gene selection algorithm with a single threshold value chosen by cross-validation. Golub *et al.* [6] select the genes that provide best distinction between the "standardized" means of two classes (although their standardization is not the typical kind of standardization in statistics). Dudoit *et al.* [3] propose to select genes that display the largest ratios of between-group to within-group sums of squares, which is applicable to gene selection for multi-type classification problems.

Numerous methods have been proposed to classify tissue samples based on gene expression data. Some are restricted to binary classification, such as the weighted voting scheme of Golub *et al.* [6], while others are applicable to multi-type classifications. Techniques of machine learning, such as nearest neighbor classifiers [3], and cluster analysis methods, including hierarchical clustering [1, 12], have been entertained. Classification trees or aggregation of classifiers built from perturbed versions of the training set using boosting, bagging or convex pseudo-data methods of perturbing the training set [3], are some other examples.

There are yet other classification techniques that are applicable to multi-type classification problems; these are based on modeling the class densities, such as the linear and quadratic discriminant analysis of Dudoit *et al.* [3], or the naive Bayes methods of

Keller *et al.* [8]. For a comprehensive review of the methods, see Keller *et al.* [8] and Dudoit *et al.* [3].

In this article, we propose a classification method based on modeling the gene expression profile of a test sample as arising from a mixture of distributions, each of which characterizes the expression profiles within a class. We believe that this general modeling framework is a flexible scheme for multi-type classification. It could also be extended to accommodate class discovery in addition to classification to known classes. We also explore several measures for gene selection, including a mixture modeling approach based on Markov chain Monte Carlo (MCMC) estimation of parameters.

2 A Multi-type Classification Method

Mixture modeling of test samples

Let K denote the number of known classes (or sub-types, *e.g.*, leukemia sub-types) for which training samples exist. We use $Y_{ki} = (Y_{ki1}, \cdots, Y_{kiG})'$ to denote the column vector of gene expressions of the ith sample from class k, where G is the number of genes. Hence $\{Y_{ki}, i = 1, \cdots, T_k\}$ is the collection of data from class k, where T_k is the sample size, $k = 1, \cdots, K$. For each $i = 1, \cdots, T_k$, we assume $Y_{ki} \sim f_k(. \mid \theta_k)$, where θ_k is the vector of parameters of the component density function, which can be estimated, for example, from the training samples.

Let $\{X_i = (X_{i1}, \cdots, X_{iG})', i = 1, \cdots, T\}$ denote the gene expression data from T test samples, whose class membership assignments are unknown and the subject of interest. We model X_i as i.i.d. observations from a mixture distribution with component density functions f_k but unknown component weights π_k, $k = 1, \cdots, K$, $\sum_{k=1}^{K} \pi_k = 1$. That is,

$$f(X_i \mid \theta) = \sum_{k=1}^{K} \pi_k f_k(X_i \mid \theta_k), i = 1, \cdots, T,$$

where θ is the vector of unknown parameters including the π_k.

Two likelihood formulations are considered. If we assume that the parameters of each component density are to be estimated from the corresponding training samples, then the likelihood formulation is based on known component densities. The parameter vector is thus $\theta = \{\pi_k, k = 1, \cdots, K\}$, with the constraint $\sum_{k=1}^{K} \pi_k = 1$, and will be estimated using the test samples only. The likelihood function is

$$L_1(\theta) = \prod_{i=1}^{T} \left[\sum_{k=1}^{K} \pi_k f_k(X_i \mid \theta_k) \right]. \tag{1}$$

Alternatively, we can estimate the parameters θ_k in each component density together with the component weight π_k using data from training samples and test samples jointly. The parameter vector is thus $\theta = \{\pi_k, \theta_k, k = 1, \cdots, K\}$, again with the constraint

$\sum_{k=1}^{K} \pi_k = 1$. The likelihood function is then

$$L_2(\theta) = \prod_{k=1}^{K} \prod_{i=1}^{T_k} f_k(Y_{ki} \mid \theta_k) \prod_{i=1}^{T} \left[\sum_{k=1}^{K} \pi_k f_k(X_i \mid \theta_k) \right]. \tag{2}$$

Under the latter formulation, data from the test samples also contribute to the estimation of the parameters in each component density. In the next section, we focus on estimation of the parameters under this formulation. Parameter estimations under formulation (1) can be carried out similarly.

EM estimation of parameters

We assume that each component density is multivariate normal with mean vector $\mu_k = (\mu_{k1}, \cdots, \mu_{kG})$ and variance-covariance matrix Σ_k, that is, $\theta_k = \{\mu_k, \Sigma_k\}$. We further assume that the expression levels among different genes are independent, therefore $\Sigma_k = \text{diag}(\sigma_{k1}^2, \cdots, \sigma_{kG}^2)$ is a diagonal matrix of the variances. To find the maximum likelihood estimates (MLEs) of the parameters in (2) with normal component densities, the EM algorithm is highly suited [10].

Let Z_i, which takes a value from the set $\{1, 2, \cdots, K\}$, denote the unobserved class assignment for test sample $i = 1, \cdots, T$. Then $\{(X_i, Z_i), i = 1, \cdots, T\} \cup \{(Y_{ki}, k), k = 1, \cdots, K, i = 1, \cdots, T_k\}$ can be regarded as a representation of the complete data. The corresponding complete data likelihood is

$$L_c(\theta) = \prod_{k=1}^{K} \prod_{i=1}^{T_k} f_k(Y_{ki} \mid \theta_k) \prod_{i=1}^{T} \prod_{k=1}^{K} [\pi_k f_k(X_i \mid \theta_k)]^{I(Z_i = k)},$$

where $I(Z_i = k)$ is the indicator function that takes the value 1 if $Z_i = k$ and 0 otherwise. The EM iterates for the parameters are easily obtained and are given by:

$$\pi_k^{(t+1)} = \sum_{i=1}^{T} E[I(Z_i = k) \mid X_i, \theta^{(t)}]/T, \tag{3}$$

$$\mu_k^{(t+1)} = \frac{\sum_{i=1}^{T_k} Y_{ki} + \sum_{i=1}^{T} E[I(Z_i = k) \mid X_i, \theta^{(t)}] X_i}{T_k + \sum_{i=1}^{T} E[I(Z_i = k) \mid X_i, \theta^{(t)}]}, \tag{4}$$

$$(\sigma^2)_{kg}^{(t+1)} = \frac{\sum_{i=1}^{T_k} (Y_{kig} - \mu_{kg}^{(t)})^2 + \sum_{i=1}^{T} E[I(Z_i = k) \mid X_i, \theta^{(t)}](X_{ig} - \mu_{kg}^{(t)})^2}{T_k + \sum_{i=1}^{T} E[I(Z_i = k) \mid X_i, \theta^{(t)}]}, \tag{5}$$

$$k = 1, \cdots, K, g = 1, \cdots, G,$$

where

$$E[I(Z_i = k) \mid X_i, \theta^{(t)}] = \frac{\pi_k^{(t)} f_k(X_i \mid \theta_k^{(t)})}{\sum_{k=1}^{K} \pi_k^{(t)} f_k(X_i \mid \theta_k^{(t)})}.$$

From a starting parameter configuration $\theta^{(0)}$, we compute the sequence of estimates $\theta^{(t)}$ iteratively using equations (3)-(5) until convergence. The resulting parameter configuration is the estimated MLEs and is denoted by $\hat{\theta}$. Note that the estimates of the component density parameters involve data from the test samples as well as those from the training samples, as pointed out earlier.

A Classification Scheme

For each sample to be classified, we compute the predictive probabilities that it belongs to each of the known classes given the observed expression data and the parameter estimates. Then the sample is assigned to the class that has the largest predictive probability. That is, we compute

$$P(Z_i = k \mid X_i, \hat{\theta}) \propto \hat{\pi}_k f_k(X_i \mid \hat{\theta}_k), k = 1, \cdots, K. \tag{6}$$

Then

$$\hat{Z}_i = \text{argmax}_k \{P(Z_i = k \mid X_i, \hat{\theta}), k = 1, \cdots, K\}.$$

For a test set with T samples with known underlying class assignments $z_i, i = 1, \cdots, T$, the quantities $r = \sum_{i=1}^{T} I(\hat{Z}_i = z_i)/T$ and $e = \sum_{i=1}^{T} I(\hat{Z}_i \neq z_i)$ give the prediction accuracy rate and the number of samples that are misclassified, respectively.

3 Methods for Building Classifier

Gene selection measures and cross-validation

Gene selection measures are summary statistics used to order or select genes according to the perceived importance in discriminating among known classes. Four measures are described below and their performances are evaluated. Other gene selection measures are also considered; see the Discussion section for details.

T: This measure is applicable to two-class discriminant problems only. The measure T_g is simply the two sample t-statistic for each gene $g = 1, \cdots, G$. That is,

$$T_g = \frac{\bar{Y}_{1.g} - \bar{Y}_{2.g}}{\sqrt{S_{1.g}^2/T_1 + S_{2.g}^2/T_2}},$$

where $\bar{Y}_{1.g} = \sum_{i=1}^{T_1} Y_{1ig}/T_1$ and $S_{1.g}^2 = \sum_{i=1}^{T_1} (Y_{1ig} - \bar{Y}_{1.g})^2/(T_1 - 1)$ are the sample mean and sample variance of class 1, respectively, and similarly for $\bar{Y}_{2.g}$ and $S_{2.g}^2$. Then the N genes with the largest absolute T_g values are selected to form the T classifier of size N. We discuss how to select N through cross-validation below.

NPT: This is the non-parametric counterpart of T, and thus is also applicable only to two-class problems. Let $R_g =$ rank $\{Y_{1ig}, i = 1, \cdots, T_1, Y_{2ig}, i = 1, \cdots, T_2\}$ denote the vector of the ranks of all the samples, among both classes, of gene g. The *NPT* measure is defined as the difference of the average rank of the samples in class one ($\bar{R}_{1.g}$) and that in class two ($\bar{R}_{2.g}$), that is, $NPT_g = \bar{R}_{1.g} - \bar{R}_{2.g}, g = 1, \cdots, G$. The N genes with the

largest absolute NPT_g values are selected to form the NPT classifier of size N. This classifier is more robust than T to outlying expression levels.

BW: This is a classifier based on the ratio of between-class sum of squares to within-class sum of squares, as proposed by Dudoit *et al.* [3]. This classifier is applicable to multi-type classification problems. The BW classifier is equivalent to the T classifier for two-class problems when the sample sizes in the two classes are equal, and hence, it may be viewed as a generalization of the T classifier. Specifically, define

$$BW_g = \frac{\sum_{k=1}^{K} T_k(\bar{Y}_{k.g} - \bar{Y}_{..g})^2}{\sum_{k=1}^{K}\sum_{i=1}^{T_k}(Y_{kig} - \bar{Y}_{k.g})^2},$$

where $\bar{Y}_{k.g}$ is the sample mean of class k, and $\bar{Y}_{..g}$ is the overall mean of all samples across all classes. Then the N genes with the largest BW_g values are selected to form the BW classifier of size N.

NPBW: This is the non-parametric counterpart of the BW classifier, which is also applicable to multi-type problems. The gene selection measure $NPBW_g$ is similarly defined as in BW_g but with the individual expression levels or the means replaced by their corresponding ranks (across all samples) and the corresponding average ranks. Like NPT, this classifier is robust to outlying expression levels.

For each type of classifier, after the genes are ordered according to their relative importance in discriminating among known classes, the number of genes N to use for classifying new samples must be selected. This task is accomplished by Leave-One-Out Cross-Validation (LOOCV). For each competing classifier, we estimate the parameters of the mixture model using data from the training samples, but leaving one out as a test sample. Since the true class assignment of the test sample is known, we can score whether correct assignment is made. After cycling through all the training samples one at a time, the prediction accuracy rate may be computed. A classifier with high prediction accuracy rate from LOOCV will be used as a candidate for classification of new samples.

An MCMC classifier

An alternative approach to gene selection for a two-class classifier is through mixture modeling and MCMC estimation and model selection. Suppose $(\bar{Y}_{k.g}, S^2_{k.g})$ are the sample mean and variance of gene g in class $k = 1,2; g = 1,\cdots,G$. If the sample size T_k is reasonably large, then $\bar{Y}_{k.g} \sim N(\mu_{kg}, S^2_{k.g}/T_k)$ approximately. Hence,

$$Y_g = \frac{\bar{Y}_{1.g} - \bar{Y}_{2.g}}{\sqrt{S^2_{1.g}/T_1 + S^2_{2.g}/T_2}}$$

follows a normal distribution with mean $\mu_{1g} - \mu_{2g}$, approximately. For a gene that is not differentially expressed in the two classes, $\mu_{1g} - \mu_{2g} = 0$. Thus, one may model Y_g as from a mixture of (univariate) normal distributions $(N(\mu_\lambda, \sigma^2_\lambda), \lambda = 1,\cdots,\Lambda)$ with an unknown number ($\Lambda \geq 1$) of components, with one of the components having mean zero

(referred to as the null component), representing those genes that are not differentially expressed.

The MCMC reversible jump method of Green [7] and Richardson and Green [13] is used to estimate the parameters of the model, including the number of components of the mixture. Then for each gene g, we compute the predictive probability that it belongs to each component of the mixture given Y_g and the estimated model, using a formula similar to (6). The gene is assigned to a component other than the null component if the predictive probability for that component is the largest and also larger than the weight of the null component. The collection of genes assigned to components other than the null component forms the MCMC classifier.

Following Richardson and Green [13], weak informative priors, chosen for computational convenience, are used for the model parameters. The priors for the means and variances of the component densities are assumed to be independent normals ($\mu_\lambda \sim N(\xi, \kappa^2)$) and inverse gammas ($\sigma_\lambda^2 \sim IG(\alpha, \beta)$), respectively. The hyperparameters ξ and κ are chosen to be the midpoint and half of the range (R) of the data interval, respectively, to make the prior for μ_λ to be rather flat. For σ_λ^2, we let $\alpha = 2$ and allow β to further follow a gamma distribution $G(l, h)$ with $l = 0.2$, and $h = 10/R^2$, to make σ_λ^2 similar but without being informative in their absolute size. The prior for the number of components (Λ) is assumed to be uniform between 1 and the pre-specified maximum number of components, taken to be 10 in our application. For the component weights, the prior is taken to be Dirichlet $D(1, 1, \cdots, 1)$. Further details can be found in Richardson and Green [13].

4 Leukemia dataset

The Leukemia dataset of Golub *et al.* [6] is the result of monitoring the expression levels of 7129 genes in two types of acute leukemia using Affymetrix high-density oligonucleotide array technology. The dataset consists of a training set which contains 27 samples of acute lymphoblastic leukemia (ALL), and 11 samples of acute myeloblastic leukemia (AML), and a test set comprising 20 ALL and 14 AML samples. The ALL samples could be further classified as ALL-B or ALL-T, depending on whether they arise from a B or T cell lineage. The 27 ALL training samples contain 19 ALL-B, and 8 ALL-T samples, while the 20 ALL test samples contain 19 ALL-B and one ALL-T sample. We refer to the problem of discriminating between ALL and AML as the two-class problem. Discriminating among ALL-B, ALL-T, and AML is referred to as the three-class problem.

5 Results

Checking the normality assumptions

In our mixture modeling of test samples, we assume that each component density

of the mixture follows a normal distribution. Therefore, we first checked whether this assumption is reasonable for the leukemia dataset. We assigned the test samples to the appropriate classes, because the true underlying class assignments for these samples are in fact known. Then, for the samples in each class, we computed the standardize expression levels (by subtracting the sample mean and dividing by the sample standard deviation within each class) for each gene, and they are plotted against normal scores. For the three-class problem, the results are shown in Figure 1. There are some obvious departures from normality, although they do not seem to be sufficiently bad to cast serious doubt on the validity of the assumption. The normality plots for the two-class problem are similar.

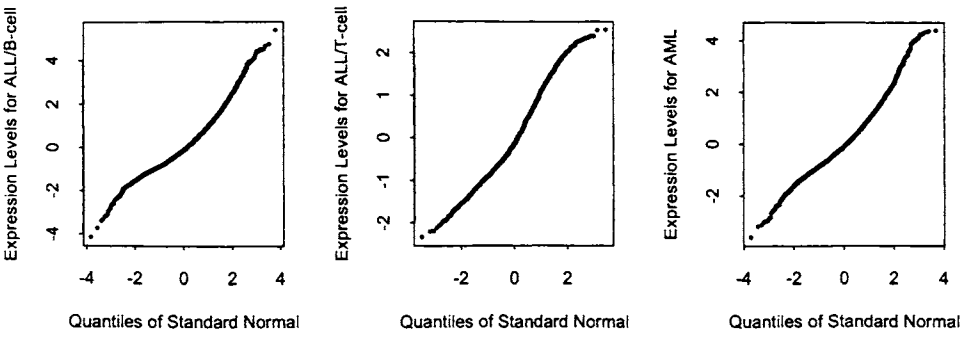


Figure 1: Normal probability plots for the samples in three classes, ALL-B, ALL-T, and AML.

Predictions for the two-class problem

For each of the four types of classifiers (*T*, *NPT*, *BW*, and *NPBW*) and a range of classifier sizes (1–200 genes), predictions of class assignments were carried out both for the training samples (through LOOCV) and the test samples using the mixture modeling approach. Figure 2(a) plots the prediction accuracy rates for the LOOCV of the training samples. The mixture modeling approach with the two types of non-parametric classifiers (*NPT* and *NPBW*) perform similarly; prediction accuracy rates of 1 are achieved in most of the range when the numbers of genes in the classifiers are more than 30. The mixture procedure does not perform as well with the parametric classifiers (*T* and *BW*), but the accuracy rates are still about 95% in most of the range. In summary, the results from LOOCV indicate that the prediction accuracy rates using the mixture modeling approach is not very sensitive to the type of classifier (among the four types that are considered here) nor the number of genes in a classifier, as long as the number is not very small.

Figure 2(b) plots the prediction accuracy rates for the test samples. Prediction accuracy rates of 1 are achieved only for the T classifiers with 19, 20, and 22 genes. Similar to the results in LOOCV, the performances of the procedures are not very sensitive to

the number of genes in the classifier as long as the number is not too small. Apart from *BW*, the results are not very sensitive to the types of classifiers in a wide range. The *BW* classifiers have not performed as well as the others.

A simulation study similar to that of Dudoit *et al.* [3] was carried out to further evaluate the performance of the mixture modeling procedure and to compare the four types of classifiers, including the effect of the size of a classifier. A total of 200 simulations (replications) were performed. For each simulation, 2/3 of the samples in each class (31 out of 47 ALL and 17 out of 25 AML) were randomly selected as the training samples, while the remaining served as the test samples. Each of the four types of classifiers with the sizes ranging from 1–200 was considered. The prediction results from the mixture procedures with all four types of classifiers are given in Figures 2(c) and 2(d). Specifically, the summary statistics for the number of test samples misclassified among the 200 replications for the T classifiers are plotted in Figure 2(c). For classifiers that are not very small, the results show that (1) there are no prediction errors in more than 25% of the replications, and (2) there are at most one prediction error in more than 75% of the replications. The performances for the three other types of classifiers are similar; full results are available from our web site (URL provided at the end of the article). The medians of the numbers of genes classified incorrectly (among 200 replications) for all four types of classifiers are plotted in Figure 2(d). We observe consistent results in all four classifier types for a wide range of classifier sizes.

We further examine the results from the simulation study by looking at each replication separately, instead of looking at the summary statistics, hoping to gain more insight into the relative performances of the four types of classifiers. Four pairs of classifiers are examined: *T* and *NPT*, *BW* and *NPBW*, *T* and *BW*, *NPT* and *NPBW*. One could examine other pairs, or higher number of classifiers jointly, but these four pairs seem the most appropriate ones to consider. For each pair and each replication, we classify the outcome into one of three categories: classifier 1 is better (the same, or worse) than classifier 2, depending on whether the number of samples incorrectly assigned under classifier 1 is smaller (the same, or larger) than that under classifier 2. The results are given in Table 1. Classifier *T* is slightly better than classifier *NPT*, while classifier *NPBW* is slightly better than classifier *BW*, consistently for the three sizes of the classifiers examined. Furthermore, *T* seems to be better than *BW*, while their nonparametric counterparts perform almost exactly the same.

The results shown thus far are obtained under the mixture modeling formulation (2); that is, data in both the training samples and the test samples contribute to the estimation of mixture component density parameters as well as the mixing proportions. Results using formulation (1) are similar, especially for LOOCV as expected, although not quite as good in predicting the original test samples, which is not surprising either since there are almost as many test samples as there are training samples. The full results can be obtained from our web site.

MCMC classifier. A total of 100,000 iterations were performed. The first 50,000 iterations were discarded to allow for convergence; the remaining realizations were then

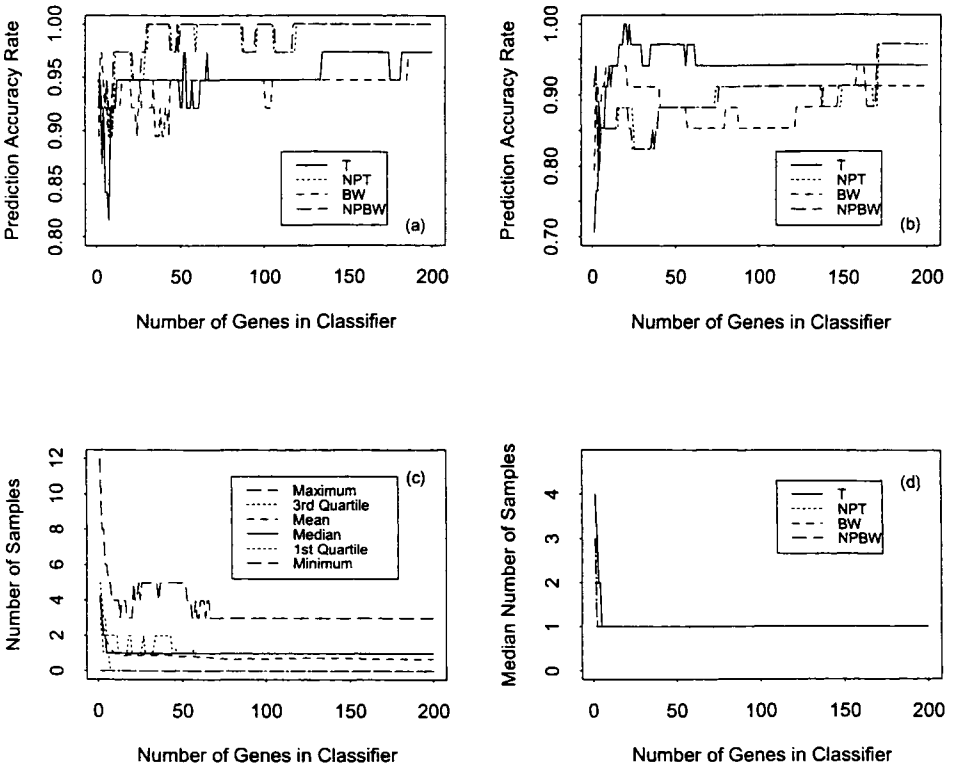


Figure 2: Prediction accuracy rates, or equivalently, the number of samples misclassified, for the training samples (through LOOCV) (a), the original test samples (b), and the simulated test samples (c and d), of the leukemia two-class problem. Figure 2(c) gives the summary statistics for classifiers based on T, and figure 2(d) plots the medians for T, NPT, BW, and $NPBW$.

used for inference. About 95% of the iterations picked three as the number of components for the mixture, with the second component corresponding to the distribution for genes that do not exhibit differential expressions (the null component), *i.e.*, with mean=0 for the component density. By applying our gene selection criterion, 23 genes were selected for the MCMC classifier. Class predictions for the test samples (using mixture formulation (2)) were then performed using the MCMC classifier. Out of the 34 samples in total, only one is classified incorrectly, giving a prediction accuracy rate

Table 1: Comparisons of classifiers for the two-class problem using the simulated data based on the leukemia dataset

#Genes	C1	C2	C1>C2[a]	C1=C2[b]	C1<C2[c]
25	*T*	*NPT*	59	96	45
	BW	*NPBW*	7	143	50
	T	*BW*	76	95	29
	NPT	*NPBW*	0	200	0
50	*T*	*NPT*	64	102	34
	BW	*NPBW*	15	161	24
	T	*BW*	72	99	29
	NPT	*NPBW*	4	196	0
100	*T*	*NPT*	58	100	32
	BW	*NPBW*	7	149	44
	T	*BW*	87	86	27
	NPT	*NPBW*	2	197	1

[a]This column gives the number of replications that result in smaller number of misclassified samples under classifier 1 than classifier 2.

[b]This column gives the number of replications that result in the same number of misclassified samples under both classifiers.

[c]This column gives the number of replications that result in larger number of misclassified samples under classifier 1 than classifier 2.

of 97%. On the other hand, under mixture formulation (1), in which only the training samples are used to estimate the component densities, five test samples are classified incorrectly.

Predictions for the three-class problem

Since *T* and *NPT* are applicable only to binary classification problems, they are not considered for further discriminating between ALL-B and ALL-T. For each multi-type feasible classifier (*BW* and *NPBW*) and a wide range of sizes (1–200 genes), the mixture modeling approach under formulation (2) were applied to classify the training samples (through LOOCV) as well as the test samples. Figure 3(a) and 3(b) plot the prediction accuracy rates for the LOOCV of the training samples and the test samples, respectively. Behavior similar to that observed in Figures 2(a) and 2(b) (for the two-class problem) is apparent in these figures. Namely, the prediction accuracy rates are not very sensitive to the type of classifier, nor the size of the classifier, and the class assignments of the samples are predicted quite accurately. *NPBW* performs better in smaller classifiers, especially in predicting the test samples, while *BW* does slightly better in LOOCV of the training samples. Overall, though, the performances of the two types of classifiers are similar.

A similar simulation study to that for the two-class problem was carried out. For

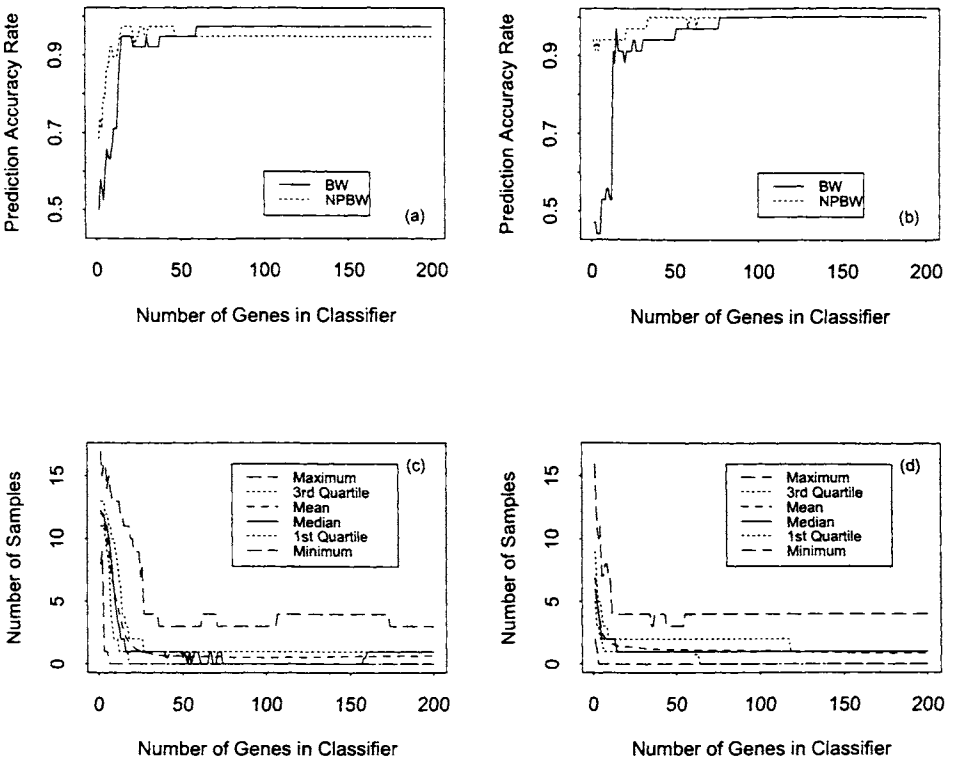


Figure 3: Prediction accuracy rates for the training samples (through LOOCV) (a), the original test samples (b), and the simulated test samples (c and d), of the leukemia three-class problem. Summary statistics for classifiers based on *BW* and *NPBW* are shown in (c) and (d), respectively.

each of the 200 replications, 2/3 of the samples in each class were randomly selected to form the training samples, and the remaining were assigned as test samples. For each classifier, the mixture approach under formulation (2) was applied to predict the class assignments of test samples. Summary statistics for the number of test samples classified incorrectly are plotted in Figure 3(c) for the *BW* classifiers, and 3(d) for the *NPBW* classifiers. Again, the mixture modeling approach yields good results for classifiers that are not too small, and *NPBW* performs slightly better for very small classifiers.

Table 2: Comparisons of classifier C1 (*BW*) and classifier C2 (*NPBW*) for the three-class problem using the simulated data based on the leukemia dataset

#Genes	C1>C2[a]	C1=C2[b]	C1<C2[c]
10	36	25	139
25	91	69	40
50	103	79	18
100	90	99	11
150	63	118	19
200	49	133	18

[a,b,c] See the footnotes of Table 1.

We further compare the performances of the two types of classifiers by examining each replication individually, in addition to the summary statistics across replications. For each replication, the outcome is classified into one of three categories: *BW* yielding smaller (same, larger) number of misclassified samples than *NPBW*. The results are shown in Table 2. We observe that, for smaller classifiers, there is a larger discrepancy between the two classifiers. Since *NPBW* is more robust to outlying expression levels, it is not surprising to see that it outperforms *BW* for the smallest classifier considered. As the number of genes in the classifiers increases, the two types of classifier become more similar, although *BW* continued to slightly outperform *NPBW* for larger classifiers.

The results shown thus far for the three-class problem are obtained using the mixture modeling formulation (2). Results using formulation (1) are similar, although not quite as good in predicting the original test samples, as what was observed for the two-class problem (full results available from our web site).

6 Discussion

In this article, we propose a method for classification of tissue samples by modeling the (multivariate) distribution of gene expression levels in a test sample as a mixture of distributions, each characterizing the distribution of the levels of gene expressions in a known class. This method can be paired with many gene selection methods (*i.e.*, methods for building classifiers) to reduce the dimensionality of the problem. Several classifiers are studied; results on *T*, *NPT*, *BW*, *NPBW*, and the MCMC classifier are presented in the current article, while results on several other binary classifiers can be found at our web site. Among the classifiers that are applicable to two-class problems, *T* performs well compared to the others in terms of prediction accuracy rates for the test samples (*T* achieving 100% accuracy rates for three classifier sizes) and the simulated samples in the leukemia dataset using the mixture modelling approach for classification. The MCMC classifier also performs well, with one prediction error out of a total of 34

test samples. Although the work on the MCMC classifier is still very preliminary, we are encouraged by these promising results, and effort is underway to extend it to handle multi-type classification problems. For multi-type feasible classifiers, *BW* is generally better than *NPBW* for predicting the training samples (through LOOCV) and the simulated samples, although *NPBW* is better in predicting the test samples, and *NPBW* classifiers were usually better than *BW* classifiers for smaller classifiers, again for the leukemia dataset. Note that the sizes of the classifiers that perform well are usually larger for the three-class problem than for the two-class problem, although they are all quite small (< 200) compared to the original number of genes. For predictions using the mixture modeling approach without first doing gene selection, three and four test samples are misclassified for the two-class and three-class problem, respectively, confirming the importance of gene selection.

Due to the lack of true test samples in the leukemia dataset, we were able to explore prediction accuracy rates for the test samples for a range of classifier sizes. In a real data analysis situation, however, we would proceed with the classification procedure proposed in this article in the following fashion. First, one would perform LOOCV with the training samples for a wide range of classifiers and sizes. Then a small set of classifiers that had performed well would be selected for classifying the test samples. We strongly recommend using more than one classifier so that consistency of prediction results can be checked. If several classifiers that had performed equally well in cross-validation had also produced consistent results in classifying the test samples, it would be an indication of satisfactory results, although there is no guaranteer that all assignments were correct. On the other hand, if discrepancies occur, then the biologists might be able to study the samples that caused the discrepancies more closely using other information.

Mixture modeling of test samples is a flexible means for multi-type classification of tissue samples. We have investigated two alternative formulations of the likelihood. It is not surprising to see that the one utilizing both the training samples and the test samples for parameter estimations (formula (2)) outperforms the one based on training samples only to estimate the parameters of the component densities, in many cases. Compared to other methods that have also been proposed for multi-type classifications, our approach performs at least as well with the leukemia dataset. For example, for predicting class membership of test samples, our approach yielded results with no prediction errors with medium-size classifiers (both *BW* and *NPBW*). The naive Bayes approach of Keller *et al.* [8] also yielded no misclassifications, for a small number of classifiers. Among the approaches discussed in Dudoit et al. [3], the best results had one misclassification of the simulated test samples, for both the median and the third-quartile, out of a total of 200 replications. Our simulation study using *BW* (for most of the classifiers ranging from 40 to 160 genes) resulted in zero and one misclassifications for the median and third-quartile, respectively, also out of 200 replications. Although one dataset and a limited simulation study do not warrant general conclusions, the results that we have obtained thus far show that the mixture modeling approach, coupled

with a gene selection measure such as *BW* (or its non-parametric counterpart if extreme observations are present), is promising. We plan to further evaluate its performance, especially its ability for classifying with larger numbers of classes.

The mixture formulation is also flexible in that it can be extended to handle situations where there are no training samples (class discovery problems) or when there are training samples but some of the test samples do not belong to any of the known classes (joint analysis of classification and class discovery). The key is to modify the mixture likelihood so that it allows for components that do not correspond to any known classes.

In our demonstration of the usage of the mixture modeling approach, each component density is assumed to be multivariate normal. This assumption was made for convenience. This assumption was also made in other methods, such as the methods based on maximum likelihood discriminant analysis [3]. Although good results were obtained from our analyses of the leukemia dataset, we could have used other distributions that fit the data better, as our figures show that there are obvious departures from normality. If the EM procedure for obtaining maximum likelihood estimates is no longer feasible, other methods for obtaining the MLEs may be used, including MCMC methods. Furthermore, we assume that the genes in a classifier are independent. Again, this assumption can be lifted, as the likelihood formulation is completely general; the component densities can be true multivariate distributions.

Electronic-Database Information

The URL for the supplementary material is: `http://www.stat.ohio-state.edu/~statgen/PAPERS/GeneExpression.html`.

Acknowledgments

Shili Lin would like to thank Professor Terry Speed for his mentoring and encouragement during her years at Berkeley and beyond. This work was supported in part by NSF grant DMS–9971770 (to S. Lin).

Shili Lin, Department of Statistics, Ohio State University, Columbus,
`shili@stat.ohio-state.edu`

Roxana Alexandridis, Department of Statistics, Ohio State University, Columbus,
`roxana@stat.ohio-state.edu`

References

[1] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore,

J. Hudson Jr, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–511, 2000.

[2] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini. Tissue Classification with Gene Expression Profiles. *Journal of Computational Biology*, 7:559–584, 2000.

[3] S. Dudoit, J. Fridlyand, and T. P. Speed. Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. *Journal of the American Statistical Association*, 97:77–87, 2002.

[4] S. Dudoit, Y. H. Yang, M. J. Callow, and T. P. Speed. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, 12:111–139, 2002.

[5] The Chipping Forecast. *Supplement to Nature Genetics*, 21:1–60, 1999.

[6] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, 286:531–537, 1999.

[7] P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732, 1995.

[8] A. D. Keller, M. Schummer, L. Hood, and W. L. Ruzzo. Bayesian Classification of DNA Array Expression Data. Technical report, UW-CSE-2000-08-01, Department of Computer Science and Engineering, University of Washington, 2000.

[9] D. J. Lockhart, H. L. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, and H. Horton. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 14:1675–1680, 1996.

[10] G. J. McLachlan and T. Krishnan. *The EM algorithm and extensions*. John Wiley & Sons, Inc. New York, 1997.

[11] M. A. Newton, C. M. Kendziorski, C. S. Richmond, F. R. Blattner, and K. W. Tsui. On Differential Variability of Expression Ratios: Improving Statistical Inference about Gene Expression Changes from Microarray Data. *Journal of Computational Biology*, 8:37–52, 2001.

[12] D. A. Notterman, U. Alon, A. J. Sierk, and A. J. Levine. Transcriptional Gene Expression Profiles of Colorectal Adenoma, Adenocarcinoma, and Normal Tissue Examined by Oligonucleotide Arrays. *Cancer Research*, 61:3124–3130, 2001.

[13] S. Richardson and P. J. Green. On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, Series B*, 59:731–758, 1997.

[14] E. E. Schadt, C. Li, C. Su, and W. H. Wong. Analyzing High-Density Oligonucleotide Gene Expression Array Data. *Journal of Cellular Biochemistry*, 80:192–202, 2000.

[15] M. Schena, editor. *DNA Microarrays: A Practical Approach*. Oxford University Press, 1999.

[16] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science*, 270:467–470, 1995.

[17] Y. H. Yang, S. Dudoit, P. Luu, and T. P. Speed. Normalization for cDNA microarray data. In M. L. Bittner, Y. Chen, A. N. Dorsel, and E. R. Dougherty, editors, *Microarrays: Optical Technologies and Informatics, Proc. SPIE*, volume 4266, pages 141–152, 2001.