STUDIES IN CLASSIFICATION,
DATA ANALYSIS,
AND KNOWLEDGE ORGANIZATION

Classification and Multivariate Analysis for Complex Data Structures

Bernard Fichet
Domenico Piccolo
Rosanna Verde
Maurizio Vichi
Editors

Springer

Studies in Classification, Data Analysis, and Knowledge Organization

For further volumes:
http://www.springer.com/series/1564

Bernard Fichet · Domenico Piccolo ·
Rosanna Verde · Maurizio Vichi
Editors

Classification and Multivariate Analysis for Complex Data Structures

Springer

Editors
Professor Bernard Fichet
Laboratoire d'Informatique
Fondamentale
Université d'Aix-Marseille II
163, Av. de Luminy – case 901
13288 Marseille cedex 9
France
Bernard.Fichet@lif.univ-mrs.fr

Professor Rosanna Verde
Facoltà di Studi Politici "Jean Monnet"
and Dipartimento di Studi Europei e
Mediterranei
Seconda Università di Napoli
Via del Setificio, 15
Sito Reale del Belvedere di San Leucio
81100 Caserta
Italy
rosanna.verde@unina2.it

Professor Domenico Piccolo
Dipartimento di Scienze Statistiche
Università di Napoli "Federico II"
Via Leopoldo Rodinò 22
80138 Naples
Italy
domenico.piccolo@unina.it

Professor Maurizio Vichi
Facoltà di Scienze Statistiche
Università di Roma "La Sapienza"
P. le Aldo Moro 5
00185 Rome
Italy
Maurizio.Vichi@uniroma1.it

This book has been printed with the grant of the Facoltà di Studi Politici e per l'Alta Formazione Europea e Mediterranea "Jean Monnet" and the Dipartimento di Studi Europei e Mediterranei of the Seconda Università di Napoli

ISSN 1431-8814
ISBN 978-3-642-13311-4 e-ISBN 978-3-642-13312-1
DOI 10.1007/978-3-642-13312-1
Springer Heidelberg Dordrecht London New York

Cover design: WMX Design, Heidelberg

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

This volume provides the latest advancements in statistical methods for multidimensional data analysis which can have a complex structure and collects a selection of revised papers presented at the first Joint Meeting of the Société Francophone de Classification and the Classification and Data Analysis Group of the Italian Statistical Society (SFC-CLADAG 2008) which was held in Caserta, June 11–13, 2008. Bernard Fichet and Domenico Piccolo co-chaired the Scientific Programme Committee and Rosanna Verde chaired the Local Organising Committee.

The meeting brought together a large number of scientists and experts, especially from Italy and francophone countries. It was a highly appreciated opportunity of discussion and mutual knowledge exchange about *techniques and tools for analyzing, classifying and summarizing statistical information, as well as for discovering and characterizing trends, and for automatically bagging anomalies.*

Special attention was paid to new methodological contributions from both the theoretical and the applicative point of views, in the fields of Clustering, Classification, Time Series Analysis, Multidimensional Data Analysis, Knowledge Discovery from Large Datasets, Spatial Statistics.

Upon conclusion of the joint meeting a cooperation agreement between the SFC and CLADAG was signed by Gilles Venturini, president of the SFC and Andrea Cerioli, president of CLADAG for a continuation of the scientific collaboration between the two groups around subjects related to the Classification and Multivariate Data Analysis.

The volume is structured in nine sections. The first contains the key note speakers papers, with eminent contributions in Classification and Data Analysis fields in plenary and semi-plenary sessions, by Edwin Diday, Carlo Lauro, Jacqueline Meulman, Paolo Giudici, André Hardy and Roberta Siciliano. The extended versions of some of their contributions were then written in collaboration with other colleagues.

Furthermore, the other 43 selected papers have been collected in 8 sections according to the following macro-topics:

- Classification and discrimination
- Data mining
- Robustness and classification
- Categorical data and latent class approach
- Latent Variables and related methods

- Symbolic, multi-valued and conceptual data analysis
- Spatial, temporal, streaming and functional data analysis
- Bio and health science

We wish to thank the authors for their contributions and the referees who carefully reviewed the papers: S. Balbi (Università di Napoli Federico II), F. Bartolucci (Università di Urbino "Carlo Bo"), Ch. Biernacki (Université LILLE I), M. Bini (Università di Firenze), H.-H. Bock (RWTH Aachen), A. Cerioli (Università di Parma), M. Chiodi (Università di Palermo), A. Chouakria-Douzal (Université Joseph Fourier Grenoble), M. Civardi (Università di Milano Bicocca), M. Corduas (Università di Napoli Federico II), F.T.A. De Carvalho (UFPE Brazil), M.R. D'Esposito (Università di Salerno), F. Domenach (University of Nicosia), V. Esposito Vinzi (ESSEC Paris), L. Fabbris (Università di Padova), L. Ferré (Université Toulouse Le Mirail), J. Gama (University of Porto), P. Giudici (Università di Pavia), A. Giusti (Università di Firenze), R. Gras (Université de Nantes), A. Guénoche (IML Marseille), G. Hébrail (ENST Paris), S. Ingrassia (Università di Catania), F.-X. Jollois (Université Paris Descartes) P. Kuntz (LINA, Université Nantes), M. La Rocca (Università di Salerno), Y. Lechevallier (INRIA Rocquencourt), V. Makarenkov (Université de Montréal), S. Mignani (Alma Mater Studiorum Università di Bologna), A. Mineo (Università di Palermo), M. Noirhomme (Université de Namur), F. Palumbo (Università di Napoli Federico II), C. Rampichini (Università degli Studi di Firenze), M. Rémon (Facultés Universitaires Notre-Dame de la Paix, Namur), M. Riani (Università di Parma), R. Rocci (Università di Roma Tor Vergata), F. Rossi (ENST Paris), G. Scepi (Università di Napoli Federico II), R. Siciliano (Università di Napoli Federico II), N. Torelli (Università di Trieste), Claus Weihs (Universität Dortmund), S. Zani (Università di Parma), D. Zighed (Université de Lyon).

We gratefully acknowledge the *Facoltà di Studi Politici e per l'Alta Formazione Europea e Mediterranea "Jean Monnet"* and the *Dipartimento di Studi Europei e Mediterranei* of the *Seconda Università di Napoli* for financial and organisation support and *EPT, Scuola Superiore della Pubblica Amministrazione, Sovraintendenza ai Beni Culturali* for the authorization for the use of the rooms and the theatre of the Royal Palace of Caserta for the conference sessions.

Special thanks are due to the members of the local Organising Committee for their work and especially to Antonio Irpinio and Antonio Balzanella for their efficient assistance in managing the web review process.

We are grateful to Hans-Hermann Bock for his help and fruitful support during the reviewing process and to have facilitated our contact with Springer-Verlag

Finally we would like to thank Dr. M. Bihn and her colleagues from Springer-Verlag, Heidelberg, for the excellent cooperation in publishing this volume.

Marseille, France — Bernard Fichet
Napoli, Italy — Domenico Piccolo
Caserta, Italy — Rosanna Verde
Roma, Italy — Maurizio Vichi
Maggio 2010

Contents

Contributors

R. Abdesselam ERIC EA 3038, University of Lyon2, 69676 Bron Cedex, France, rafik.abdesselam@univ-lyon2.fr

Giada Adelfio Department of Statistical and Mathematical Sciences, University of Palermo, 90128 Palermo, Italy, adelfio@dssm.unipa.it

Jaromír Antoch Department of Probability and Statistics, Charles University of Prague, CZ-186 75 Prague 8, Czech Republic, jaromir.antoch@mff.cuni.cz

Phipps Arabie Faculty of Management, Rutgers University, New Brunswick, NJ, USA, arabie@andromeda.rutgers.edu

Massimo Aria Department of Mathematics and Statistics, University of Naples Federico II, Naples, Italy, aria@unina.it

Marie-Aude Aufaure Laboratoire MAS, Ecole Centrale Paris, 92295 Châtenay-Malabry, France, Marie-Aude.Aufaure@ecp.fr

Luigi Augugliaro Dipartimento di Scienze Statistiche e Matematiche, University of Palermo, 90128 Palermo, Italy, augugliaro@dssm.unipa.it

Hanane Azzag Computer Science Laboratory CNRS – UMR 7030, Paris XIII University, France, hanane.azzag@lipn.univ-paris13.fr

Dunarel Badescu Département d'informatique, Université du Québec à Montréal, Montréal, Québec H3C 3P8, Canada, badescu.dunarel@courrier.uqam.ca

Antonio Balzanella Universitá degli Studi di Napoli Federico II, 80126 Napoli, Italy, balzanella2@alice.it

Stefano Bertelli Divisione Banche Estere, Intesa Sanpaolo, Milano, Italy, stefano.bertelli@intesasanpaolo.com

Matilde Bini Department of Economics, European University of Rome Via degli Aldobrandeschi, 190 00163 Roma, mbini@unier.it

Mathieu Blanchette School of Computer Science, McGill Centre for Bioinformatics, McGill University, Montréal, Québec H3A 2B4, Canada, blanchem@mcb.mcgill.ca

Giuseppe Boari Dipartimento di Scienze statistiche, Università Cattolica del Sacro Cuore, Milano, Italy, giuseppe.boari@unicatt.it

Concetto Elvio Bonafede Univesity of Pavia, Pavia, Italy, ingc.bonafede@gmail.com

Wender L.P. Bredie University of Copenhagen, Copenhagen, Denmark, wb@life.ku.dk

D. Bruzzese Department of Preventive Medical Sciences, Federico II University, Naples, Italy, dario.bruzzese@unina.it

Marine Cadoret Laboratoire de mathématiques appliquées, Agrocampus Ouest, 35042 Rennes Cedex, France, marine.cadoret@agrocampus-ouest.fr

Daniela G. Calò Department of Statistics, University of Bologna, 40126 Bologna, Italy, danielagiovanna.calo@unibo.it

Gabriele Cantaluppi Dipartimento di Scienze statistiche, Università Cattolica del Sacro Cuore, Milano, Italy, gabriele.cantaluppi@unicatt.it

Gabriele Cantaluppi Dipartimento di Scienze statistiche, Università Cattolica del Sacro Cuore, Milano, Italy, gabriele.cantaluppi@unicatt.it

Paola Cerchiello Univesity of Pavia, Pavia, Italy, paola.cerchiello@unipv.it

Marcello Chiodi Department of Statistical and Mathematical Sciences, University of Palermo, 90128 Palermo, Italy, chiodi@unipa.it

S. Cieply CREM UMR CNRS 6211, University of Caen-Basse Normandie, 14032 Caen Cedex, France, sylvie.cieply@unicaen.fr

Aldo Corbellini Economics Department, Università di Parma, Parma, Italy, aldo.corbellini@unipr.it

Marcella Corduas Dipartimento di Scienze Statistiche, Universitá di Napoli Federico II, Via L. Rodinò 22, 80138 Napoli, Italy, corduas@unina.it

Michele Costa Dipartimento di Scienze Statistiche, Università di Bologna, Bologna, Italy, michele.costa@unibo.it

Lisa Crosato Department of Statistics, Università di Milano-Bicocca, Milano, Italy, lisa.crosato@unimib.it

Carles M. Cuadras Department of Statistics, University of Barcelona, Barcelona, Spain, ccuadras@ub.edu

Daniel Cuadras IDIBELL, Barcelona, Spain, danicuadras@gmail.com

Antonio D'Ambrosio Department of Mathematics and Statistics, University of Naples Federico II, Naples, Italy, antdambr@unina.it

Giulio D'Epifanio Faculty of Political Science, University of Perugia, 06100 Perugia, Italy, ggiulio@stat.unipg.it

Maria Rosaria D'Esposito Dipartimento di Scienze Economiche Statistiche, Universitá di Salerno, 84084 Fisciano, Salerno, Italy, mdesposito@unisa.it

Luca De Angelis Dipartimento di Scienze Statistiche, Università di Bologna, Bologna, Italy, l.deangelis@unibo.it

Domenico De Stefano Dipartimento di Scienze Economiche e Statistiche, Università di Salerno, 84084 Fisciano Salerno, Italy, d.destefano@unina.it

Emanuele Dequarti Department of Statistics and Applied Economics Libero Lenti, University of Pavia, Pavia, Italy, emanuele.dequarti@unipv.it

Agostino Di Ciaccio Department of Statistics, Probability and Applied Statistics, Sapienza, University of Rome, Rome, Italy, agostino.diciaccio@uniroma1.it

Abdoulaye Baniré Diallo Département d'informatique, Université du Québec à Montréal, Montréal, Québec H3C 3P8, Canada; School of Computer Science, McGill Centre for Bioinformatics, McGill University, Montréal, Québec H3A 2B4, Canada, diallo.abdoulaye@uqam.ca

Edwin Diday CEREMADE, Université Paris-Dauphine, Paris, France, diday@ceremade.dauphine.fr

Sébastien Duplessis Centre INRA Nancy, Champenoux, France, duplessi@nancy.inra.fr

Luigi Fabbris Statistics Department, University of Padua, Padua, Italy, luigi.fabbris@unipd.it

Miguel Fernández Departamento de Estadística, Universidad de Valladolid, 47005 Valladolid, Spain, miguelaf@eio.uva.es

Antonio Forcina Dipartimento di Economia, Finanza e Statistica, 06100 Perugia, Italy, forcina@stat.unipg.it

U. Giani Department of Preventive Medical Sciences, Federico II University, Naples, Italy, fdario.bruzzesegfumberto.gianig@unina.it

Giuseppe Giordano University of Salerno, 84084 Fisciano, Salerno, Italy, ggiordan@unisa.it

Francesco Giordano Department of Economics and Statistics, University of Salerno, 84084 Fisciano, Salerno, Italy, giordano@unisa.it

Paolo Giudici Department of Statistics and Applied Economics Libero Lenti, University of Pavia, Pavia, Italy, giudici@unipv.it

M.G. Grassia Department of Mathematics and Statistics, University "Federico II" Naples, 80125 Naples, Italy, mgrassia@unina.it

Leonardo Grilli Department of Statistics "G. Parenti", University of Florence, Florence, Italy, grilli@ds.unifi.it

Christiane Guinot CE.R.I.E.S, Neuilly-sur-Seine, 92521 France, christiane.guinot@ceries-lab.com

André Hardy University of Namur, 5000 Namur, Belgium, andre.hardy@fundp.ac.be

Lawrence J. Hubert Department of Psychology, University of Illinois at Urbana-Champaign Champaign, IL, USA, lhubert@cyrus.psych.uiuc.edu

Mauro Iacono DEM, Seconda Universitá degli Studi di Napoli Belvedere Reale di San Leucio, Caserta, Italy, mauro.iacono@unina2.it

Antonio Irpino DEM, Seconda Universitá degli Studi di Napoli Belvedere Reale di San Leucio, Caserta, Italy, antonio.irpino@unina2.it

Mehdi Kaytoue LORIA, Campus Scientifique, Vandoeuvre-lés-Nancy, France, kaytouem@loria.fr

Mohsen Kompany-Zareh Institute of Advanced Studies in Basic Sciences, Zanjan, Iran, kompanym@iasbs.ac.ir

Michele La Rocca Department of Economics and Statistics, University of Salerno, 84084 Fisciano, Salerno, Italy, larocca@unisa.it

Carlo Natale Lauro Università degli Studi di Napoli "Federico II". Monte Sant' Angelo, 80125 Napoli, Italy, carlo.lauro@unina.it

Julien Lavergne Computer Science Laboratory, François-Rabelais University, Tours, France, julien.lavergne@univ-tours.fr

Sébastien Lê Laboratoire de mathématiques appliquées, Agrocampus Ouest, 35042 Rennes Cedex, France, sebastien.le@agrocampus-ouest.fr

Yves Lechevallier Projet AxIS, INRIA Paris Rocquencourt, Domaine de Voluceau, Rocquencourt, B.P. 105, 78153 Le Chesnay Cedex, France, Yves.Lechevallier@inria.fr

Bénédicte Le Grand Laboratoire d'Informatique de Paris 6, 75016 Paris, France, Benedicte.Le-Grand@lip6.fr

A.L. Le Nadant CREM UMR CNRS 6211, University of Caen-Basse Normandie, 14032 Caen Cedex, France, anne-laure.lenadant@unicaen.fr

Umberto Magagnoli Dipartimento di Scienze statistiche, Università Cattolica del Sacro Cuore, Milano, Italy, umberto.magagnoli@unicatt.it

Vladimir Makarenkov Département d'informatique, Université du Québec à Montréal, Montréal, Québec H3C 3P8, Canada, makarenkov.vladimir@uqam.ca

Marina Marino Department of Agricultural Engineering and Agronomy, University of Naples Federico II, 80055 Portici, Naples, Italy, marina.marino@unina.it

Mariagiulia Matteucci Statistics Department "Paolo Fortunati", University of Bologna, 40126 Bologna, Italy, m.matteucci@unibo.it

Giovanna Menardi Department of Economics and Statistics, University of Trieste, Trieste, Italy, giovanna.menardi@econ.units.it

Jacqueline J. Meulman Department of Mathematics, Leiden University, Leiden, The Netherlands, jmeulman@math.leidenuniv.nl

R. Miele Department of Mathematics and Statistics, University "Federico II" Naples, 80125 Naples, Italy, rafmiele@unina.it

Raffaele Miele Department of Mathematics and Statistics, University of Naples Federico II, 80126 Napoli, Italy, rafmiele@unina.it

Angelo M. Mineo Dipartimento di Scienze Statistiche e Matematiche, University of Palermo, 90128 Palermo, Italy, elio.mineo@dssm.unipa.it

Isabella Morlini Dipartimento di Scienze Sociali, Cognitive e Quantitative, Università di Modena e Reggio Emilia, 42100 Reggio Emilia, Italy, isabella.morlini@unimore.it

Massimo Mucciardi Department D.E.S.Ma.S. "V. Pareto", University of Messina, Messina, Italy, massimo.mucciardi@unime.it

Amedeo Napoli LORIA, Campus Scientifique, Vandoeuvre-lés-Nancy, France, napoli@loria.fr

D. Nappo Department of Matematic and Statistics, University "Federico II" Naples, 80125 Naples, Italy, daniela.nappo@unina.it

Jérôme Pagès Laboratoire de mathématiques appliquées, Agrocampus Ouest, 35042 Rennes Cedex, France, jerome.pages@agrocampus-ouest.fr

Francesco Palumbo Dipartimento di Istituzioni Economiche e Finanziarie, Universitá di Macerata, 62100 Macerata, Italy, francesco.palumbo@unimc.it

Maria Lucia Parrella Department of Economics and Statistics, University of Salerno, 84084 Fisciano, Salerno, Italy, mparrella@unisa.it

Marcello Pecoraro CRM – Intesa Sanpaolo Group, Piazza S. Carlo 156, Turin (Italy), marcello.pecoraro@unina.it

Marco Pitzalis Dipartimento di Ricerche Economiche e Sociali, Università di Cagliari, Cagliari, Italy, pitzalis@unica.it

Mariano Porcu Dipartimento di Ricerche Economiche e Sociali, Università di Cagliari, Cagliari, Italy, mrporcu@unica.it

Giancarlo Ragozini Dipartimento di Sociologia "G. Germani", Universitá Federico II, 80138 Napoli, Italy, giragoz@unina.it

Rosaria Romano Seconda Universitá degli Studi di Napoli, Napoli, Italy, romano.rosaria@gmail.com

Elvira Romano Seconda Universitá degli Studi di Napoli, 81100 Caserta, Italy, elvroman@unina.it

Fabrice Rossi Institut Télécom, Télécom ParisTech, LTCI – UMR CNRS 5141, 75013 Paris, France, Fabrice.Rossi@telecom-paristech.fr

Cristina Rueda Departamento de Estadística, Universidad de Valladolid, 47005 Valladolid, Spain, miguelaf@eio.uva.es

Giorgio Russolillo Università degli Studi di Napoli "Federico II", Napoli, Italy, giorgio.russolillo@unina.it

Bonifacio Salvador Departamento de Estadística, Universidad de Valladolid, 47005 Valladolid, Spain, miguelaf@eio.uva.es

Germana Scepi University of Naples, Monte Sant'Angelo, Naples, Italy, scepi@unina.it

Roberta Siciliano Department of Mathematics and Statistics, University of Naples Federico II, Naples, Italy, roberta@unina.it

Simona Signoriello Department of Matematic and Statistics, University of Naples Federico II, 80125 Naples, Italy, simona.signoriello@unina.it

Michel Soto Laboratoire d'Informatique de Paris 6, 75016 Paris, France, Michel.Soto@lip6.fr

Isabella Sulis Dipartimento di Ricerche Economiche e Sociali, Università di Cagliari, Cagliari, Italy, isulis@unica.it

Federico Tedeschi Department of Economics and Statistics, University of Trieste, Trieste, Italy

Nicola Torelli Department of Economics and Statistics, University of Trieste, Trieste, Italy

Laura Trinchera Department of Signal Processing & Electronic Systems, SUPELEC, Gif-sur-Yvette, France, laura.trinchera@supelec.fr

Valerio A. Tutore Department of Mathematics and Statistics, University of Naples Federico II, Naples, Italy, v.tutore@unina.it

Bernard P. Veldkamp Department of Research Methodology, Measurement and Data Analysis, University of Twente, 7500 AE, Enschede, The Netherlands, b.p.veldkamp@gw.utwente.nl

Margherita Velucchi Dipartimento di Statistica "G. Parenti", Università di Firenze, 50134 Firenze, Italia, velucchi@ds.unifi.it

Gilles Venturini Computer Science Laboratory, François-Rabelais University, Tours, France, gilles.venturini@univ-tours.fr

Rosanna Verde Seconda Universitá degli Studi di Napoli, Facoltà di Studi Politici “Jean Monnet” and Dipartimento di Studi Europei e Mediterranei, 81100 Caserta, Italy, rosanna.verde@unina2.it

Jannie S. Vestergaard University of Copenhagen, Copenhagen, Denmark, jve@life.ku.dk

Part I
Key Notes

Principal Component Analysis for Categorical Histogram Data: Some Open Directions of Research

Edwin Diday

Abstract In recent years, the analysis of symbolic data where the units are categories, classes or concepts described by interval, distributions, sets of categories and the like becomes a challenging task since many applicative fields generate massive amount of data that are difficult to store and to analyze with traditional techniques [1]. In this paper we propose a strategy for extending standard PCA to such data in the case where the variables values are "categorical histograms" (i.e. a set of categories called bins with their relative frequency). These variables are a special case of "modal" variables (see for example, Diday and Noirhomme [5]) or of "compositional" variables (Aitchison [1]) where the weights are not necessarily frequencies. First, we introduce "metabins" which mix together bins of the different histograms and enhance interpretability. Standard PCA applied on the bins of such data table loose the histograms constraints and suppose independencies between the bins but copulas takes care of the probabilities and the underlying dependencies. Then, we give several ways for representing the units (called "individuals"), the bins, the variables and the metabins when the number of categories is not the same for each variable. A way for representing the variation of the individuals, for getting histograms in output is given. Finally, some theoretical results allow the representation of the categorical histogram variables inside a hypercube covering the correlation sphere.

1 Introduction

Recent advances in Symbolic Data Analysis (SDA) as Billard and Diday [2], Diday and Noirhomme-Fraiture [5] have motivated the development of strategies for the analysis of data where the variable values (called "symbolic data") are intervals, categorical sequences sometimes weighted, distributions or histograms. Application fields include socio demographic surveys, medical information management, biology, forecasting and climate monitoring, telecommunications.

E. Diday (✉)
CEREMADE, Université Paris-Dauphine, Paris, France,
e-mail: diday@ceremade.dauphine.fr

B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-13312-1_1,

Here we are interested on what we have called "categorical histogram" data, where each variable value for each individual is a set of the relative frequencies associated to the categories (or bins) of this variable. That is why the sum of the weights is equal to 1 and these numbers can be considered as the probability of their associated category for the corresponding individual. Therefore, these data are a case of what has been called in the SDA framework "modal multi-valued" data where the value taken by each variable (called "modal") for each individual is a sequence of weighted categories. Categorical histogram data are also a case of compositional data (Aitchison [1, 3]) where sum of the weights of the categories for a given variable remains a constant when the individuals vary. Nagabushan et al. [11] approach consists of doing the standard PCA of each table T_k for $k = 1, \ldots, m$ where each individual is described by its kth bin for each of the p variables Other approaches like Rodriguez et al. [13] or Ichino [7, 8] are based on the transformation of the histograms in distributions which is not possible in the case of non ordinal bins. Makosso Kallyth et al. [10] approach needs the same number of bins supposed ordered for each variable. The method presented in this paper can be applied to ordinal or nominal (i.e. non ordinal) bins with different number of bins for each variable. We first present the so called "categorical histogram data", we define and build examples of metabins which join together bins taken in each variable. Then, we present the standard PCA and the "Copular PCA" which improves the standard one by taking care of the dependencies between the bins. We give tools for the representation of individuals, individual × variables, individual × metabins in its factorial space with histograms at output following in that way one of the SDA framework principle saying that the output data have to be of the same kind than the input data. Finally, we give tools for the representation of variables or metabins in the correlation sphere of a PCA inside a hypercube.

2 The Categorical Histogram Data Table

The set of individuals (i.e. observations) is denoted $\Omega = \{\omega_1, \ldots, \omega_n\}$, the set of the histogram value variables is $Y = \{Y_1, \ldots, Y_p\}$. Each variable Y_j has m_j bins (i.e. categories) and is associated to its bin variables $\left(Y_{j1}, \ldots, Y_{jm_j}\right)$. The variables are such that the value of Y_j for an individual ω_i is $Y_j(\omega_i) = \left(Y_{j_1}(\omega_i), \ldots, Y_{jm_j}(\omega_i)\right) = \left(r_{ij_1}, \ldots, r_{ijm_j}\right) \in IR^{m_j}$. Hence $Y_{jk}(\omega_i) = r_{ijk}$ is the categorical histogram value of ω_i for the variable Y_j and its bin k. Therefore, we have for any individual i and variable j: $\sum_{k=1}^{m_j} r_{ijk} = 1$. In other words we obtain the table of Fig. 1. For example, if the data table describes the teams of the world cup and if the variable Y_j represents the nationality, the bins Y_{jk} are associated to countries and the individuals ω_i are associated to the teams, we can have for example:

NATIONALITY (*Spanish Team*) = [(0.8) *Spanish*, (0.1) *Brazilian*, (0.1) *French*]

Fig. 1 Initial categorical histogram data table characterized by $\sum_{k=1}^{m_j} r_{ijk} = 1$

	Y_1			...	Y_p		
	Y_{11}	...	Y_{1m_1}	...	Y_{p_1}	...	Y_{pm_p}
ω_1	r_{111}	...	r_{11m_1}	...	r_{1p_1}	...	r_{1pm_p}
...	...	...	...	...	...	...	...
ω_n	r_{n11}	...	r_{n11}	...	r_{np_1}	...	r_{npm_p}

where the weights as (0.8) represent the frequencies of the nationalities in this team. We can have two kinds of categorical histogram value variables: the "nominal" case where the bins are not ordered as for example countries in the example of the nationality variable and the ordinal case where the bins are ordered as the intervals of age in the following example:

$$AGE\ (Spanish\ Team) = [(0.5)[20, 25[, (0.3)[25, 30[, (0.2)[30, 35]]\,.$$

3 Building "Metabins" by Scoring the Bins in Case of Nominal Histogram Variables

3.1 What Are "Metabins"?

It is possible to sort the bins of nominal histogram value variables by many ways. The simplest way is to sort the bins by the means of their frequencies on the set of individuals. For example having:

$$AGE\ (Spanish\ Team) = [(0.8)\ Spanish, (0.1)\ Brazilian, (0.1)\ French]$$
$$AGE\ (French\ Team) = [(0.2)\ Spanish, (0.2)\ English, (0.6)\ French]$$

We get the following mean and range:

$$[(0.5)\ Spanish, (0.35)\ French, (0.1)\ English, (0.05)\ Brazilian].$$

In practice, it is more interesting to sort the bins in such a way that the first bins, second bins etc. of all the variables be "linked" within the meaning of their (standard, Kendall, Spearman, Guttman, etc.) correlation denoted "cor". In Sect. 3.2 we give examples of such criteria. We denote $K = \left((k_{11}, \ldots, k_{1m_1}), \ldots, (k_{p1}, \ldots, k_{pm_p})\right)$ an ordered set of integers such that $k_{js} \leq m_j$ associated to the following bins vector: $\left((Y_{1k_{11}}, \ldots, Y_{1k_{pm_1}}), \ldots, (Y_{pk_{p1}}, \ldots, Y_{pk_{pm_p}})\right)$ which represents a reordering of the bins of each categorical histogram variable. We call "metabin" a set of p bins one for each variable.

Example From the table given Fig. 1 in case of $p = 4$, $S = \{Y_{21}, Y_{41}, Y_{61}, Y_{43}\}$is a metabin. But $S' = \{Y_{25}, Y_{43}, Y_{41}, Y_{34}\}$ is not a metabin as two bins belong to the same variable Y_4.

3.2 Metabins Quality Criteria, Correlation or Copulas Based

Many quality criteria of metabins can be defined. We can set, for example, that the best order of the bins is the one which maximizes the following criterion: $W(K) = \sum_{j,j'=1}^{p} \sum_{m=1}^{min(m_j,m_{j'})} cor^2(Y_{jk_{jm}}, Y_{j'k_{j'm}})$ which means that the "link" between the bins in same metabin for each pair of variables must be maximized.

Another kind of criterion based on "copulas" (see the Annex for a short recall on copulas) can be defined in the following way. First we define a joint probability distribution F_s based on a metabin $s = \{Y_{1k_{1m}}, \ldots, Y_{pk_{pm}}\}$ by $F_s(x_1, \ldots, x_p) = Cop(G_{1s}(x_1), \ldots, G_{ps}(x_p))$ where $G_{js}(x_j) = Prob(Y_{jk_{jm}} \leq x_j) = card\{w/Y_{jk_{jm}}(\omega) \leq x_j\}/n$ and $F_s(x_1, \ldots, x_p) = Prob(\cap_{j=1,p} Y_{jk_{jm}} \leq x_j)$.
Knowing the F_s and the G_{js}, the copula Cop can be estimated among a parametric copula family (as the Clayton or Frank family) which allows finding F_s by just knowing the marginals G_{js} and a criterion to optimize. An example of criterion W likelihood based can be written in the following way:

$W(K) = \sum_{m=1}^{\max_j(m_j)} L(F_m)$ or $W(K) = \sum_{s=1}^{\max_j(m_j)} Log(L(F_s))$ where $L(F_m)$ is the likelihood of F_s given by $L(F_m) = \prod_{i=1}^{n} F_s(Y_{1k_{1m}}(\omega_i), \ldots, Y_{pk_{pm}}(\omega_i))$.

Many non parametric copulas are also possible as the "product" which expresses independencies: $F_s(Y_{1k_{1m}}(\omega_i), \ldots, Y_{pk_{pm}}(\omega_i))) = \prod_{j=1}^{p} Prob(Y_{jk_{jm}} \leq Y_{jk_{jm}}(\omega_j))$. A simpler possible choice is $F_s(Y_{1k_{1m}}(\omega_i), \ldots, Y_{pk_{pm}}(\omega_i)) = \underset{j=1,p}{Min} \left\{Prob(Y_{jk_{jm}} = Y_{jk_{jm}}(\omega_j))\right\}$. By using Archimedean copulas (see for example, Nelsen [12]) many other criteria can be also used.

4 PCA for Histogram Data Table Using Copulas

4.1 The Standard PCA on Histogram Data

After having centered and reduced the initial data table given in Fig. 1, we denote $X = \{X_{ij}\}_{i=1,n;\, j=1,p}$ its induced matrix of rows defined by the vectors $X_{ij} = (x_{ij1}, x_{ijk}, \ldots, x_{ijm_j})$ for $j = 1, p$. Its rows can be also associated to a vector X_i of $\sum_{k=1}^{p} m_k$ dimension, such that:

$$X_i^T = (x_{i11}, \ldots, x_{i1m_1}, x_{i21}, \ldots, x_{i2m_2}, \ldots, x_{ip1}, \ldots, x_{ipm_p}).$$

In other words, if the initial coordinates are denoted r_{ijk}, we have $x_{ijk} = (r_{ijk} - r_{.jk}/n)/(s_{jk}\sqrt{n})$ where $r_{.jk}/n$ and s_{jk} are respectively the mean and the mean

square of the bin variable Y_{jk}. As the bin variables Y_{jk} are centered and reduced we get $\sum_{i=1}^{n} Y_{jk}(\omega_i) = \sum_{i=1}^{n} x_{ijk} = 0$ and $\sum_{i=1}^{n} x_{ijk}^2 = 1$ for any j and k. Therefore, the condition $\sum_{k=1}^{m_j} x_{ijk} = 1$ is no more available. Standard PCA can be applied to the numerical centered and reduced data table of Fig. 1. In all the following we use the standard formulas given in Lebart et al. [9]. We denote D a diagonal matrix $p \times p$ where the ith term of the diagonal is a weight p_i associated to the ith individual such that $\sum_{i=1}^{n} p_i = 1$ and we look for a vector:

$$Z^T = (z_{11}, \ldots, z_{1m_1}, z_{21}, \ldots, z_{2m_2}, \ldots, z_{p1}, \ldots, z_{pm_p})$$

of $\sum_{k=1}^{p} m_k$ coordinates, such that $\sum_{i=1}^{n} p_i (X_i M Z)^2$, where M is a metric between individuals, be maximized for any i under the constraint $Z^T M Z = 1$. In other words we wish to maximize

$$\sum_{i=1}^{n} p_i X_i M Z X_i M Z = \sum_{i=1}^{n} p_i Z^T M X_i^T X_i M Z = Z^T M X^T D X M Z = Z^T M \Gamma Z = \lambda$$

where λ is a positive number and $\Gamma = X^T D X M$ is the standard correlation matrix of dimension $\sum_{k=1}^{p} m_k \times \sum_{k=1}^{p} m_k$. It is easy to see that a solution of the equation $Z^T M \Gamma Z = \lambda$ is $\Gamma Z = \lambda Z$ as $Z^T M (\Gamma Z) = \lambda Z^T M Z = Z^T M (\lambda Z)$. Hence, the solution is given by the Eigen vector of Γ which maximizes λ. This Eigen vector is called the first factor and denoted $Z^{(1)}$. The other best orthogonal solutions are given by the other Eigen vectors of W and denoted $Z^{(2)}, \ldots, Z^{(p)}$.

In the following, in order to simplify, we consider that M is reduced to the identity matrix even if all the results can be easily generalized to any metric defined by M. The coordinate i of the principal component $F^{(\alpha)}$ is by definition the projection of the ith individual ω_i on the factor $Z^{(k)}$. Therefore, its value is: $F_i^{(\alpha)} = \sum_{j=1}^{p} \sum_{k=1}^{m_j} x_{ijk} z_{jk}^{(\alpha)}$. For which it results more generally that $F^{(\alpha)} = X Z^{\alpha}$. The correlation between each variable and each factor gives its coordinates inside the so called "correlation sphere" of raw length equal to 1. More precisely, the coordinate of the bin variable Y_{ij} on the factor k is defined by: $cor(Y_{jk}, F^{(\alpha)}) = \sum_{i=1}^{n} x_{ijk} F_i^{(\alpha)} / \sqrt{\lambda_\alpha}$. Therefore, we get: $\sum_{\alpha=1}^{p} cor^2(Y_{jk}, F^{(\alpha)}) = 1$. Finally this method allows the mapping of the variables inside the correlation circle.

4.2 The Copular PCA

In the standard PCA developed in Sect. 4.1 we have not used the fact that probabilities can be associated to the frequency of each bin. If we wish to take care of that, "copulas" are needed. In order to simplify, we suppose that D is the $1/n$ diagonal matrix and M is the identity matrix. It results that Γ is the correlation matrix and so its generic term is:

$$C_{jj'} = \sum_{i=1}^{n} x_{ijk} x_{ij'k} = \sum_{i=1}^{n} (r_{ijk} - r_{.jk}/n)(r_{ij'k} - r_{.j'k}/n)/(ns_j s_{j'}).$$

Let R_{ij} be the random variable associated to the variable Y_j for the individual i such that $Prob(R_{ij}(k)) = r_{ijk}$. We can now see that in $C_{jj'}$ there are many products $r_{ijk} r_{ij'k} = Prob(R_{ij} = k) Prob(R_{ij'} = k')$. Let H be the joint probability between R_{ij} and $R_{ij'}$. Under the hypotheses of independence between R_{ij} and $R_{ij'}$, we get $H(R_{ij} = k, R_{ij'} = k') = Prob(R_{ij} = k) Prob(R_{ij'} = k') = r_{ijk} r_{ij'k}$. This means that the product get a high value if both r_{ijk} and $r_{ij'k}$ are high. In practice, this is not the case when there is no independencies between R_{ij} and $R_{ij'}$. In order to express the joint probability in a more realistic way, we need to introduce copulas which aim is to do the link between the marginals and the joint by the following formula (Shweizer and Sklar (1983)), for more details on copulas see annex):

$$H(R_{ij} = k, R_{ij'} = k') = Cop_{\theta}(Prob(R_{ij} = k), Prob(R_{ij'} = k')),$$

where H is the joint probability:

$$H(R_{ij} = k, R_{ij'} = k') = Prob((R_{ij} = k) \cap (R_{ij'} = k')),$$

where only Cop and θ are unknown and has to be estimated. The copula Cop exists and is unique under some hypothesis given in the Sklar theorem [15] recalled in annex. If the initial data are the native data from which the histograms have been build are known, and the copula family has been chosen, the parameter θ can sometimes be estimated by using the Kendall Tau between the random variables R_{ij} and $R_{ij'}$ (see Nelsen [12] for examples of such estimations).

5 Data Tables Derived from the Initial Categorical Histogram Data Table

5.1 The Use of the Derived Data Tables

Until now we have just used the standard PCA on the normalized histogram data table (given in Fig. 1) extended to copulas inside the correlation matrix. From the Fig. 2 we can see that many other data tables are also involved in this framework. The Tables 2–6 are based on the standard or copular PCA presented in Sects. 4.1

Variables	Individuals			Ind. × bins			Ind. × metabins			Ind. × Variables		
	ω_1	ω_i	ω_n	$\omega_{1.1}$	$\omega_{i.k}$	$\omega_{n.m_p}$	$\omega_{1.1}$	$\omega_{i.m}$	$\omega_{i.M}$	$\omega_{11.}$	$\omega_{ij.}$	$\omega_{np.}$
Bins:												
$Y_{.11}$	Table1			Table4			Table5			Table6		
$Y_{.jk}$												
$Y_{.pm_p}$												
Variables												
$Y_{1.} = Y_1$	Table2						Table7					
$Y_{.j.} = Y_j$												
$Y_{.p.} = Y_p$												
Metabins												
S_1	Table3									Table8		
S_m												
S_M												
F^{α}												

Fig. 2 Data tables derived from the initial categorical histogram data table

Fig. 3 Table 4 where $Y_{.jk}(\omega_{t.q}) = x_{ijk}$ if $(t, q) = (j, k)$ and equal to 0 if not

	$\omega_{1.1}$	...	$\omega_{i.k}$	...	$\omega_{n.m_p}$
$Y_{.11}$	x_{111}	...	$Y_{.11}(\omega_{i.k})$	...	$Y_{.11}(\omega_{n.m_p})$
...	...	...	...	...	...
$Y_{.jk}$	...	...	x_{ijk}	...	...
...	...	...	...	...	...
$Y_{.pm_p}$	$Y_{.pm_p}(\omega_{1.1})$	...	$Y_{.pm_p}(\omega_{i.k})$	...	x_{npm_p}

and 4.2. For example, the table 4 is detailed in Fig. 3. They allow the representation of supplementary individuals and variables as will be shown in Sect. 5.2.

5.2 Representation of Supplementary Individuals, Variables and Metabins

The idea is here to use the tables shown in Fig. 2 in order to represent inside the standard or copular PCA of Table 1, the contribution of bins, variables or specific individuals by using the table 2–6. More precisely, we use the Table 2 for the projection of the variables Y_j, the Table 3 for the projection of the metabins S_m, the table 4–6 respectively for the projection of the Individual × bins, the Individual × Variables and the Individual × Metabins. We use table 2 for the representation of new variables on the principal components $F^{(\alpha)}$.

5.3 Representation of the Variation of the Individuals According to the Bins, the Metabins and the Variables

The table 4 expresses the Cartesian product between the individuals and the bins described by the bins. In this case the coordinate on the factor $Z^{(\alpha)}$ of the individual $\omega_i \times (j, k)$ associated to the initial observation ω_i and the bin k of the variable Y_j is $x_{ijk} z_{jk}^{(\alpha)}$.

Hence, each individual wi can be considered as the mean center of the projections of $M = \sum_{j=1}^{p} m_j$ bins such that: $Proj^{\alpha}\left(w_i \times (k, j)/Z^{(\alpha)}\right) = Mx_{ijk}z_{jk}^{(\alpha)}$ for all the M bins on the factor $Z^{(\alpha)}$. This is easily proved as we know that $Proj^{\alpha}\left(w_i/Z^{(\alpha)}\right) = \sum_{j=1}^{p}\sum_{k=1}^{m_j} x_{ijk}z_{jk}^{(\alpha)}$ is the projection of ω_i on the factor $Z^{(\alpha)}$ and therefore, we get:

$$Mean\left\{Proj^{\alpha}\left(\omega_i \times (k, j)/Z^{(\alpha)}\right)/j = 1, p,\ k = 1, m_j\right\}$$
$$= \sum_{j=1}^{p}\sum_{k=1}^{m_j} Proj^{\alpha}\left(\omega_i \times (k, j)/Z^{(\alpha)}\right)/M = Proj^{\alpha}\left(\omega_i/Z^{(\alpha)}\right).$$

In practice, in order to avoid too many points we can select the projection of the bins according to a given variable or to a given metabin.

The table 5 expresses the Cartesian product between the individuals and the metabins described by their bins. In this case, the coordinate on the factor $Z^{(\alpha)}$ of the unit $\omega_i \times S_m$ associated to the initial observation ω_i and the metabin S_m is: $\sum_{j=1}^{p} x_{ijk_{jm}}z_{jk_{jm}}^{(\alpha)}$. Hence, each observation ω_i according to the metabin S_m among the $K_{Max} = Max_i m_i$ metabins can be represented by a supplementary individual denoted $\omega_i \times S_m$ of projection: $Proj^{\alpha}\left(\omega_i \times S_m/Z^{(\alpha)}\right) = K_{Max}\sum_{j=1}^{p} x_{ijk_{jm}}z_{jk_{jm}}^{(\alpha)}$ on the factors $Z^{(\alpha)}$. The product by M in the preceding formula leads to

$$Mean\left\{Proj^{\alpha}\left(\omega_i \times S_m/Z^{(\alpha)}\right)/j = 1, p\right\} =$$
$$= \sum_{m=1}^{K_{Max}} Proj^{\alpha}(\omega_i \times S_m/Z^{(\alpha)})/K_{Max} = Proj^{\alpha}\left(\omega_i/Z^{(\alpha)}\right)$$

and so it results that the projection of ω_i on $Z^{(\alpha)}$ is the mean center of the projections of the supplementary individuals denoted $\omega_i \times S_m$ for $k = 1, K_{Max}$ on $Z^{(\alpha)}$.

The table 6 expresses the Cartesian product between the individuals and the variables described by the bins. In this case, the coordinate of the unit $\omega_i \times Y_j$ associated to the initial individual ω_i and the variable Y_j is $\sum_{k=1}^{m_j} x_{ijk}z_{jk}^{(\alpha)}$. Hence, each individual ω_i according to the variable Y_j can be represented by p supplementary individuals denoted $\omega_i \times Y_j$ of projection: $Proj^{\alpha}\left(\omega_i \times Y_j/Z^{(\alpha)}\right) = p.\sum_{k=1}^{m_j} x_{ijk}z_{jk}^{(\alpha)}$ on the factors $Z^{(\alpha)}$. As in the preceding cases, it results that

$$Mean\left\{Proj^{\alpha}\left(\omega_i \times Y_j/Z^{(\alpha)}\right)/j = 1, p\right\}$$
$$= \sum_{j=1}^{p} \frac{Proj^{\alpha}\left(\omega_i \times Y_j/Z^{(\alpha)}\right)}{p} = Proj^{\alpha}\left(\omega_i/Z^{(\alpha)}\right).$$

Hence, the projection of ω_i on $Z^{(\alpha)}$ is the mean center of the projections of the supplementary individuals $\omega_i \times Y_j$ for $j = 1, p$ on $Z^{(\alpha)}$.

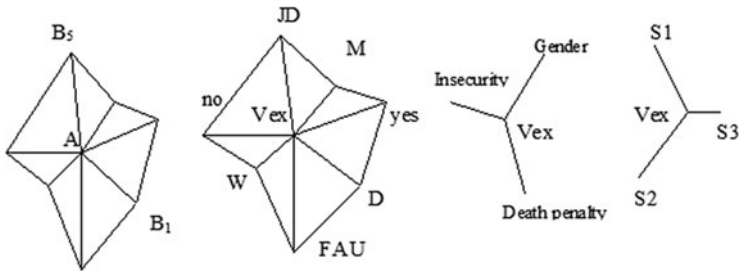


Fig. 4 Examples of "histogram stars" of bins or of variables where the sum of the length of the radius is 1 and the center is the projection of the individual is ω_i (Vex in the example)

5.4 "Histogram Stars" Associated to the Bins, the Metabins or the Variables, as Output of the PCA

We have seen in the preceding section that the representation of the individuals ω_i can be considered as the mean center of the M supplementary individuals $\times$ bins $\omega_i \times (k, j)$, of the K_{Max} supplementary individuals $\times$ metabins $\omega_i \times S_m$ or of the p supplementary individuals $\times$ variables $\omega_i \times Y_j$. These properties can be used in order to get a categorical histograms in output. In that way, we introduce the so called "histogram stars" (see Fig. 4). The centers of these stars are the projection of the n individuals ω_i on the factorial space and their vertices are the M projections of the $\omega_i \times bins$ or the p projections of the $\omega_i \times variables$ or the K_{Max} projections of the $\omega_i \times metabins$ such that the sum of the length of the radius of each star equal 1 (that is why such stars are called "histogram stars"). For example, in the case of the stars whose vertices represent the $\omega_i \times bins$ and the center represents the individual ω_i we can get them in the following way: we denote A_i for $i = 1, n$ the projection of the ω_i in the factorial space and B_k the projection of the M individual $\times$ bin in the same space. It is then easy to join A_i to the B_k and to divide the length of each radius by the sum of the length of all the A_i B_k when $k = 1, M$ in order to get a star as the ones shown in Fig. 4 where the sum of the length of the radius is equal to 1. In the Fig. 4 an individual $\times$ bin (resp. individual $\times$ variable or individual $\times$ metabin) is represented by just the name of the bin (resp. the name of the variable or the name of the metabin) as there is no ambiguity. Knowing the meaning of the factorial axis, the directions of the radius of these stars in the factorial space can enhance the interpretability of the categorical histogram PCA (standard or copular) for each chosen individual.

5.5 Representation of the Symbolic Variables in a Hypercube

We represent a variable Y_j on $F^{(\alpha)}$ by a weighted sum of the projection on each factor of the m_j bins of this variable. This weighted sum corresponds to the projection of the observations on the factor Z^{α} restricted to the bins of Y_j. Hence, $Y_j^{(\alpha)}$ is defined by:

$$Y_j^{(\alpha)} = \sum_{k=1}^{m_j} z_{jk}^{(\alpha)} Y_{jk}. \tag{1}$$

More precisely $Y_j^{(\alpha)}$ can be written $Y_j^{(\alpha)} = (Y_{1j}^{(\alpha)}, \ldots, Y_{nj}^{(\alpha)})^T$ where $Y_{ij}^{(\alpha)} = \sum_{k=1}^{m_j} x_{ijk} z_{jk}^{(\alpha)}$ for $i = 1, n$. We can then calculate the correlation between $Y_j^{(\alpha)}$ and $F^{(\alpha)}$ defined by:

$$Cor\left(Y_j^{(\alpha)}, F^{(\alpha)}\right) = Y_j^{(\alpha)} F^{(\alpha)} / \left(|Y_j^{(\alpha)}| \cdot |F^{(\alpha)}|\right) \tag{2}$$

Lemma *The projection of the bins on the principal component $F^{(\alpha)}$ is given by*

$$Y_{jk} F^{(\alpha)} = \lambda_\alpha z_{jk}^{(\alpha)} \tag{3}$$

Proof The projection of the bins on the principal component $F^{(\alpha)}$ is given by the matricial product $X' F^{(\alpha)}$. As $F^{(\alpha)} = XZ^\alpha$ we have $X' F^{(\alpha)} = X' X Z^\alpha = \lambda_\alpha Z^\alpha$ which implies $X' F^{(\alpha)} = \lambda_\alpha Z^\alpha$. As Y_{jk} and $z_{jk}^{(\alpha)}$ are respectively the kth row of X' and of Z^α , we get $Y_{jk} F^{(\alpha)} = \lambda_\alpha z_{jk}^{(\alpha)}$.

Proposition

$$Cor\left(Y_j^{(\alpha)}, F^{(\alpha)}\right) = \sqrt{\lambda_\alpha} \sum_{k=1}^{m_j} \left(z_{jk}^{(\alpha)}\right)^2 / \sqrt{\sum_{i=1}^{n} \left(Y_{ij}^{(\alpha)}\right)^2 / n} \tag{4}$$

Proof We have

$$Y_j^{(\alpha)} F^{(\alpha)} = \sum_{k=1}^{m_j} z_{jk}^{(\alpha)} Y_{jk} F^{(\alpha)} \quad \text{from (1)}$$

$$Y_j^{(\alpha)} F^{(\alpha)} = \sum_{k=1}^{m_j} z_{jk}^{(\alpha)} \left(\lambda_\alpha z_{jk}^{(\alpha)}\right) \quad \text{from (3)}$$

Therefore

$$Y_j^{(\alpha)} F^{(\alpha)} = \lambda_\alpha \sum_{k=1}^{m_j} \left(z_{jk}^{(\alpha)}\right)^2 \tag{5}$$

Knowing that

$$|F^{(\alpha)}| = \sqrt{\lambda_\alpha} \tag{6}$$

and

$$|Y_j^{(\alpha)}| = \sigma\left(Y_j^{(\alpha)}\right) \tag{7}$$

with $\sigma^2\left(Y_j^{(\alpha)}\right) = \sum_{i=1}^{n}\left(Y_{ij}^{(\alpha)}\right)^2/n$ where $Y_{ij}^{(\alpha)} = \sum_{k=1}^{m_j} x_{ijk} z_{jk}^{(\alpha)}$ due to the fact that $Y_{ij}^{(\alpha)}$ is centred as linear combination of centred variables. By replacing in (3) the formulas (5), (6) and (7) we get $Cor\left(Y_j^{(\alpha)}, F^{(\alpha)}\right) = \lambda_\alpha \sum_{k=1}^{m_j}\left(z_{jk}^{(\alpha)}\right)^2 / \left(\sigma\left(Y_j^{(\alpha)}\right)\sqrt{\lambda_\alpha}\right)$. From which results (3).

The "symbolic hypercube": From this proposition, it is possible to associate a "global" vector OG_j to each symbolic variable Y_j. This vector OG_j is defined by its coordinates on the principal components axes $F^{(\alpha)}$ which value is $Cor\left(Y_j^{(\alpha)}, F^{(\alpha)}\right)$. In other words $OG_j = \left(Cor\left(Y_1^{(\alpha)}, F^{(\alpha)}\right), \ldots, Cor\left(Y_p^{(\alpha)}, F^{(\alpha)}\right)\right)$. Therefore, each coordinate of this vector varies between -1 to 1 and therefore the vectors OG_j are inside the hypercube of p dimension whose projection on each plane defined by two axes $F^{(\alpha)}$, $F^{(\beta)}$ is a square with vertices of coordinates (1, 1), $(-1, 1)$, $(-1, -1)$, $(1, -1)$. This hypercube is called "symbolic hypercube" as it contains the symbolic variables.

6 Conclusion

We have presented a set of open directions of research for methods allowing the extraction of several kinds of information from a categorical histogram data table. We have shown that standard PCA can be applied but loose the probabilistic aspect of these kind of data due to its underlying assumption of independency between the bins. The copular approach solves this question by allowing the estimation of other models closer from the data. Copular approach can be applied to ordinal histogram data and has to be extended to standard numerical histograms and more generally to distributions of functional data. We have also given ways for the representation of individuals, individuals x variables, individuals $\times$ metabins at the vertices of what has been called "histogram stars". For the representation of the categorical histogram variables, we have provided new theoretical results allowing their representation in a hypercube of the correlation PCA space. This work open doors for solving the difficult question of PCA on nominal histogram data but much remain to be done to improve in practice the proposed directions.

Annex

Definition of a k-copula (Schweizer and Sklar [14], Nelsen [12]): A k-copula is a function C from $[0, 1]^k$ to $[0, 1]$ with the following properties:

1. for every u in $[0, 1]^k$, $C(u) = 0$ if at least one coordinate of u is 0;
2. if all coordinate of u are 1 except u^* then $C(u) = u^*$;
3. the number assigned by C to each hyper-cube $[a_1, a_2] \times [b_1, b_2] \times \ldots \times [z_k, z_k]$ is non negative.

For example, in two dimensions ($k = 2$), the third condition gives: $C(a_2, b_2) - C(a_2, b_1) - C(a_1, b_2) + C(a_1, b_1) \geq 0$. In the following, we denote $RanG$ as the range of the mapping G. Sklar [15] gave the following theorem:

Theorem *Let H be a k-dimensional distribution with marginal distributions $G_1, \ldots, G_k$. Then there exists a k-copula C such that for all $(x_1, \ldots, x_k) \in [0, 1]^n$,*

$$H(x_1, \ldots, x_k) = C(G_1(x_1), \ldots, G_k(x_k)). \quad (1)$$

Moreover, if $G_1, \ldots, G_k$ are continuous, then C is unique; otherwise C is uniquely determined on $RanG_1 \times \ldots \times RanG_k$. Conversely, if $G_1, \ldots, G_k$ are distribution functions and C is a copula, the function H defined by (1) is a k-dimensional distribution function with marginal distributions $G_1, \ldots, G_k$.

Parametric families of copulas: the most simple copulas denoted M, W and Π are $M(u, v) = min(u, v)$, $\Pi(u, v) = uv$ and $W(u, v) = max(u + v - 1, 0)$. These copulas are special cases of some parametric families of copulas as the followings: $C_b(u, v) = max([u^{-b} + v^{-b} - 1]^{-1/b}, 0)$ discussed by Clayton [4] has the following special cases: $C_{-1} = W$, $C_0 = \Pi$, $C_\infty = M$. Frank [6] has defined $C_b(u, v) = -1/b \ ln\left(1 + \left(e^{-bu} - 1\right)\left(e^{-bv} - 1\right) / \left(e^{-b} - 1\right)\right)$ which has the following special cases: $C_{-\infty} = W$, $C_0 = \Pi$, $C_\infty = M$. We denote τ the Kendall tau between R_{ij} and $R_{ij'}$.The Clayton family is defined by $Cop_\theta(u, v) = max\left(\left(u^{-\theta} + v^{-\theta} - 1\right)^{-1/\theta}, 0\right)$ where $\theta \in [-1, 1]/0$ and $\theta = 2\tau/(1 - \tau)$. Many other copulas families are defined in Nelsen [12].

References

1. Aitchison, J.: The statistical analysis of compositional data. Chapman & Hall, London; Biometrika **70**(1), 57–65 (1996)
2. Billard, L., Diday, E.: Symbolic data analysis, Wiley, Chichester (2006)
3. Bock, H.H., Diday, E. (eds.): Analysis of symbolic data, Springer, Berlin (2000)
4. Clayton, D.G.: A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. Biometrika **65**, 141–152 (1978)
5. Diday, E., Noirhomme-Fraiture, M. (eds.): Symbolic data analysis and the SODAS software, pp. 279–311, Wiley, Chichester (2008)
6. Frank, M.J.: On the simultaneous associativity of $F(x, y)$ and $x+y–F(x, y)$. Aequationes Math. **19**, 194–226 (1979)
7. Ichino, M.: Symbolic principal component analysis based on the nested covering. In: ISI2007, Lisbon, 2007
8. Ichino, M.: Symbolic PCA for histogram-valued data. In: Proceedings IASC, Yokohama, Japan, 5–8 Dec 2008

9. Lebart, L., Morineau, A., Piron, M.: Statistique exploratoire multidimensionnelle. Dunod Editeur, Paris (1995)
10. Makosso Kallyth, S., Diday, E.: Analyse en composantes principales de variables symboliques de types histogrammes. In: D'Aubigny, G. (ed.) Proceedings SFC. IMAG, Grenoble, France (Sept 2009)
11. Nagabhsushan, P., Kumar, P.: Principal component analysis of histogram data. In: Liu, D. et al. (eds.) ISNN 2007, Part II, LNCS 4492, pp. 1012–1021. Springer, Berlin, Heidelberg (2007)
12. Nelsen, R.B.: An Introduction to copulas. In: Lecture Notes in Statistics, Springer, NewYork, NY (1998)
13. Rodriguez, O., Diday E., Winsberg, S.: Generalization of the principal component analysis to histogram data. In: Workshop on Symbolic Data Analysis of the 4th European Conference on Principles and Practice of Knowledge Discovery in Data Bases, 12–16 Sept 2000, Lyon (2001)
14. Schweizer, B., Sklar, A.: Probabilistic metric spaces, Elsevier, North-Holland, NewYork (1983)
15. Sklar, A.: Fonction de répartition à n dimensions et leurs marges. Publ. Inst. Stat. Univ. Paris **8**, 229–231 (1959)

Factorial Conjoint Analysis Based Methodologies

Giuseppe Giordano, Carlo Natale Lauro, and Germana Scepi

Abstract Aim of this paper is to underline the main contributions in the context of Factorial Conjoint Analysis. The integration of Conjoint Analysis with the exploratory tools of Multidimensional Data Analysis is the basis of different research strategies, proposed by the authors, combining the common estimation method with its geometrical representation. Here we present a systematic and unitary review of some of these methodologies by taking into account their contribution to several open ended problems.

1 Introduction

Conjoint Analysis [15, 16] is one of the most popular statistical technique used in Marketing to elicit preference functions at both individual and aggregate level. Conjoint Analysis (CA) is a methodology based on several steps starting from designing the experiment, collecting data, estimating the model and, finally, using the results for market segmentation or product positioning.

Since the early 1970s, this technique has known an even wider diffusion in different applicative fields, ranging from Trading to Health, from Agriculture to Food Industry, among others. One of the most recent field is Regulatory Impact Analysis where the aim is to set the *ideal* regulation among alternative policies [24].

In the 1998, Lauro et al. [18] proposed the use of Principal Component Analysis in order to manage dependent and explanatory variables in Conjoint Analysis. In such approach, the traditional interpretative tools of multidimensional techniques enhance the classical CA results. The underlying thought is that individual part-worth coefficients derived for each respondents can be aggregated in a set of common latent utility models, arranged in decreasing importance with respect to their explicative power (see Sect. 3).

Starting from this approach, different methodologies have been then developed and applied in the framework of Multidimensional Data Analysis. Aim of this paper

G. Giordano (✉)
University of Salerno, 84084 Fisciano, Salerno, Italy,
e-mail: ggiordan@unisa.it

B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-13312-1_2,

is to present a systematic and unitary review of these methodologies by taking into account their contribution to several open ended problems: (i) to obtain homogeneous groups of respondents (ii) to take into account information on the respondents not included in the Conjoint models, and (iii) to consider multiple criteria as response variables.

The paper is structured as follows: in Sect. 2 we refer to the Metric Conjoint Analysis model as the baseline of all successive methods. In Sect. 3 it is shown how the data structure considered in Metric Conjoint Analysis can be analyzed in the framework of exploratory multidimensional data analysis and how to read and interpret the factorial maps as preference map. In Sect. 4, it is considered the opportunity to enrich the original data structure with information about respondents. This new data set is introduced and analyzed by a new specification of the Conjoint model considering the preference system influenced by both the stimuli features and the consumer characteristics.

Thus, in order to derive ex-post cluster of homogeneous set of respondents, a peculiar approach is discussed in Sect. 5. It starts from the results of the factorial decomposition in order to derive global and local utility models. Finally (in Sect. 6), we address the problem of a multi-criteria approach to Conjoint Analysis by introducing a data structure allowing to take into account multiple set of response variables. Some conclusions and future directions are in the Sect. 7.

2 The Metric Approach to Conjoint Analysis

As starting point, we look at the metric approach to Conjoint Analysis. This allows us to consider a well defined data structure that can be analyzed with a multivariate multiple regression model, where the response variable is measured at interval scale and OLS estimation method is applied.

For instance, we consider the role played by two sets of variables: the dependent variables in the matrix $\mathbf{Y}$ $(N \times J)$, and the explicative ones in the design matrix $\mathbf{X}$ of size $N \times (K - k)$, where N is the set of Stimuli, J is the number of judges, i.e. the preference responses, and k is the number of experimental factors expanded in K attribute-levels. Let us notice that one dummy category has been dropped out for each attribute to obtain a full rank matrix design.

The Metric Conjoint Analysis model is written as the following multivariate multiple regression model:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \tag{1}$$

where $\mathbf{B}$ is the $(K - k) \times J$ matrix of individual part-worth coefficients and $\mathbf{E}$ is the $(N \times J)$ matrix of error terms for the set of J individual regression models.

Indeed, the simultaneous computation of the elements of the coefficient matrix $\mathbf{B}$ yields the same results as a set of J separate multiple regression models, since the relationships within the multiple responses are not involved in the ordinary least squares method. The OLS is seen here as a decompositive technique because

the classical assumptions on the errors are disregarded in Conjoint Analysis. Typically, some *holdout* runs are used to assess the internal validity of the model. Since researchers often deal with complex stimuli, where a large number of attributes and levels are involved, the use of saturated models is often necessary as a screening study.

Focusing on the quantitative nature of the response variables (preference rating), on the qualitative featuring of the design matrix (dummy variables) and because of the simple reading of the part-worth coefficients computed as average effects, the metric approach to Conjoint Analysis is the most used one. This model is at the basis of the methodologies proposed in the following sections.

3 The Factorial Conjoint Analysis

The Multidimensional Approach to Conjoint Analysis aims at improving the interpretation of the traditional results of this technique by proposing a new reading in the context of Exploratory Data Analysis. The main advantage is to obtain a graphical visualization of the relationships between the preference judgments (dependent variables) and the attribute-levels (independent variables) represented onto a common space.

Different techniques have been proposed in order to take into account the dimension reduction aspect of the model stated in Eq. (1). Among others, we mention the Reduced-Rank Regression Model [1, 17]; the Principal Component of Instrumental Variables [20]; the Simultaneous Linear Prediction Modeling [10]; the Redundancy Analysis [25] and the Principal Component Analysis on a Reference Subspace [5, 6]. The peculiarity of all these techniques is the possibility to link the computational aspects of the regression coefficients with the descriptive and interpretative tools of principal component or canonical variates.

Here we refer to the Principal Component Analysis on a Reference Subspace (PCAR). We consider the asymmetric role played by the two sets of variables (preferences and attributes) involved in multiattribute preference data. Note that in multidimensional data analysis asymmetry refers to the different role played by two or more set of variables when we observe a particular phenomenon. In this context, we highlight the dependence relation between the set of the J preference response variables and the set of the $(K - k)$ attribute–levels described in the design matrix. This technique allows to summarize the multivariate set of preference response variables by performing a Principal Component Analysis of the matrix $\mathbf{XB}$ – stated in model (1) – and equivalent to:

$$\hat{\mathbf{Y}} \equiv \mathbf{XB} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{2}$$

The individual part-worth coefficients are aggregated by means of a suitable weighting system (the PCAR coefficients) reflecting the preference variability:

$$\mathbf{B}'\mathbf{X}'\mathbf{X}\mathbf{B}\mathbf{u}_\alpha = \lambda_\alpha \mathbf{u}_\alpha \qquad \alpha=1,\ldots,(K-k) \tag{3}$$

It is worth noting, by comparing expression (1) and (3), that the criterion optimized with PCAR (i.e. preference variance accounted by attribute-levels) is fully consistent with the metric Conjoint Analysis data structure. Namely, we define this method as Factorial Conjoint Analysis (FCA).

The PCAR geometrical interpretation allows to enrich even more Conjoint Analysis by joint plots of attribute-levels, judges and stimuli on the first two or three factorial axes. Additional information on judges (e.g. a priori cluster or social-demographic characteristics) can also be shown on the plot.

Traditional interpreting tools of Conjoint Analysis can be read in the context of multidimensional data analysis too. For instance, the relative importance of each attribute are derived by looking at the range of the attribute-level coordinates on each factorial axis. Each factorial axis is a synthesis of the preference variables. They describe the preference of a homogenous subset of respondents towards the attribute levels. The first factorial axis determine the maximum agreement system within judges while the successive ones establish alternative preference patterns of judges subsets.

Considering the expression (2), the principal axes of inertia are obtained as solution of the following characteristic equation under orthonormality constraints:

$$\hat{\mathbf{Y}}'\hat{\mathbf{Y}}\mathbf{u}_\alpha = \lambda_\alpha \mathbf{u}_\alpha \qquad \mathbf{u}'_i\mathbf{u}_i=1; \quad \mathbf{u}'_i\mathbf{u}_j=0 \quad \{i,j\}\in\alpha=1,\ldots,(K-k); \tag{4}$$

which is a Principal Component Analysis of the matrix $\hat{\mathbf{Y}}$.

The eigenvectors $\mathbf{u}_\alpha$ are the weights for the J respondents in the aggregated preference model:

$$\bar{\hat{\mathbf{Y}}}_\alpha = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\mathbf{u}_\alpha = \mathbf{X}\mathbf{B}\mathbf{u}_\alpha \tag{5}$$

Since there are at most $(K-k)$ different weighting systems with decreasing order of importance, we refer to $\alpha = (1, 2)$ as the principal judgment system and define the first factorial plan as a *Preference Scenario*. The (5) is used in computing the coordinates of the N stimuli. The holdouts stimuli can be represented as supplementary points Y^+:

$$Coor(Y^+) = X^+\mathbf{B}\mathbf{u}_\alpha \tag{6}$$

where X^+ are the $\mathbf{X}$ rows containing the attribute levels combinations describing any new products. The coordinates of the attribute-levels are:

$$Coor(X) = (\mathbf{X}'\mathbf{X})^{-1/2}\mathbf{X}'\mathbf{Y}\mathbf{u}_\alpha \tag{7}$$

The level coordinates in Eq. (7) are computed as linear combination of the individual part-worth coefficients and assuming the relation between judge and factorial axis as weighting system (i.e. the vector $\mathbf{u}$), they represent different synthesis of the

estimated part-worth coefficients. In this way, we obtain different synthesis of the individual estimates instead of the unique average which is traditionally considered.

The coordinates of the J respondents are:

$$Coor(Y) = \sqrt{\lambda_\alpha}\mathbf{u}_\alpha \tag{8}$$

which give the directions where pointing out the individual preference models.

The most important feature of this approach is the possibility to synthesize the part-worth coefficients in an optimal way according to hierarchical patterns of preferences.

The advantage of this technique with respect to similar approach, carried out by means of Multidimensional Scaling techniques, is the possibility to recover and interpret the different role of all the *objects* involved in the analysis (e.g. Stimuli, Attributes, Levels and Preference Scores). Furthermore, *holdout* stimuli not involved in the analysis can be represented as supplementary points on the factorial plan which is here interpreted as a *Preference Map*.

4 The FCA with Two Informative Structures

In the multivariate regression model introduced in Eq. (1) we have considered two groups of variables: a set of dependent variables (in $\mathbf{Y}$), judges' preference, described by a set of explicative variables (in $\mathbf{X}$) which are the dummy coded attributes of the stimuli.

In Marketing, for example, with the aim of defining a peculiar market strategy, consumers can be a-priori classified on the basis of several socio-demographical characteristics. Starting from this point of view, it is possible to undertake the Taguchi's categorization between controllable versus noise factors proposed in Design of Experiment for Total Quality Control and introduce it in Preference Data Analysis [12]. In particular, the set of a-priori information on judges can be considered as *external* factors and the attribute-levels describing different stimuli as controllable or *internal* factors. Therefore, it is possible to introduce the data matrix $\mathbf{Z}$ of size $(H - h) \times J$, holding socio-demographical characteristics (h nominal variables expanded in full disjunctive binary coding) observed on J judges.

We refer to the design matrix $\mathbf{X}$ as *internal* informative structure or *Inner Array* in Taguchi's notation. while the matrix $\mathbf{Z}$ can be seen as *external* informative structure or *Outer Array*. With the aim of studying the relationships between the two different informative structures, the influence of these two kinds of information on the response variables and, finally, the relationships within each data structure, an extension of the Factorial Conjoint Analysis approach has been applied.

In particular, the two data matrices can be regarded as two different sets of explicative variables in two separated multivariate regression models. The first one is the model (1) above defined and the second one is defined by considering the respondents as statistical units in the model:

$$\mathbf{Y}' = \mathbf{Z}'\mathbf{D} + \mathbf{F} \tag{9}$$

where $\mathbf{D}$ is the $(H - h) \times J$ matrix of coefficients and $\mathbf{F}$ is the $(J \times N)$ matrix of error terms. In order to relate the information of the two designs ($\mathbf{X}$ and $\mathbf{Z}'$ and their own effects on the values in $\mathbf{Y}$, the coefficients matrices $\mathbf{B}$ and $\mathbf{D}$ have been used in the models (10) and (11) which share common solutions in the matrix $\boldsymbol{\Theta}$.

$$\hat{\mathbf{B}} = \mathbf{Z}'\Theta + \mathbf{V} \tag{10}$$

$$\hat{\mathbf{D}} = \mathbf{X}\Theta' + \mathbf{W} \tag{11}$$

where $\mathbf{V}$ and $\mathbf{W}$ are the corresponding matrices of error terms. The generic elements of [(H−h)×(K−k)] can be regarded as a coefficient showing the relationship between the two sets of explicative factors. The common OLS solution for Θ is:

$$\hat{\Theta} = (\mathbf{Z}\mathbf{Z}')^{-1}\mathbf{Z}\mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}. \tag{12}$$

In order to synthesise and represent the information in Θ, the Singular Value Decomposition SVD) with respect to two different metrics [13] allows to produce a factorial plan where to show the relationships between the users' characteristics and the service features.

This approach allows us to enrich the results of Conjoint Analysis by considering the elements of Θ as inter-reference coefficients while the elements of $\mathbf{B}$ and $\mathbf{D}$ can be regarded as intra-reference coefficients.

The coefficients in $\mathbf{B}$ are useful to answer to questions as: *What happens if we substitute an attribute level with another one?*

The coefficients in $\mathbf{D}$ answer to: *How a category of respondents value a product/service with respect to another category?* The Θ coefficients help the researchers to answer at question as: *What is the effect of changing the attribute level when we target a peculiar category of respondents?*

In this way, for example, we may simulate potential market segments characterised by both consumers and products characteristics.

5 Cluster Based Factorial Conjoint Analysis

The results obtained by Factorial Conjoint Analysis (see Sect. 3) give an aggregate model derived by the total variability within preference judgments. Let us note that judges are represented as variables (judge-vectors) in the FCA subspace.

Starting from FCA results, a strategy which alternates steps of factorial analyses and clustering procedures is proposed in Lauro et al. [19].

With the aim of defining utility models for homogeneous group of judges, we used a method of variables clustering that split the set of judgements into hierarchical clusters. In a Customer Satisfaction strategy, for example, the obtained classes

can be considered as market segments and specialized products are offered to these segments in order to maximize customer satisfaction.

Therefore, for each class a local utility model is derived. Thus, in order to define an aggregate model coherent with the local models and, at the same time, reflecting the preferences of most judges, a weighted PCA is carried out. This analysis aims to synthesize the local preference models in a single model by considering both the number of units in a class and their variability. In this way, the final synthesis put emphasis on more homogenous and larger classes. As a result, this aggregate model is different from the initial aggregate model furnished by the first axis of the FCA; it gets the information from the classes for defining an ideal scenario representative of the preferences of most respondents.

Indeed, each factorial axis obtained through Eq. (3) is a synthesis of the whole set of J preference models. According to standard PCA interpretation rules, vector-variables pointing in the same direction are highly correlated and represent similar preference. Starting from results shown in Eq. (8), a variable clustering (i.e. the VARCLUS procedure of the SAS/Stat system) is performed, so that the individual judgments are aggregated to form homogenous classes. For each class, a local utility model is derived and a standardized scoring coefficient is assigned to each variable to determine the membership to the class. In this step the number of members in each class and a measure of variability explained by each class is also retained. The number of classes C is chosen by exploring the tree structure (dendrogram) or can be set ex ante by the researcher. In a further step we use this information to derive an optimal aggregated utility function, as a synthesis of the local models. The aggregated model is different from the initial PCAR results because it takes into account the relationships among clusters instead of individual judgments. At this aim it is defined the matrix $\mathbf{S}$ $(J \times C)$ holding the standardized scoring coefficients for each judge, and the diagonal weighing matrix $\mathbf{W}(C \times C)$ of generic term defined as:

$$w_c = \frac{j_c/\sigma_c}{\sum_c j_c/\sigma_c}; \quad c=1,\ldots,C; \quad 0 \le w_c \le 1; \quad \sum\nolimits_c w_c = 1 \tag{13}$$

where j_c is th number of judges in the cth class and σ_c is the variability explained by the cth class. By considering the initial coefficients $\mathbf{B}$, a weighted Principal Component Analysis is defined by the following eigen-equation:

$$\mathbf{WS}'\mathbf{B}'\mathbf{BSv}_\alpha = \upsilon_\alpha \mathbf{v}_\alpha \tag{14}$$

where the υ_α are the eigenvalues associated to the corresponding eigenvectors v_α. The matrix product $\mathbf{BS}$ $[(K-k) \times C]$ retrieves the importance of each part-worth coefficients in the C segments.

The first principal component obtained by the (14) represents a synthesis of the local preference models. So the new aggregate model can just defined by considering this component. We highlight that the weighing matrix $\mathbf{W}$ used in Eq. (14)

allows to stress on local models with a large number of units (all judges sharing the same model) and allows to give importance to more homogeneous clusters (i.e. market segments).

Therefore, the direction of the new principal component depends on (i) the correlations between the local models; (ii) the size and (iii) the variability within the clusters.

The main results of this strategy are:

- the definition of local preference models;
- the definition of a synthesis model taking into account the local preference models;
- the graphical representations of preferences of both the local and the aggregate models as useful tools for interpreting results.

The local preference models allow to represent utility functions for subset of respondents. For example, in marketing, they could be very useful to establish a marketing specialization strategy. The global preference model is complementary to the average model and to the principal component model. The more homogeneous is the market as a whole, the more the syntheses will tend to the average model. In presence of strong variability among the judgements the principal component model improve the average model. Whereas there exists subsets of consumers that define market segments, the global model is the best suited for a covering market strategy. The graphical representation allows to better visualize the results of the Conjoint experiment and simulation study for product positioning and market simulation.

We warn that internal validity should always achieved by cross-validation techniques, and more important in application fields, the actual use of the methods should produce feedback information to assess external validity and reliability of results.

6 Multi Criteria Factorial Conjoint Analysis

The developments of Conjoint Analysis discussed above are defined on the concept of utility function and applied in the context of Marketing Research, Customer Satisfaction and Customer Relationship Management. Recently, the concept of function of value has taken new meanings. Since one cannot measure utility directly, and attempts to derive it based on preferences (Conjoint Analysis relies on the *Neumann-Morgenstern* theory) could not work because the idea of utility is ambiguous in Social Choice theory where you are speaking about what is useful to society in general. Anyway, *Which are society values?* and *What do you value for society?* In general, different criteria could be taken into account when evaluating some concepts from the socio-political point of view. In this view, we think that Conjoint Analysis can be easily adapted to understand the importance (or value) of different attribute-levels in defining a new product/service, as well as a new policy or politics [24]. All we need, is a value system, different from *Utility*, able to describe

the new concept. Some example are the *Efficacy* or the *Sustainability* in a broader sense. The definition of such analytical functions and their estimates through the use of the Conjoint Analysis approach, will provide a new tool for comparing and evaluate the gaps between what is expected and what is possible to get.

We plan and administer a questionnaire for collecting simultaneously opinions of same judges on a set of stimuli on the basis of m different criteria such as *expected benefits*, *expected utility*, *strategic priority*, and so on.

With this aim, we extend the metric model of Conjoint Analysis (1) by introducing several response matrices:

$$\begin{cases} \mathbf{Y}_1 = \mathbf{X}\mathbf{B}_1 + \mathbf{E}_1 \\ \quad \vdots \\ \mathbf{Y}_m = \mathbf{X}\mathbf{B}_m + \mathbf{E}_m \end{cases} \tag{15}$$

where $\mathbf{B}_m$ is the coefficient matrix related to the mth criterion and $\mathbf{E}_m$ is the corresponding error matrix. Therefore, m sets of OLS part-worth coefficients have been calculated by considering each single criterion separately. Let us notice that the Design Matrix $\mathbf{X}$ is the same in the m criterion.

We are interested in define a similarity measure among the J judges with respect to the m criteria simultaneously. A straightforward approach is to carry out the Factorial Conjoint Analysis on a given criterion and then project (as supplementary points) on the obtained subspace the others $m-1$ criteria. However, this choice has several issues because of the subjectivity of the reference criterion and the absence of a reference subspace where all criteria play an equal (as well as weighted) role.

Different methods allow to face with these issues. They aim at obtaining a synthesis of the multiple criteria directly on a factorial plan. From the others, we refer to STATIS [9], Principal Matrices Analysis [4, 21] and Multiple Factorial Analysis [8] and, from a non-symmetrical point of view, to an extension of PCAR [5] and to a non-symmetrical version of Principal Matrices Analysis [3], which considers the case of multiple observational designs.

The Multi Criteria Factorial Conjoint Analysis deals with a peculiar data structure where the design matrix is the same in the different occasions while the response matrix changes. Therefore, we propose to apply the MFA to the coefficient matrices $\mathbf{B}_i (i = 1, \ldots, m)$ and (according to Eq. 2) to interpret it in the frame of a non-symmetrical data analysis.

Multiple factor analysis analyzes observations described by several "blocks" or sets of variables. MFA seeks the common structures present in all of these sets. MFA is performed in two steps. First a principal component analysis (PCA) is performed on each data set which is then "normalized" by dividing all its elements by the square root of the first eigenvalue obtained from each PCA. Second, the normalized data sets are merged to form a unique matrix and a global PCA is performed on this matrix. The individual data sets are then projected onto the global space to analyze communalities and discrepancies.

In analogy with MFA, we carry out m PCA's, one for each separated criterion, and the first eigenvalue is retrieved. So we normalize each $\mathbf{B}_m (i = 1, \ldots, m)$ and juxtapose them in order to obtain a unique matrix. A global PCA is performed on this matrix. In this way a synthesis of the coefficients related to all criteria is achieved. On this common plan, we can compare the different criteria and we can project the judges for analyzing their differences and similarities respect to the different criteria. The relationships among the different criteria and between the criteria and the global solution can be analyzed by computing the partial inertia of each analysis for each dimension of the global analysis. In this way we are able to understand the importance of each criterion in the definition of the global solution and we can define the ideal combination of attribute-levels (product, service, policy and so on) by selecting the levels with the larger coordinates on the global plan.

7 Some Conclusions

Different other contributions to Factorial Conjoint Analysis have been proposed by the authors with different aims and from different perspectives [2, 7, 11, 23]. A further development consists in searching a new distance for comparing metric Conjoint Analysis models. In particular, Romano et al. [22] define an inter-models distance which takes into account both the analytical structure of the model (through coefficient deviations) and the information about the model fitting (through the difference between the adjusted R^2 related to each pair of models). The so defined *Model Distance* is parameterized by a trimmer value that allows to take into account the extent to which the model fitting enter in the definition of the distance. This new metric allows to cluster judges in terms of individual models by taking into account both the information on the structural component of the model and on the error term. As Green et al. [14] say:

> [...] despite its maturity, conjoint analysis is still far from stagnant, because the methods deal with the pervasive problem of buyer preferences and choices.

In this point of view, we think that new methodological development and applications can be performed in Factorial Conjoint Analysis context. In particular, it will be interesting to investigate the possibility to extend the proposed methods to different structures of data, such as for example, interval value data.

References

1. Anderson, T.W.: Estimating linear restrictions on regression coefficients for multivariate normal distributions. Ann. Math. Stat. **22**, 327–351 (1951)
2. Balbi, S., Giordano, G.: A factorial technique for analyzing textual data with external information. In: Borra, S., Rocci, R., Vichi, M., Schader, M. (eds.) Studies in Classification Data Analysis and Knowledge Organization, pp. 166–176. Springer, Heidelberg (2000)
3. Balbi, S., Lauro, N.C., Scepi, G.: A multiway data analysis technique for comparing surveys. Methodologica **3**, 79–91 (1994)

4. D'Alessio, G.: Multistep principal component analysis in the study of panel data. In: Coppi, R., Bolasco, S. (eds.) Multiway Data Analysis, pp.375–381. North-Holland, Amsterdam (1989)
5. D'Ambra, L., Lauro, N.C.: Analisi in Componenti Principali in rapporto ad un sottospazio di riferimento. Stat. Appl. **15**, 51–67 (1982)
6. D'Ambra, L., Lauro, C.: Non-symmetrical exploratory data analysis. Stat. Appl. **4**, 511–529 (1992)
7. Davino, C., Giordano, G., Lauro, N.C.: Analisi dei dati e reti neurali per la Conjoint Analysis. In: Atti del Convegno: La Statistica per le imprese, pp. 299–306. SIS, Torino (1997)
8. Escofier, B., Pagés, J.: Multiple factor analysis. Comput. Stat. Data Anal. **18**, 121–140 (1990)
9. Escoufier, Y.: Three-mode data analysis: The STATIS method. In: Fichet, B., Lauro, C. (eds.) Methods for Multidimensional Data Analysis, pp.259–272. ECAS, Napoli (1987)
10. Fortier, J.J.: Simultaneous linear prediction. Psychometrika **31**, 369–381 (1966)
11. Gettler Summa, M., Giordano, G., Verde, R.: Symbolic interpretation in a clustering strategy on multiattribute preference data. Stat. Appl. **12**(4), 473–495 (2000)
12. Giordano, G., Scepi, G.: Different informative structures for quality design. J. Ital. Stat. Soc. **8**(2–3), 139–149 (1999)
13. Gower, J.C., Hand, D.J.: Biplots. Chapman & Hall, London (1996)
14. Green, P.E., Krieger, A.M., Wind, Y.: Thirty years of conjoint analysis: Reflections and prospects. Interfaces **31**(3), 56–73 (2001)
15. Green, P.E., Srinivasan, V.: Conjoint analysis in consumer research: issues and outlook. J. Consum. Res. **5**(2), 103–123 (1978)
16. Green, P.E., Srinivasan, V.: Conjoint analysis in marketing: New developments with implication for research and practice. J. Mark. **25**, 319 (1990)
17. Izenman, A.J.: Reduced rank regression for the multivariate linear model. J. Multivar. Anal. **5**, 248–264 (1975)
18. Lauro, N.C., Giordano, G., Verde, R.: A multidimensional approach to conjoint analysis. Appl. Stoch. Model Data Anal. **14**, 265–274, Wiley (1998)
19. Lauro, N.C., Scepi, G., Giordano, G.: Cluster based conjoint analysis. In: Proceedings of 6th International Conference on Social Science Methodology, RC Logic & 33 Methodology, Amsterdam 17–20 Aug 2004
20. Rao, C.R.: The use and the interpretation of principal component analysis in applied research. Sankhya A. **26**, 329–358 (1964)
21. Rizzi, A.: On the synthesis of the three-way data matrices. In: Coppi, R., Bolasco, S. (eds.) Multiway Data Analysis, pp.143–154. North-Holland, Amsterdam (1988)
22. Romano, E., Giordano, G. Lauro, N.C.: An inter-models distance for clustering utility functions. Stat. Appl. **17**, 2 (2006)
23. Scepi, G., Giordano, G., Ramunno, I.: A new technique for dealing with complex stimuli in conjoint analysis. Stat. Appl. **18**(1), 105–118, Rocco Curto Editore (2006)
24. Scepi, G., Lauro, N.C., Giordano, G.: The ex-ante evaluation of regulatory impact by conjoint analysis: some developments. In: Book of Short Abstracts of the 7th International Conference on Social Science Methodology RC33 Logic and Methodology in Sociology, Jovene Editore (2008)
25. van den Wollenberg, A.L.: Redundancy analysis. An alternative for canonical correlation analysis. Psychometrika **42**, 207–219 (1977)

Ordering and Scaling Objects in Multivariate Data Under Nonlinear Transformations of Variables

Jacqueline J. Meulman, Lawrence J. Hubert, and Phipps Arabie

Abstract An integrated iterative method is presented for the optimal ordering and scaling of objects in multivariate data, where the variables themselves may be transformed in the process of optimizing the objective function. Given an ordering of objects, optimal transformation of variables is guaranteed by the combined use of majorization (a particular (sub)gradient optimization method) and projection methods. The optimal sequencing is a combinatorial task and should not be carried out by applying standard optimization techniques based on gradients, because these are known to result in severe problems of local optima. Instead, a combinatorial data analysis strategy is adopted that amounts to a cyclic application of a number of local operations. A crucial objective for the overall method is the graphical display of the results, which is implemented by spacing the object points optimally over a one-dimensional continuum. An indication is given for how the overall process converges to a (possibly local) optimum. As an illustration, the method is applied to the analysis of a published observational data set.

1 Introduction

In this paper we consider a set of iterative methods that are integrated to perform a particular data analysis task. The available data will typically consist of nonnumerical variables, e.g., ordinal variables that provide an ordering for the objects, or nominal variables that group the objects into a limited number of classes. In either case, the distances between distinct pairs of the objects are unknown and have to be recovered as part of the overall optimization process. For such data, we require the results of the analysis to be invariant under one-to-one nonlinear transformations of the variables.

We focus on the analysis of an $n \times m$ multivariate data matrix $\mathbf{Q} = \{q_{ij}\}_{1 \leq i \leq n, 1 \leq j \leq m}$, where n denotes the number of objects in the rows of the matrix and m denotes the number of variables in the columns. Later, the columns of $\mathbf{Q}$

J.J. Meulman (✉)
Department of Mathematics, Leiden University, Leiden, The Netherlands,
e-mail: jmeulman@math.leidenuniv.nl

B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-13312-1_3,

will themselves be assumed constructed by transformations on the columns of an originally given $n \times m$ data matrix $\mathbf{Z}$, but for the moment we phrase our presentation solely in terms of the data matrix $\mathbf{Q}$. The relationship between the objects is defined by Euclidean distances between the rows of the data matrix $\mathbf{Q}$, and these are collected into an $n \times n$ proximity matrix $\mathbf{P} = \{p_{ik}\}_{1 \leq i,k \leq n}$, having entries p_{ik} ($= p_{ki} \geq 0$, and $p_{ii} = 0$). Thus, p_{ik} represents the proximity between the objects O_i and O_k, where a large value corresponds to a small similarity and a large distance, and can be given explicitly as:

$$p_{ik} \equiv \sqrt{d_{ik}^2(\mathbf{Q})} \equiv \sqrt{(\mathbf{e}_i - \mathbf{e}_k)'\mathbf{Q}\mathbf{Q}'(\mathbf{e}_i - \mathbf{e}_k)}\,, \tag{1}$$

with $\mathbf{e}_i$ denoting the ith column vector of the $n \times n$ identity matrix $\mathbf{I}$, and $d_{ik}^2(\mathbf{Q})$ the squared Euclidean distance between O_i and O_k. The basic data analytic problem is to represent the (high-dimensional) proximities between objects by distances between object points in a low-dimensional representation space $\mathbf{X}$ of order $n \times s$, for some suitably chosen value of s.

There are numerous approaches that could be considered if the representation space $\mathbf{X}$ has $s \geq 2$ dimensions. (We mention these very briefly; details can be found in [22].) One of these is the classical approach usually associated with Torgerson [30] and Gower [9], which is an eigenvalue technique based on Young and Householder [37]. Here a scalar product matrix $\mathbf{R}$, constructed by taking $-\frac{1}{2}$ the double-centered squared proximities, is approximated by another scalar product matrix of lower rank $\mathbf{X}\mathbf{X}'$. To optimize the approximation by using distances $D(\mathbf{X})$, and not scalar products $\mathbf{X}\mathbf{X}'$, gradient [16, 17, 23, 24], subgradient [5, 10], and majorization methods [6] have been developed to minimize iteratively the objective function

$$STRESS(\mathbf{X}) = \|\mathbf{P} - D(\mathbf{X})\|^2, \tag{2}$$

over $\mathbf{X}$ in $s < r$ dimensions, and where $\|\cdot\|$ denotes the usual Frobenius (Euclidean) norm. Basically, these methods are variations (with fixed or optimal stepsizes) of iteratively computing a successor configuration $\mathbf{X}^+$ for $\mathbf{X}$ of the same order, and satisfying

$$\|\mathbf{P} - D(\mathbf{X}^+)\|^2 \leq \|\mathbf{P} - D(\mathbf{X})\|^2, \tag{3}$$

for fixed $\mathbf{X}$. Kruskal [16, 17] proposed a gradient method and re-estimated a stepsize γ in each iteration. The method proposed by Guttman [10] implicitly used the same algorithm, but with a fixed stepsize. (Again, see Meulman, Hubert, and Arabie [22] for details regarding these candidate successor configurations.) De Leeuw and Heiser [5] showed that the procedure proposed by Guttman is equivalent to a subgradient method, and a convergence proof was given, while De Leeuw and Heiser [6] interpreted the associated update $\mathbf{X}^+$ as a majorization procedure, and offered a new, alternative convergence proof.

2 Our Data Analytic Problem

In the present paper, we use a set of iterative methods that are integrated to minimize the objective function

$$STRESS(\mathbf{Q}; \mathbf{x}; c) = \|D(\mathbf{Q}) + c - D(\mathbf{x})\|^2. \tag{4}$$

Here, the vector $\mathbf{x}$ defines a one-dimensional scale to be fitted to the proximities in $D(\mathbf{Q})$, and c is a constant to be estimated; $D(\cdot)$ contains Euclidean distances $d_{ij}(\cdot)$ according to (1) between all objects i and j, $i, j = 1, \ldots, n$. $\mathbf{Q}$ denotes the data matrix, with columns allowed to be transformed according to a prespecified transformation level. For ordinal data, the transformation level would typically be monotonic, using either the full parameter space, or a monotonic spline function with a prespecified degree and number of interior knots.

The combination of multivariate analysis and optimal scaling (i.e., exchanging the variables in $\mathbf{Z}$ with a set of transformed variables in $\mathbf{Q}$), has a long history in psychometrics. The major impetus was the extension of the loss function in (2) to include optimization over only the ordinal information present in the proximities in $\mathbf{P}$, an approach pioneered by Shepard [27, 28] and Kruskal [16, 17]. Subsequently, the principle of optimal transformation was transferred from proximity data to multivariate data, with optimal scaling of the variables instead of the proximities. Selected highlights from the early psychometric literature on optimal scaling include Kruskal [18], Shepard [29], Roskam [26], De Leeuw [4], Kruskal and Shepard [19], Young, De Leeuw, and Takane [38], Young, Takane, and De Leeuw [39], Winsberg and Ramsay [34, 35], Van der Burg and De Leeuw [31], Van der Burg, De Leeuw, and Verdegaal [32], and Ramsay [25]. Approaches to systematization are the "ALSOS" system [36], and the Leiden "Albert Gifi" system [8]. The more mainstream statistical literature has acknowledged optimal scaling by the papers by Breiman and Friedman [1], Ramsay [25], Buja [2], and Hastie, Buja, and Tibshirani [11].

For the data analytic task of the present paper, optimal scaling has to be integrated with the task of finding an optimal ordering of objects and spacing the objects as points along a one-dimensional continuum (along with an estimate for the constant c).

2.1 The Three Subtasks and Corresponding Optimization Methods

Finding an optimal ordering of objects is a combinatorial optimization task, and it is known, for example, from Defays [3], De Leeuw and Heiser [5], and Hubert and Arabie [12] that (sub)gradient and majorization methods are prone to identifying seriously suboptimal solutions whenever applied to scaling tasks that are limited to one dimension.

We begin with a given $n \times n$ symmetric proximity matrix $\mathbf{P} = \{p_{ik}\}$, whose entries have a distance interpretation in which larger values reflect the more

dissimilar objects. An ordering of the n objects is sought, represented by a permutation of the first n integers, $\rho(\cdot)$, such that the reordered matrix $\mathbf{P}_\rho \equiv \{p_{\rho(i)\rho(k)}\}$ is as "close as possible", defined by an explicit objective function, to a so-called anti-Robinson matrix (see Hubert and Arabie [13] and the references therein for the historical background to the use of the term "anti-Robinson"). Formally, an $n \times n$ symmetric matrix, $\Delta \equiv \{\delta_{ik}\}$, has an anti-Robinson form if the entries within the rows and columns never decrease as we move away from the main diagonal in any direction, i.e.,

$$\begin{aligned} &\delta_{ik} \leq \delta_{i(k+1)} \text{ for} 1 \leq i < k \leq n-1; \\ &\delta_{ik} \leq \delta_{i(k-1)} \text{ for } 2 \leq k < i \leq n. \end{aligned} \tag{5}$$

Except for small values of n, a complete enumeration strategy that would exhaustively consider all $n!$ permutations is computationally infeasible. To construct a general search procedure that would work for any reasonable size for n, a heuristic strategy is adopted based on suggestions from the quadratic assignment literature (e.g., see Hubert and Schultz [14]), and which has been successfully applied in Hubert and Arabie [13] for fitting anti-Robinson matrices to given proximity matrices.

Specifically, we begin by computing the cross-product $\Gamma(\rho(\cdot)) = \sum_{i,k} p_{\rho(i)\rho(k)}\delta_{ik}$ for some (randomly) chosen permutation $\rho(\cdot)$ and equally-spaced anti-Robinson target matrix $\Delta = \{|i-k|\}$, and attempt by a series of local operations to produce a sequence of permutations that increase $\Gamma(\cdot)$ until no local operation can improve on its value. The strategy includes local operations from three classes: (i) all pairwise interchanges of the n objects; (ii) all insertions of l consecutive objects between any two existing objects (or at the beginning and end of the permutation); (iii) all complete reversals of the orderings of l consecutive objects, $l = 2, \ldots, n$. When the best permutation, say, $\rho^*(\cdot)$, is found, a "new" anti-Robinson target matrix Δ is reconstructed by a least-squares fit to the proximity matrix $\mathbf{P}_{\rho^*}$, and the whole process repeated, until stability is eventually achieved both in the identification of $\rho^*(\cdot)$ and Δ.

Optimal spacing of the object points amounts to obtaining scale values for the (ordered) objects along a one-dimensional continuum. To obtain scale values and to construct a graphical display, requires additional constraints, to be imposed on the fitting of the anti-Robinson matrix Δ for the identified permutation $\rho^*(\cdot)$. Explicitly, Δ is constrained as $\Delta = D(\mathbf{x})$, with $D(\mathbf{x})$ the matrix containing all pairs of Euclidean (or more generally, Minkowski) distances, $d_{ik}(\mathbf{x}) = |x_i - x_k|$, for some collection of coordinates $x_1 \leq \cdots \leq x_n$. Given the ordering $\rho^*(\cdot)$, the scale values (in $\mathbf{x}$) and the constant (c) can be obtained in several ways. In the present paper, we use Dykstra's [7] iterative projection approach. For a fixed permutation $\rho(\cdot)$, the set of all $n \times n$ matrices equal to the reordered proximity matrix $\{p_{\rho(i)\rho(k)}\}$ up to an additive constant c will be denoted by $\mathcal{P}_\rho$, and $\mathcal{D}$ denotes the set of all $n \times n$ matrices that give the interpoint distances between the objects along a one-dimensional continuum for $x_1 \leq \cdots \leq x_n$. The sets $\mathcal{P}_\rho$ and $\mathcal{D}$ are both closed and convex; thus, projections of any $n \times n$ symmetric matrix onto either $\mathcal{P}_\rho$ or $\mathcal{D}$ can be constructed.

Because we need to find both the coordinates in $\mathbf{x}$ and the additive constant c simultaneously, an alternating projection strategy is implemented between the two sets $\mathcal{P}_\rho$ and $\mathcal{D}$. The projection of any given matrix onto $\mathcal{P}_\rho$ to generate an optimal value for c is straightforward, but the projection onto $\mathcal{D}$ is somewhat more complicated. This minimization requires the satisfaction of several linear inequality/equality constraints:

(i) $0 \leq d_{ik}(\mathbf{x})$ for $1 \leq i \neq k \leq n$;
(ii) $0 = d_{ii}(\mathbf{x})$ for $1 \leq i \leq n$;
(iii) $d_{ik}(\mathbf{x}) = d_{ki}(\mathbf{x})$ for $1 \leq i \neq k \leq n$;
(iv) $d_{i(i+1)}(\mathbf{x}) + \cdots + d_{(k-1)k}(\mathbf{x}) = d_{ik}(\mathbf{x})$ for $1 \leq i < k \leq n$, where $2 \leq k - i$.

The details for solving this particular optimization problem can be found in Hubert, Arabie, and Meulman [15]. Basically, the constraints are considered sequentially and checked as to whether each is satisfied; if not, the current update is projected onto the subset that does satisfy the constraint. In this cyclic process, the previous changes are "undone" when each constraint is reconsidered (see Dykstra [7] for a proof of convergence to the desired projection onto the closed convex set).

Optimal scaling of each variable implies finding a transformation $\mathbf{q}_j = \varphi(\mathbf{z}_j)$ with maximal fit for fixed $\mathbf{x}$. Thus, the optimal scaling process is guided by the objective function

$$STRESS(\mathbf{x}; c) = \min_{\mathbf{q}_j \in C_j} \|D(\mathbf{Q}) + c - D(\mathbf{x})\|^2, \tag{6}$$

with $\mathbf{q}_j'\mathbf{q}_j = 1$. For a given matrix $D(\mathbf{x})$ and constant c, finding the optimal $\mathbf{Q}$ can be solved by an optimization strategy based on majorization (which is equivalent to a subgradient method) and projection (see De Leeuw and Heiser [6] for the basic principles).

In our algorithm, the majorization step finds for each current $\mathbf{Q}$ an unconstrained update $\mathbf{Y}$ of the same order as $\mathbf{Q}$ that satisfies

$$\|D(\mathbf{Y}) - D(\mathbf{x})\|^2 \leq \|D(\mathbf{Q}) - D(\mathbf{x})\|^2, \tag{7}$$

for fixed $\mathbf{x}$. This task is carried out by choosing $\mathbf{Y}$ as $\mathbf{Y} = n^{-1}\mathbf{B}(\mathbf{Q})\mathbf{Q}$. Here, the $n \times n$ matrix $\mathbf{B}(\mathbf{Q})$ is defined as $\mathbf{B}(\mathbf{Q}) = \mathbf{B}_t(\mathbf{Q}) - \mathbf{B}_r(\mathbf{Q})$; the $n \times n$ matrix $\mathbf{B}_r(\mathbf{Q})$ has elements $b^r_{ik}(\mathbf{Q}) = d_{ik}(\mathbf{x})/d_{ik}(\mathbf{Q})$ if $i \neq k$, and $b^r_{ik}(\mathbf{Q}) = 0$ if $i = k$ or $d_{ik}(\mathbf{Q}) = 0$. The $n \times n$ diagonal matrix, $\mathbf{B}_t(\mathbf{Q})$, has elements $b^t_{ii}(\mathbf{Q}) = \mathbf{u}'\mathbf{B}_r(\mathbf{Q})\mathbf{e}_i$ along its main diagonal. The required transformation is constructed by a metric projection of the unconstrained update $\mathbf{Y}$ onto the space that satisfies the constraints; explicitly, if a class of cones is defined for all admissible transformations of the variables in $\mathbf{Q}$, we write the metric projection of $\mathbf{Y}$ onto the cone as $\mathbf{P}_C(\mathbf{Y}) = \hat{\mathbf{Q}} = \{\hat{\mathbf{q}}_1, \ldots, \hat{\mathbf{q}}_j, \ldots, \hat{\mathbf{q}}_M\}$, where $\hat{\mathbf{Q}}$ minimizes $\|\mathbf{Y} - \mathbf{Q}\|^2$ over all $\mathbf{q}_j \in C_j$. The convergence results of De Leeuw and Heiser [6] can be used to show that $\hat{\mathbf{Q}}$ will yield the desired result:

$$\|D(\hat{\mathbf{Q}}) - D(\mathbf{x})\|^2 \leq \|D(\mathbf{Q}) - D(\mathbf{x})\|^2. \tag{8}$$

The metric projection $\mathbf{P}_C(\mathbf{Y})$ needs to be specified by a particular choice of transformation. If we choose monotonic spline transformations and construct an I-spline basis matrix $\mathbf{S}_j$ for the j^{th} variable in $\mathbf{Z}$ (see Ramsay [25] for details),

$$\|\mathbf{y}_j - \mathbf{S}_j\mathbf{b}_j\|^2, \tag{9}$$

is minimized, under the conditions that the spline coefficient vector $\mathbf{b}_j$ contains nonnegative entries (to guarantee monotonic I-splines) and $\mathbf{b}_j'\mathbf{S}_j'\mathbf{S}_j\mathbf{b}_j = 1$ (to give the transformed variable a unit length). To ensure nonnegativity of the entries in $\mathbf{b}_j$, the problem is further partitioned by separating the t^{th} column of the spline basis matrix $\mathbf{S}_j$ (denoted by $\mathbf{s}_t^j$, with the remaining matrix denoted by $\mathbf{S}_{-t}^j$) and the t^{th} element from the spline coefficient vector $\mathbf{b}_j$ (denoted by b_t^j, with the remaining vector denoted by $\mathbf{b}_{-t}^j$). We minimize iteratively:

$$\|(\mathbf{y}_j - \mathbf{S}_{-t}^j\mathbf{b}_{-t}^j) - \mathbf{s}_t^j b_t^j\|^2, \tag{10}$$

over elements $b_t^j \geq 0, t = 1, \ldots, M$ (where M is dependent on the degree of the spline and the number of knots), and for $j = 1, \ldots, m$. After an update for $\mathbf{Q}$ has been determined, the procedure returns, and the three steps described in this section are cycled through repeatedly until no change occurs.

3 An Empirical Illustration

To illustrate our iterative method, we use a data set that concerns the perceived effectiveness, safety, availability, and convenience of 15 birth control methods. Four groups of respondents (two groups consisting of female respondents, and two of male respondents) ranked the methods from 1 to 15 according to the four criteria. Each group consisted of seven individuals, and their rankings were aggregated [33]. For the present analysis, it is important to use the aggregate rankings per group as the variables in the analysis, and the 15 contraceptive methods as the row objects (thus we analyze the transpose of the usual data matrix.) The objective of the analysis is to find an optimal ordering of the 15 methods displaying a consensus among the respondents, on the basis of optimal transformation of the rankings for each of the eight groups of respondents. The rankings were optimally transformed with a quadratic spline function with one interior knot (thus, for every ranking, three parameters were fitted in the transformation). The solution found by our iterative method accounts for 92.6% of the total variance in the distances between the objects according to the transformed variables. One ordering and scaling is depicted in Fig. 1. The sequence is from the extreme birth control method *abortion* via the other surgeries, *hysterectomy*, *tubal ligation*, and *vasectomy*, to the other extreme *abstinence* via *rhythm*, *withdrawal*, and *oral sex*. The *condom*, *pill*, and *iud* are in

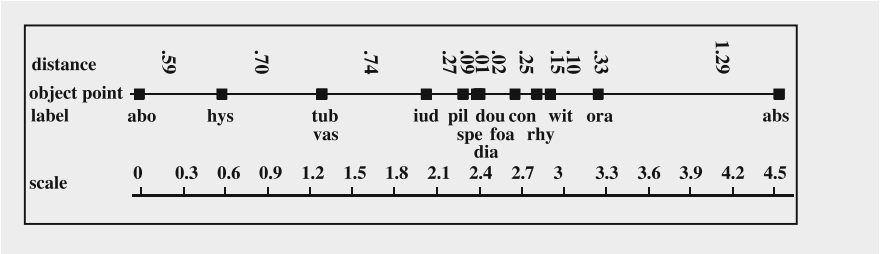


Fig. 1 The proximity between 15 birth control methods fitted along a one-dimensional continuum. The horizontal axis displays the ordering and the spacing of the objects, showing the consensus among the four groups of respondents on the four different criteria

the center of the continuum. The spacing between the points corresponds to the fitted distances; the scale values are given as well. The largest distance between adjacent objects is for *abstinence* and *oral sex* (1.29). At the other end of the continuum, *hysterectomy* is 0.59 from *abortion* and 0.70 from *tubal ligation* and *vasectomy*, which coincide and are 0.74 from *intra-uterine device*. The latter is much closer (0.27) to the *pill*, but the smallest distances are from the *pill* to *spermicide* (0.09) to *diaphragm* (0.01) to *douche* (0.02) and to *foam*, tied with *douche*. At some distance (0.25) we find *condom* to *rhythm* (0.15), and *withdrawal* (0.10), and finally we have *oral sex* at distance 0.33, and *abstinence* at the end of the scale.

Next, we consider the transformation of the (aggregate) rankings from 7 (7×1) to 105 (7×15) to their optimally scaled versions (in Figs. 2 and 3). In the left two panels of Fig. 2, the transformations for the women groups are displayed; the two panels on the right are transformations for the male groups. The two top panels are transformations according to *effectiveness*, followed by *safety* at the bottom. It is clear that there is a great deal of consensus, with the overall exception being the second male group. Although the women and the first male group believe that *oral sex* and *abstinence* is efficient and safe, the second male group ranks *abstinence* and *oral sex* much lower on *effectiveness*, and *oral sex* much lower on *safety*. The corresponding plots for *convenience*, and *availability* are given in Fig. 3. Here it can be seen that the women judge *oral sex* and *abstinence* as available and convenient, but both male groups believe that *oral sex* is not available, and the second male group that *oral sex* is not convenient either.

An alternative display of the optimal scale values from Fig. 1 is given in Fig. 4. Here, the horizontal axis displays the objects sequenced according to the scale values (and equally spaced), while the vertical axis displays the optimal spacing according to the scale values. The transformations in Figs. 2 and 3 were found separately for each of the 4 groups and according to each of the four criteria. As an interpretation, Fig. 3 displays a "consensus" transformation over all groups of respondents and criteria. Overall, the consensus ordering and scaling in $\mathbf{x}$ is most associated (according to the Pearson correlations) with SF1 (0.95), SM1 (0.91), CF1 (0.90), AF2 (0.89), AF1 (0.85), CF2 (0.81), SF2 (0.73), and SM2 (0.70), and least with EM2 (-0.28), EF1 (0.0), EF2 (0.17), AM1 (0.41), CM1 (0.48), AM2 (0.51),

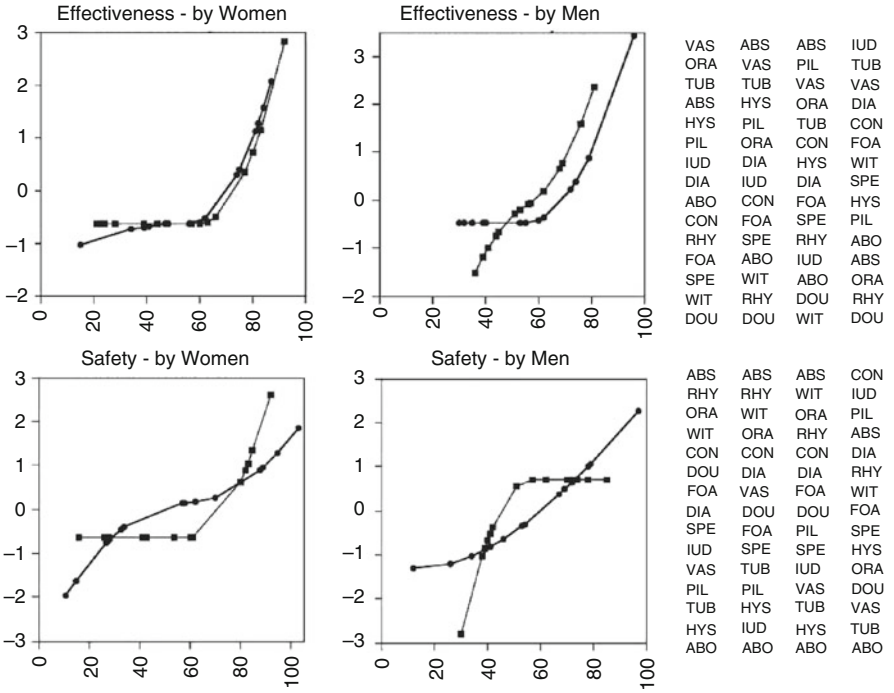


Fig. 2 The optimally scaled rankings (vertical axis) versus original ranking scores (horizontal axis) giving transformation plots. The transformations were found for two groups of men and two groups of women on four criteria; above are the optimally scaled rankings according to Effectiveness and Safety. The transformations for the two female and the two male groups were combined into a single plot each. The sequence of the 15 contraceptive methods - ordered from high to low for each of the four groups of judges - is given at the right-hand side

EM1 (0.56), and CM2 (0.67). Thus, *safety, convenience*, and *availability* according to women and *safety* according to men are closest to the consensus ordering in **x**, while *availability* and *convenience* according to men, and *effectiveness* according to all groups of respondents, are not. The women respondents are generally closer to the consensus ordering than the male groups, and the criterion *effectiveness* is judged differently from the other three.

To evaluate the possible presence of (suboptimal) local optima, the analysis reported above was repeated with 100 random starts for the initial ordering of the objects. Of the 100 solutions, 99 were identical, with 92.6% of the variance in the distances accounted for; the only solution that differed gave a slightly less variance-accounted-for of 92.5%. The ordering in this solution differed for the objects *withdrawal* and *rhythm*: in the optimal solution, rhythm comes before withdrawal (with scale values of 2.806 and 2.903, respectively); in the suboptimal solution, rhythm and withdrawal have almost the same scale values (2.835 and 2.834, respectively). The rest of the one-dimensional scale was virtually identical, as well as the transformed rankings per group.

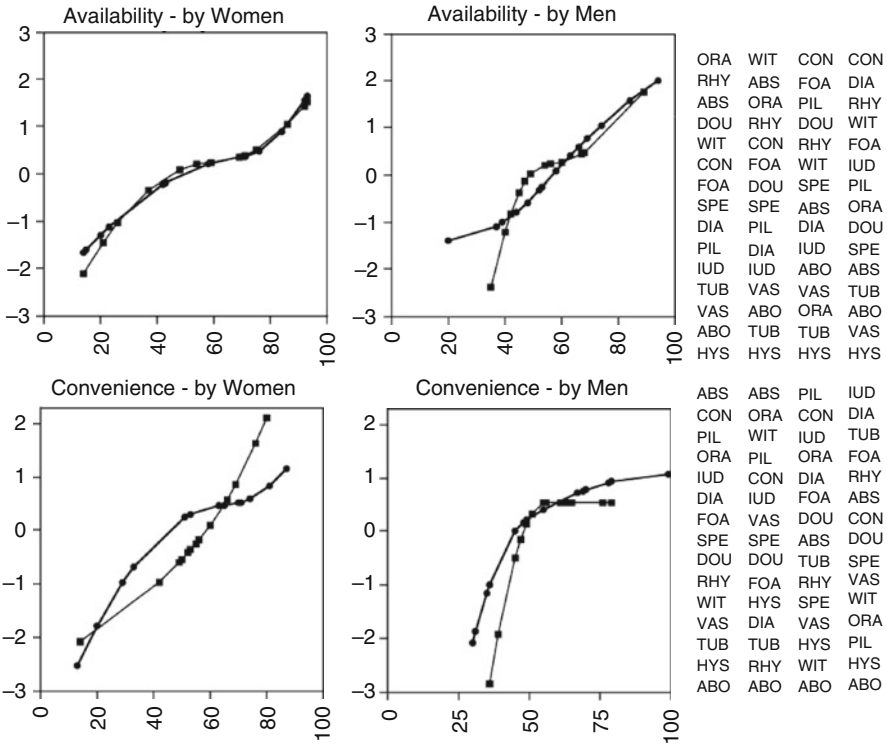


Fig. 3 Optimally scaled rankings as in Figure 2, but now for Availability and Convenience

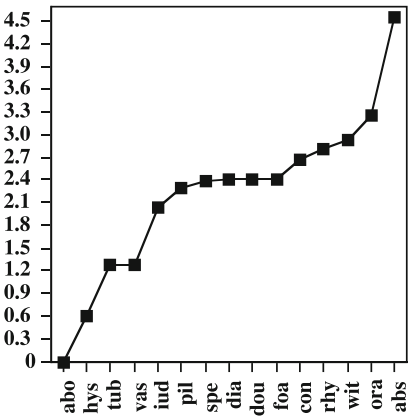


Fig. 4 The proximity between 15 birth control methods fitted along a one-dimensional continuum: the horizontal axis represents the ordering and the vertical axis represents the spacing: optimally scaled consensus ranking

4 Concluding Remarks

In the present paper, we have shown how to construct a parsimonious one-dimensional representation for a (large) number of transformed variables. The data analytic objective was to incorporate the use of discrete multivariate ordinal data,

where relationships between variables could be nonlinear, and where the data might be decidedly nonnormal in distribution. Optimal scaling was proposed as a way for dealing with this nontraditional type of multivariate data. The data analytical task is computationally a nested iterative procedure, where objects are optimally sequenced cyclically, spaced optimally, and all implemented with optimal transformation of the variables in the original data.

Using this combinatorial data analysis framework, the transformation of variables can also be combined with fitting the special structures described by Hubert, Arabie, and Meulman [15]. One extension is to the construction of a circular unidimensional scale, where in addition to the ordering, the scale values, and the additive constant, a set of inflection points must be found that indicate where the minimum distance calculation must change direction around the closed continuum. A second generalization is to the use of additive trees, defined by a structure where the fitted value d_{ik} is the length of the path that joins the two objects i and k. The fitted values satisfy what is called the four-point condition: over all distinct i, k, l, and m, among the three sums, $d_{ik} + d_{lm}$, $d_{il} + d_{km}$, and $d_{im} + d_{kl}$, the largest two are equal. As a special case of an additive tree, ultrametrics can be fitted, where for any three distinct objects i, k, and l, the largest two values among d_{ik}, d_{il}, and d_{kl}, are equal. The latter condition induces a sequence of partitions for the object set.

Finally, in the general distance analysis framework considered, we can also fit objects in a representation space with dimensionality $s > 1$, as done by Meulman [20, 21]. In contrast to their performance in the unidimensional scaling task (and their difficulty with identifying seriously suboptimal solutions), gradient and majorization methods generally work well in constructing such multidimensional representations.

References

1. Breiman, L., Friedman, J.H.: Estimating optimal transformations for multiple regression and correlation. J. Am. Stat. Assoc. **80**, 580–598 (1985)
2. Buja, A.: Remarks on functional canonical variates, alternating least squares methods and ACE. Ann. Stat. **18**, 1032–1069 (1990)
3. Defays, D.: A short note on a method of seriation. Br. J. Math. Stat. Psychol. **31**, 49–53 (1978)
4. De Leeuw, J.: Canonical analysis of categorical data, 2nd edn, 1984. Leiden University DSWO Press, Leiden (1973)
5. De Leeuw, J., Heiser, W.J.: Convergence of correction matrix algorithms for multidimensional scaling. In: Lingoes, J. (ed.) Geometric Representations of Relational Data, pp. 735–752. Mathesis Press, Ann Arbor, MI (1977)
6. De Leeuw, J., Heiser, W.J.: Multidimensional scaling with restrictions on the configuration. In: Krishnaiah, P.R. (ed.) Multivariate Analysis, vol. V, pp. 501–522. North-Holland, Amsterdam (1980)
7. Dykstra, R.L.: An algorithm for restricted least squares regression. J. Am. Stat. Assoc. **78**, 837–842 (1983)
8. Gifi, A.: Nonlinear Multivariate Analysis, 1st edn, 1981, Department of Data Theory, Leiden University. Wiley, Chicester (1990)
9. Gower, J.C.: Some distance properties of latent roots and vector methods used in multivariate analysis. Biometrika **53**, 325–338 (1966)

10. Guttman, L.: A General nonmetric technique for finding the smallest coordinate space for a configuration of points. Psychometrika **33**, 469–506 (1968)
11. Hastie, T., Tibshirani, R., Buja, A.: Flexible discriminant analysis. J. Am. Stat. Assoc. **89**, 1255–1270 (1994)
12. Hubert, L.J., Arabie, P.: Unidimensional scaling and combinatorial optimization. In: De Leeuw, J., Heiser, W.J., Meulman, J.J., Chritchley, F.(eds.) Multidimensional Data Analysis, pp. 181–196. DSWO Press, Leiden (1986)
13. Hubert, L.J., Arabie, P.: The analysis of proximity matrices through sums of matrices having (anti-)Robinson forms. Br. J. Math. Stat. Psychol. **47**, 1–40 (1994)
14. Hubert, L.J., Schultz, J.V.: Quadratic assignment as a general data analysis strategy. Br. J. Math. Stat. Psychol. **29**, 190–241 (1976)
15. Hubert, L.J., Arabie, P., Meulman, J.J.: Linear and circular unidimensional scaling for symmetric proximity matrices. Br. J. Math. Stat. Psychol. **50**, 253–284 (1997)
16. Kruskal, J.B.: Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika **29**, 1–28 (1964)
17. Kruskal, J.B.: Nonmetric multidimensional scaling: a numerical method. Psychometrika **29**, 115–129 (1964)
18. Kruskal, J.B.: Analysis of factorial experiments by estimating monotone transformations of the data. J. R. Stat. Soc. Ser. B. **27**, 251–263 (1965)
19. Kruskal, J.B., Shepard, R.N.: A nonmetric variety of linear factor analysis. Psychometrika **39**, 123–157 (1974)
20. Meulman, J.J.: The integration of multidimensional scaling and multivariate analysis with optimal transformations of the variables. Psychometrika **57**, 539–565 (1992)
21. Meulman, J.J.: Fitting a distance model to homogeneous subsets of variables: points of view analysis of categorical data. J. Classification **13**, 249–266 (1996)
22. Meulman, J.J., Hubert, L.J., Arabie, P.: Ordering and scaling objects in multivariate data under nonlinear transformations of variables (extended version), Leiden. http://www.datatheory.nl/pages/meulman.html (2008)
23. Ramsay, J.O.: Maximum likelihood estimation in multidimensional scaling. Psychometrika **42**, 241–266 (1977)
24. Ramsay, J.O.: MULTISCALE II Manual, Department of Psychology. McGill University, Montreal (1982)
25. Ramsay, J.O.: Monotone regression splines in action. Stat. Sci. **3**, 425–441 (1988)
26. Roskam, E., E., C., I.: Metric analysis of ordinal data in psychology. VAM, Voorschoten (1968)
27. Shepard, R.N.: The analysis of proximities: multidimensional scaling with an unknown distance function I. Psychometrika **27**, 125–140 (1962)
28. Shepard, R.N.: The analysis of proximities: multidimensional scaling with an unknown distance function II. Psychometrika **27**, 219–246 (1962)
29. Shepard, R.N.: Metric structures in ordinal data. J. Math. Psychol. **3**, 287–315 (1966)
30. Torgerson, W.: Theory and methods of scaling. Wiley, New York, NY (1958)
31. Van der Burg, E., De Leeuw, J.: Nonlinear canonical correlation. Br. J. Math. Stat. Psychol. **36**, 54–80 (1983)
32. Van Der Burg, E., De Leeuw, J., Verdegaal, R.: Homogeneity analysis with k sets of variables: an alternating least squares method with optimal scaling features. Psychometrika **53**, 177–197 (1984)
33. Weller, S.C., Romney, A.K.: Metric scaling: Correspondence analysis. Sage, Newbury Park, CA (1990)
34. Winsberg, S., Ramsay, J.O.: Monotonic transformations to additivity using splines. Biometrika **67**, 669–674 (1980)
35. Winsberg, S., Ramsay, J.O.: Monotone spline transformations for dimension reduction. Psychometrika **48**, 575–595 (1983)
36. Young, F.W.: Quantitative analysis of qualitative data. Psychometrika **46**, 357–387 (1981)
37. Young, G., Householder, A.S.: Discussion of a set of points in terms of their mutual distances. Psychometrika **3**, 19–22 (1938)

38. Young, F.W., De Leeuw, J., Takane, Y.: Regression with qualitative and quantitative variables: an alternating least squares method with optimal scaling features. Psychometrika **41**, 505–528 (1976)
39. Young, F.W., Takane, Y., De Leeuw, J.: The principal components of mixed measurement level multivariate data: an alternating least squares method with optimal scaling features. Psychometrika **43**, 279–281 (1978)

Statistical Models to Predict Academic Churn Risk

Paolo Giudici and Emanuele Dequarti

Abstract This paper describes a research conducted on university students careers. The purpose is to study, describe and prevent the phenomenon of abandonment (churn). Results from predictive models may be employed to start activities of personalized tutoring, aimed at preventing the phenomena

1 Introduction

In this paper we present how to analyse student careers data, in order to predict and, therefore, prevent, students abandonment. We show empirical evidence derived from real data from the Faculty of Psychology of the University of Pavia.

The faculty of psychology was chosen because we deal with strongly motivated students. This makes the analysis of the data particularly meaningful. The main objective of this paper is to study the phenomenon of churn concerning the students, in order to reduce the number of students that leave their university career without reaching the degree.

In profit companies is rather immediate to establish the number of clients that abandons a service. The distinction is usually made between voluntary and involuntary churn. Involuntary churn occurs when the company terminates the customers' contract or account - usually on the basis of a poor payment history. Voluntary churn is when the customer decides to take their business elsewhere.

Now we adapt these concepts to a situation that interests the students of a faculty (see E.6. [5]). Voluntary churn is when the student interrupts academic studies. The student could definitely leave his/her career. We will call this possibility *renounced student*. A different event is that the student could continue his/her career in another University. We will call this possibility *dismissed student*. *Renounced student* and *dismissed student* are positions officially enacted in official documents.

There are also positions of students that officially result "active", but that do not take exams anymore and they suspend the payment of the annual tax for years. The

P. Giudici (✉)
Department of Statistics and Applied Economics Libero Lenti, University of Pavia, Pavia, Italy
e-mail: giudici@unipv.it

B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-13312-1_4,

reception office considers these kind of students as "renounced student" after eight years of complete inactivity, but it is clearly possible to establish with wide advance this condition of abandonment.

It thus becomes necessary to individualize a criterion to establish what students formally held active, are instead definitely inactive.

The missed payment of the annual tax of registration, with the exception of the graduated students, clearly signals a possible risk of churn. A necessary caution is the need to verify that the student does not restart to pay taxes in following years.

The final objective is to define profiles of students with high churn risk, depending on the following factors: given credits, average mark of given exams, and the followings social- demographic variables: gender, date of birth, province of residence, type of middle school diploma.

These profiles could be used for beginning tutorial and counselling activities in order to prevent churn risk.

For relevant contributions about analysis on the performances of the students see also [2] [7] [8] [13] [14] [15].

To reach this objective we employ descriptive and inferential models such us Kaplan Meier survival function and Cox model (see E.6. [3]).

The paper is organised as follows: in Sect. 2 we present the data available, in Sect. 3 our methodological proposal aimed at estimating for each student churn risk; finally in Sect. 4 empirical evidence is given. Section 5 reports the conclusion and further ideas of research.

2 Data Set

The data available for our analysis contain information about the Degree of Psychology. The data set contains 845 observations: for each row (statistical units), which represents in this case a different student, we report variables useful to describe the university career of each student and his/her social-demographic characteristics.

The variables are:

- ID: (univocal code of identity for each student).
- $Dateofbirth$: (year).
- $Gender$: (female, male).
- $ProvRec$: (italian province of residence).
- $Diploma$: type of middle school diploma (professional institute, technical institute, classical high school, linguistic high school, scientific high school, teacher's college, Other).
- $Position$: current position of the student (Active, Dismissed, Interruption, Graduated, Renounced, Declined).
- $Credits$: (credits maturated by the student in the exams for each year of his career). The minimum value is 5, the maximum is 185.
- $Averagemark$: (average mark of the given exams for each year of the career). The minimum value is 18, the maximum is 30.
- Tax: type of payment of the tax for each year (none, only first, full).

The period of time considered is: 2002–2006. Missing values are present for the $Credits$ and the $Averagemarkvariables$. This means that the student did not take any exam in that year.

The response variable is called *Churn*. We will use it to examine the distribution of times between two events, the beginning and the end of the career. This kind of data includes some censored cases. Censored cases are cases for which the second event is not recorded. In this case the censored cases include students graduated and students *Active*, that are regularly continuing the university career and paying the tax. The status variable *Churn* is dichotomic and it presents these values:

- 1: the (churn) event has occurred
- 2: the event is censored.

3 Methodological Proposal

A number of components can generate a churn behaviour:

- a static component, determined by the characteristics of the students;
- a dynamic component, that encloses trend and the contacts of the students with the university career;
- a seasonal part, linked to the period of exams;
- external factors.

The goal of the university is to identify students that are likely to leave and join a new university or a job. This objective is well perceived by the university top management, which considers lowering churn one of the key targets.

The churn models currently used to predict churn are logistic regression and classification trees. However, in order to obtain a predictive tool which is able to consider the fact that churn data is ordered in calendar time, the use of new methods is necessary. We propose to use a survival analysis approach, as a longitudinal data analysis method aimed at predicting students' churn behavior. For lack of space, we do not present details of our methodology here, but refer the reader to the paper [6], that contains the methodology employed and a more extensive treatment of the context, as also explained by Fleming and Harrington [9].

4 Empirical Evidence

We now proceed with survival analysis modeling. Through the analysis of the payments of the taxes and the activities of the students we classify the careers, defining the variable churn, that points out abandons. As shown in Table 1, on 845 cases, 188 students experiment the event abandonment, while 657 students, 77, 8 %, are considered censored cases.

In Fig. 1 and 2 we show the survival and hazard functions, relatively to the careers of the students of psychology in the five analyzed years.

In Table 2 we report the results of the application of the Kaplan Meier estimator [12] to the data.

Table 1 Distribution of the events of abandonment and censored cases, on Psychology

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Abandonment	188	22.2	22.2	22.2
	Active/censored	657	77.8	77.8	100.0
	Total	845	100.0	100.0	

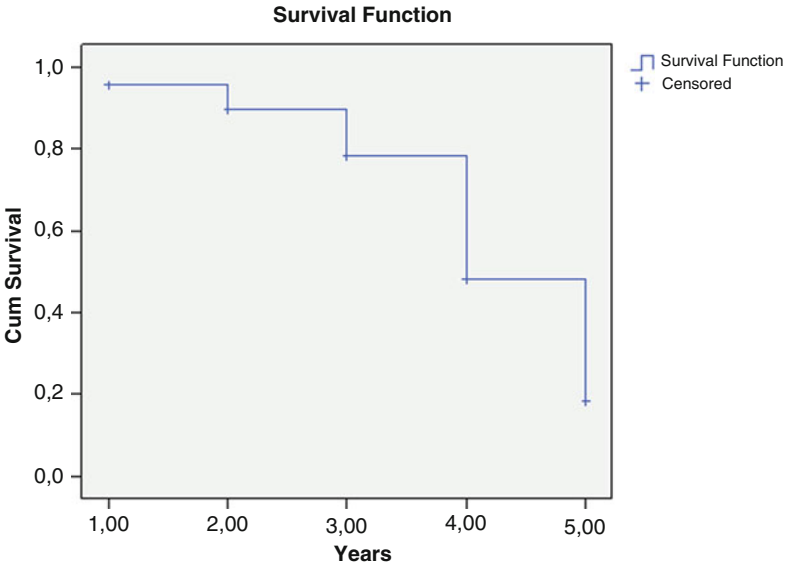


Fig. 1 Survival function

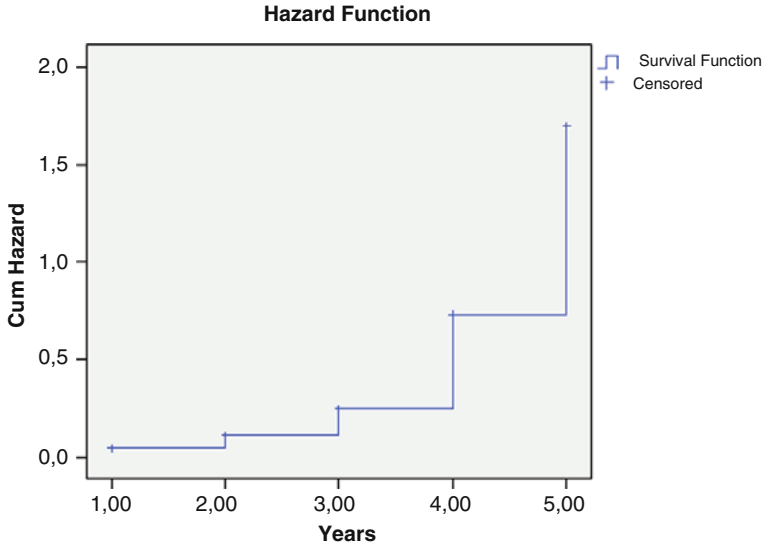


Fig. 2 Hazard function

From Fig. 2 and Table 2, note that the probabilities of survival, after one year from the beginning of the studies, are of 95, 6%, for a total of 37 subjects that experiment the event abandonment. Probabilities moderately decrease, then, in the second and third year. They lower in more evident way from the fourth year, being equal to 48, 1%. An intervention of tutoring for the students that have not completed the career after the three years could be meaningful.

Table 3 shows that the highest percentage of students focuses in the band between 141 and 170 credits.

Table 2 Kaplan Meier estimator

Time	n.risk	n.event	Survival	Std.err	Lower 95% CI	Upper 95% CI
1	845	37	0.956	0.00704	0.943	0.97
2	707	44	0.897	0.01091	0.876	0.918
3	448	57	0.783	0.01703	0.75	0.817
4	96	37	0.481	0.04026	0.408	0.567
5	21	13	0.183	0.05323	0.104	0.324

Table 3 Distribution of the variable credits banded

	Frequency	Percent	Valid percent	Cumulative percent
<= 20,00	24	2,8	2,8	2,8
21,00 – 50,00	44	5,2	5,2	8
51,00 – 80,00	136	16,1	16,1	24,1
81,00 – 110,00	171	20,2	20,2	44,4
111,00 – 140,00	188	22,2	22,2	66,6
141,00 – 170,00	277	32,8	32,8	99,4
171,00+	5	0,6	0,6	100
Total	845	100	100	

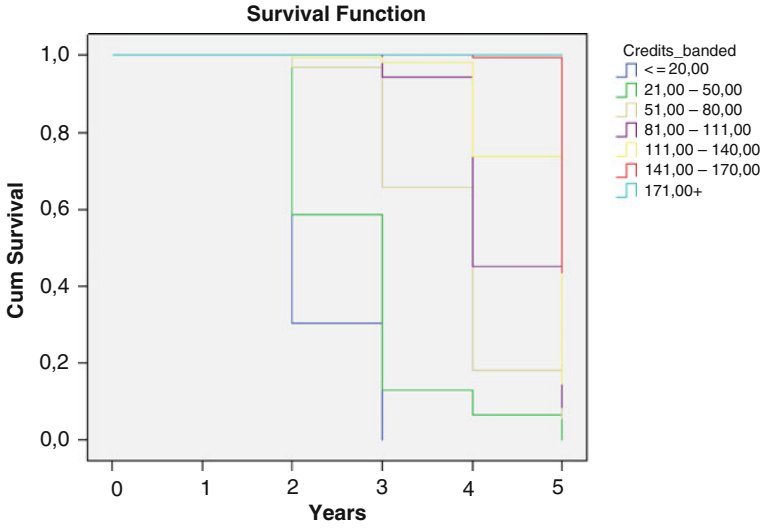


Fig. 3 Survival function for the factor Credits

Figure 3 and Table 4 show that the Kaplan-Meier method, applied to the data, allows us to compare overall survival rates between different groups of student, according to different factors:

- Credits
- Average mark
- Date of birth
- Gender
- Italian province of residence
- Diploma

Table 4 Kaplan Meier estimator for the factor credits

Credits_banded=<= 20.00						
Time	n.risk	n.event	Survival	Std.err	Lower 95% CI	Upper 95% CI
1	24	16	0.333	0.0962	0.189	0.587
2	6	6	0	NA	NA	NA
Credits_banded = 21.00 – 50.00						
Time	n.risk	n.event	Survival	Std.err	Lower 95% CI	Upper 95% CI
1	44	17	0.614	0.0734	0.4854	0.776
2	21	16	0.146	0.0597	0.0656	0.325
3	4	2	0.073	0.0472	0.0206	0.259
4	2	2	0	NA	NA	NA
Credits_banded = 51.00 – 80.00						
Time	n.risk	n.event	Survival	Std.err	Lower 95% CI	Upper 95% CI
1	136	3	0.978	0.0126	0.9536	1
2	48	14	0.693	0.0648	0.5767	0.832
3	25	17	0.222	0.0679	0.1216	0.404
4	5	2	0.133	0.0634	0.0523	0.338
5	3	3	0	NA	NA	NA
Credits_banded = 81.00 – 110.00						
Time	n.risk	n.event	Survival	Std.err	Lower 95% CI	Upper 95% CI
2	168	6	0.964	0.0143	0.937	0.993
3	39	18	0.519	0.0774	0.388	0.695
4	12	7	0.216	0.0806	0.104	0.449
5	3	3	0	NA	NA	NA
Credits_banded = 111.00 – 140.00						
Time	n.risk	n.event	Survival	Std.err	Lower 95% CI	Upper 95% CI
1	188	1	0.995	0.0053	0.9843	1
2	186	2	0.984	0.00917	0.9662	1
3	106	19	0.808	0.03742	0.7375	0.884
4	27	18	0.269	0.07432	0.1567	0.462
5	6	3	0.135	0.06634	0.0512	0.354
Credits_banded = 141.00 – 170.00						
Time	n.risk	n.event	Survival	Std.err	Lower 95% CI	Upper 95% CI
3	269	1	0.996	0.00371	0.989	1
4	45	8	0.819	0.05686	0.715	0.939
5	6	4	0.273	0.15878	0.0874	0.854

The curves of survival confirm an high risk for students in low bands of credits, in particular for the bands <= 20.00 and 21.00 − 50.00 credits.

Similar results can be obtained for the Gender variable, as in Fig. 4 and Table 5.

The performance of female students is always better than male colleagues, in terms of survival probabilities, as shown in Table 5. We now report the performances of students coming from different Italian provinces, in Fig. 5. The greater part of the students arrives from the same region where the faculty is situated, Lombardy. The students resident in the province of Alessandria and Sondrio have underlined positive results, while the ones from the province of Genoa have negative results.

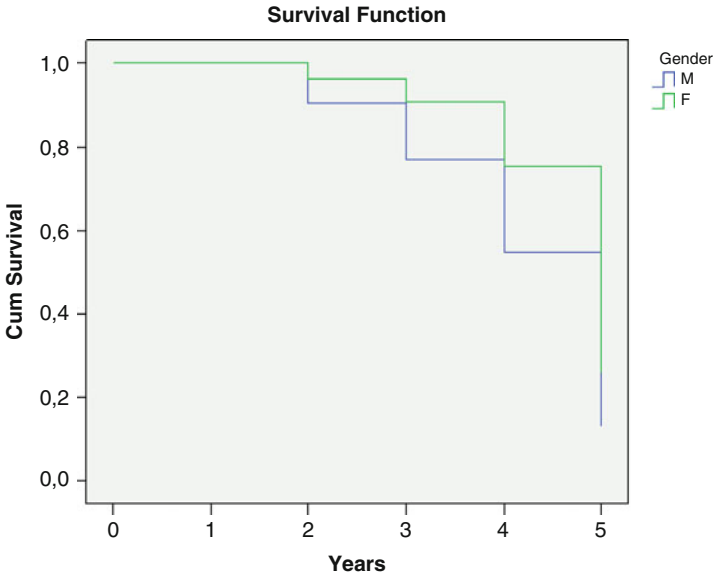


Fig. 4 Survival function for the factor Gender

Table 5 Kaplan Meier estimator for the factor Gender

Gender=Male						
Time	n.risk	n.event	Survival	Std.err	Lower 95% CI	Upper 95% CI
1	143	13	0.909	0.024	0.863	0.957
2	113	15	0.788	0.0357	0.721	0.862
3	73	15	0.626	0.0469	0.541	0.725
4	16	7	0.352	0.082	0.223	0.556
5	4	4	0	NA	NA	NA
Gender=Female						
Time	n.risk	n.event	Survival	Std.err	Lower 95% CI	Upper 95% CI
1	702	24	0.966	0.00686	0.952	0.979
2	594	29	0.919	0.01075	0.898	0.94
3	375	42	0.816	0.01775	0.782	0.851
4	80	30	0.51	0.04553	0.428	0.607
5	17	9	0.24	0.06533	0.141	0.409

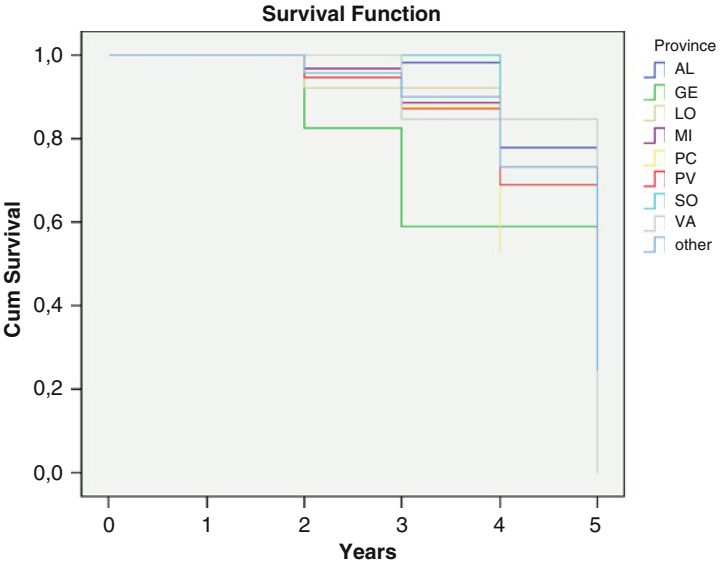


Fig. 5 Survival function for the factor Province

Concerning high school diplomas, Fig. 6 shows that the best results have been noticed for the students coming from high school and teacher's college.

The application of Cox models, model [3], as a multivariate analysis tool to estimate the risk of abandonment, considers the covariates: gender, province of residence, diploma, average mark, sustained credits. The results of the analysis

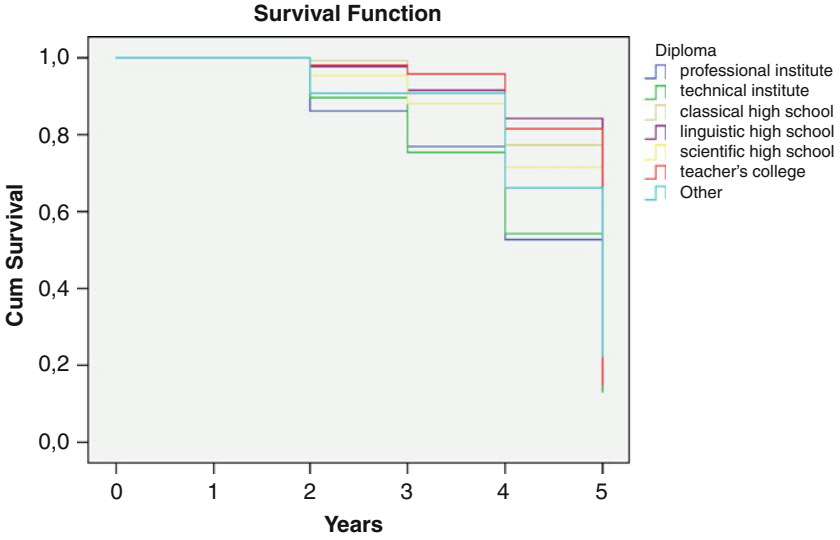


Fig. 6 Survival function for the factor Diploma

confirm the previous findings: the covariates gender, diploma and credits are significant. meaningful. The best performances in comparison to the curve of hazard are recorded for female students, with the maximum of credits and coming from a classical high school and teacher's college. The more subjects to risk are males, from professional institutes and with few credits sustained during the university career. For more examples of this methodology see [11].

We have presented a methodology aimed at individuating, for each degree study, the typologies of students more subject to the risk, considering the characteristics of their careers and socio-demographic variables. The structures of the university in Pavia devoted to the actions of tutoring can thus choose to which students' groups address their activities. An important research development, that we are investigating, concerns the integration between this "objective" data and data coming from students' questionnaires, that measure the subjective perceptions of students. As explained by Giudici [10]. This may allow, when integrated in a statistical model, a more refined predictive tool.

Acknowledgments The Authors acknowledge support from the Centre of Orientation (COR), Prof. Laura Pagani and Prof. Assunta Zanetti for useful discussion and the data. We also acknowledge useful discussion on the methodological aspects with Silvia Figini.

References

1. Anderson, P.K., Borgan, O., Gill, R.D., Keiding, N.: Statistical Models Based on Counting Processes. Springer, New York, NY (1993)
2. Carnell, L J.: The effect of a student-designed data collection project on attitudes toward statistics. J. Stat. Educ. **16**(1) (2008)
3. Cox, D.R.: Regression models and life-tables (with discussion). J. R. Stat. Soc. B. **34**, 187–220 (1972)
4. Cox, D.R., Oakes, D.: Analysis of Survival Data. Chapman and Hall, London (1984)
5. Dawson, R.J.M.: The 'unusual episode' data revisited. J. Stat. Educ. [Online] **3**(3). http://www.amstat.org/publications/jse/v3n3/datasets.dawson.html (1995)
6. Dequarti, E., Figini, S., Giudici, P.: Churn risk mitigation models for students' behaviour. To appear in Ejasa, Electron. J. Appl. Stat. Anal. **2**(1), 37–57 (2009)
7. Dutton, J., Dutton, M.: Characteristics and performance of students in an online section of business statistics. J. Stat. Educ. **13**(30) (2005)
8. Dutton, J., Dutton, M., Perry, J.: Do online students perform as well as lecture students? J. Eng. Educ. **90**, 131–136 (2001)
9. Fleming, T.R., Harrington, D.P.: Counting Processes and Survival Analysis. Wiley, New York, NY (1991)
10. Giudici, P.: Applied Data Mining. Wiley, London (2003)
11. Hurvich, C.M., Tsai, C.-L.: Regression and time series model selection in small samples. Biometrika **76**, 297–307 (1989)
12. Kaplan, E.L., Meier, P.: Nonparametric estimation from incomplete observations. J. Am. Stat. Assoc. **53**, 457–481 (1958)
13. Simonoff, J.S.: The unusual episode and a second statistics course. J. Stat. Educ. **5**(1) (1997)
14. Spooner, F., Jordan, L., Algozzine, B. Spooner, M.: Student ratings of instruction in distance learning and on-campus classes. J. Educ. Res. **92**, 132–140 (1999)
15. Wallace, D.R., Mutooni, P.: A comparative evaluation of world wide web-based and classroom teaching. J. Eng. Educ. **86**, 211–219 (1997)

The Poisson Processes in Cluster Analysis

André Hardy

Abstract This paper aims to review some use of the point processes in cluster analysis. The homogeneous Poisson process is, in many ways, the simplest point process, and it plays a role in point process theory in most respects analogous to the normal distribution in the study of random variables. We first propose a statistical model for cluster analysis based on the homogeneous Poisson process. The clustering criterion is extracted from that model thanks to maximum likelihood estimation. It consists in minimizing the sum of the Lebesgue measures of the convex hulls of the clusters. We also present a generalization of that model to the non-stationary Poisson process, as well as some monothetic divisive clustering methods also based on the Poisson processes. On the other hand, it is usually considered that the central problem of cluster validation is the determination of the best number of natural clusters. We present two likelihood ratio tests for the number of clusters based on the Poisson processes. Most of these clustering methods and tests for the number of clusters have been extended to symbolic data.

1 Introduction

The objective of cluster analysis is to identify a natural structure within a data set, if any. Most of clustering methods are based on the choice of a dissimilarity or a distance between objects. Recently spatial statistics have been increasingly used in cluster analysis, discriminant analysis and pattern recognition. We propose to construct statistical models for cluster analysis based on point processes (Karr (1991)). The simplest point process is one in which points occurs totally randomly. Poisson processes are usually considered as good models of randomness (Cox and Isham [3]). This paper surveys some use of the Poisson processes in cluster analysis. Section 2 presents the starting problem of the research (the estimation of a convex

A. Hardy (✉)
University of Namur, 5000 Namur, Belgium,
e-mail: andre.hardy@fundp.ac.be

B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-13312-1_5,

set), describes the model for cluster analysis based on the homogeneous Poisson process and the corresponding clustering criterion (the Hypervolumes clustering criterion). Section 3 generalizes the statistical model to the non-stationary Poisson process. Section 4 concentrates on two hypotheses tests for the number of clusters based on the Hypervolumes clustering criterion. Section 5 deals with the presentation of monothetic divisive clustering methods based on the Poisson processes. Section 6 reports some concluding remarks.

2 A Statistical Model for Cluster Analysis Based on the Homogeneous Poisson Process

We introduce the definition of the homogeneous Poisson process and the conditional uniformity property of that process. We present the starting problem of the research, the estimation of a convex set, and we show how that result leads to the Hypervolumes clustering method.

2.1 The Homogeneous Poisson Process

We assume that the data are the observation of a n random sample $x_1,\ x_2, \cdots, x_n$ inside some measurable convex domain D (with $0 < m(D) < \infty$) which is a subset of the p-dimensional Euclidean space R^p. m is the Lebesgue measure. Let us denote $X_1, X_2, \cdots, X_n$ the random sample associated with these n data points.

2.1.1 Definition

The Poisson process is a point process in which points occur totally randomly. The homogeneous Poisson process N with intensity q ($q \in R$) on a set $D \subset R^p$ ($0 < m(D) < \infty$) is characterized by the following two properties (Cox and Isham [3])

- If the sets $A_1, A_2, \ldots, A_k \subset D$ are disjoint sets, then the random variables $N(A_1), N(A_2), \ldots, N(A_k)$ are independent ($k \in \mathbb{N}_0$).
- $\forall\ A \subset D, \quad \forall k \in \mathbb{N}_0, \quad P(N(A) = k) = \dfrac{(q\, m(A))^k}{k!}\, e^{-q\, m(A)}$.
  That is for each A, $N(A)$, the random number of points in A, has a Poisson distribution with mean $qm(A)$ where $m(.)$ is the p-dimensional Lebesgue measure.

2.1.2 Conditional Uniformity Property

Given that $N(D) = n$, then the n data points $\{x_1,\ x_2,\ \cdots,\ x_n\}$ are independently and uniformly distributed over D.

This conditional uniformity property allows us to write the density function associated with the homogeneous Poisson process

$$f(x) = \frac{1}{m(D)} \; I_D(x) \tag{1}$$

where I_D is the indicator function of the set D. $I_D(x) = 1$ if $x \in D$ and 0 elsewhere. If $x = (x_1, \; x_2, \cdots, \; x_n)$, the likelihood function L_D takes the form

$$L_D(x_1, \; x_2, \cdots, \; x_n) = \frac{1}{(m(D))^n} \prod_{i=1}^{n} I_D(x_i). \tag{2}$$

2.2 Starting Problem: The Estimation of a Convex Set

The research was initiated with the following problem formulated by D.G. Kendall: "Given a realization of a homogeneous planar Poisson process of unknown intensity within a compact convex set D, find D". When p = 1, that problem is the well-known bus or taxi problem of estimating an interval from which an observed set of data points is assumed to have been drawn uniformly. $H(x_1, \; x_2, \; \cdots, \; x_n)$, the convex hull of sample $x_1, \; x_2, \; \cdots, \; x_n$, is both a sufficient statistics and a maximum likelihood estimate of the domain D. The unbiaised estimate of D proposed by Ripley and Rasson [22] is a dilatation of the convex hull of the points about its centroid. An estimate C of the coefficient of dilatation is given by Moore [17] and can be computed as $C = \sqrt{\frac{n}{n-V_n}}$ where n is the number of points and V_n the number of points on the convex hull.

2.3 The Hypervolumes Clustering Method

The homogeneous Poisson process has the property of complete randomness in two senses: its intensity is constant, and given that $N(A) = n$, the n points are independently and uniformly located in A. For that reason the homogeneous Poisson process is of central importance among point processes, both in theory and in practice. The homogeneous Poisson process also provides a natural starting point for the construction of models for cluster analysis.

Consequently, the challenge was to construct a model for cluster analysis based on the homogeneous Poisson process and on the estimation of a convex set. The Hypervolumes clustering method (Hardy and Rasson [13], Hardy [7]) assumes that the n p-dimensional observation points $x_1, \; x_2, \; ..., \; x_n$ are independant realizations of a homogeneous Poisson process N in a convex domain D included in the Euclidean space R^p (with $0 < m(D) < \infty$). The set D is supposed to be the union of k disjoint convex domains $D_1, \; D_2, \; ..., \; D_k$ (with k fixed; k and D_i are unknown). The problem is to estimate the unknown domains D_i in which the points were generated. We denote by $C_i \subset \{x_1, \; x_2, \; \cdots, \; x_n\}$ the subset of the data points belonging to the domain $D_i \;\; (1 \leq i \leq k)$. The likelihood function L_D of the model can be written as

$$L_{D_1,\ D_2,\ \cdots,\ D_k}(x_1,\ x_2,\ \cdots,\ x_n) = \frac{1}{(m(D_1 \cup D_2 \cup \cdots \cup D_k))^n} \prod_{i=1}^{n} I_D(x_i)$$
$$= \frac{1}{m(D_1 \cup D_2 \cup \cdots \cup D_k)^n} I_D(H(x_1,\ x_2,\ \cdots,\ x_n))$$

where $H(x_1,\ x_2,\ \cdots,\ x_n)$ is the convex hull of the data points $x_1,\ x_2,\ \cdots,\ x_n$.

The maximum likelihood estimates of the k unknown domains $D_1,\ D_2,\ ...,\ D_k$ are the k convex hulls $H(C_i)$ $(i = 1,\ 2,\ ...,\ k)$. The maximization of the likelihood function L_D is equivalent to the minimization of the sum of the Lebesgue measures of the convex hulls of the clusters (Hardy [7]) i.e.

$$\max_{D_1, D_2, \cdots, D_k} L_D(x_1,\ x_2,\ \cdots,\ x_n) \quad \Longleftrightarrow \quad \min_{P \in \mathcal{P}_k} \sum_{i\ =\ 1}^{k} m(H(C_i)) \tag{3}$$

where $\mathcal{P}_k$ is the set of all the partitions of C into k clusters and $P = \{C_1,\ C_2, \cdots,\ C_k\}$.

The Hypervolumes clustering criterion is defined by

$$W_k(P) = \sum_{i\ =\ 1}^{k} m(H(C_i)) = \sum_{i\ =\ 1}^{k} \int_{H(C_i)} m(dx) \tag{4}$$

where $H(C_i)$ is the convex hull of the points belonging to C_i and $m(H(C_i))$ is the multidimensional Lebesgue measure of that convex hull. That clustering criterion has to be minimized over the set of all the partitions of the observed sample into k clusters. So in the context of a clustering problem, we have to find the partition P^* such that

$$P^* = arg \min_{P \in \mathcal{P}_k} \sum_{i\ =\ 1}^{k} \int_{H(C_i)} m(dx). \tag{5}$$

For example, in the one-dimensional euclidean space, convex sets of points are intervals of points and the Lebesgue measure of an interval is its lenght. So, if we fix the number of clusters to k, the clustering problem consists in finding the k intervals, containing all points, such that the sum of the lenghts of the intervals is minimum. In a two-dimensional space, the Lebesgue measure of a set is the area of that set. So we've to find the k sets $C_1,\ C_2,\ \cdots,\ C_k$, containing all points, such that the sum of the areas of the convex hulls $H(C_i)$ is minimum. In a p-dimensional space, the Lebesgue measure of a set is called the hypervolume of that set.

An algorithm (Hardy [8, 10]) has been implemented in a polynomially bounded time. Furthermore the hypervolumes clustering method is not biased towards ellipsoidal or hyperspherical clusters, and it fulfils most of the admissibility conditions of Fisher and Van Ness [6].

3 The Statistical Model Based on the Non-stationary Poisson Process: The Generalized Hypervolumes Clustering Method

The simplest generalization of the homogeneous Poisson process is the non-stationary Poisson process in which the rate is a function $q(x)$ of the location. So the next step of the research was to construct a more general model for cluster analysis, based on the non-stationary Poisson point process.

3.1 The Non-stationary Poisson Process

We give the definition and the main property of the non-stationary Poisson process.

3.1.1 Definition

The non-stationary Poisson process N with intensity $q(.)$ on the measurable domain $D \subset R^p$ $(0 < m(D) < \infty)$ is characterized by the following two properties (Cox and Isham [3])

- If $A_1, A_2, \cdots, A_k$ are arbitrary disjoint sets, then the random variables $N(A_1)$, $N(A_2), \cdots, N(A_k)$ are independent $(k \in \mathbb{N}_0)$.
- $\forall A \subset D$, $N(A)$ has a Poisson distribution with mean $\int_A q(x)\, m(dx)$.

3.1.2 Conditional Property

Given that $N(D) = n$, then the n data points are independently distributed in D, with a density function proportional to $q(x)$.

So the density function associated with the non-stationary Poisson process is

$$f(x) = \frac{q(x)\, I_D(x)}{\int_D q(t)\, m(dt)} = \frac{q(x)\, I_D(x)}{\rho(D)} \tag{6}$$

where $\rho(D) = \int_D q(t)\, m(dt)$ is called the integrated intensity of the process on D. For the homogeneous Poisson process, the intensity q is constant. So $\rho(D) = \int_D q\, m(dt) = q\, m(D)$.

3.2 The generalized Hypervolumes Clustering Method

The generalized Hypervolumes clustering method (Kubushishi [15], Rasson and Granville [19]) assumes that the n p-dimensional points $x_1, x_2, \ldots, x_n$ are generated by a non-stationary Poisson process N with intensity $q(.)$ in a set $D \subset R^p$ $(0 < m(D) < \infty)$ where D is the union of k disjoint convex domains $D_1, D_2, \ldots, D_k$.

The problem is then to estimate the unknown domains D_i in which the points were generated. The likelihood function can be written as

$$L_D(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} \frac{q(x_i) I_D(x_i)}{\rho(D)}$$
$$= \frac{1}{(\rho(D))^n} \prod_{i=1}^{n} q(x_i) I_D(H(x_1, x_2, \cdots, x_n)). \quad (7)$$

The convex hull $H(x_1, x_2, \cdots, x_n)$ is again the maximum likelihood estimate of the convex domain D. The generalized Hypervolumes clustering criterion is deduced from that statistical model, using maximum likelihood estimation.

If the intensity $q(.)$ of the non-stationary Poisson process is known, the maximization of the likelihood function L_D is equivalent to the minimization of the sum of the integrated intensities over the convex hulls of the clusters i.e.

$$\max_{D_1, D_2, \cdots, D_k} L_D(x_1, x_2, \cdots, x_n) \iff \min_{P \in \mathcal{P}_k} \sum_{i=1}^{k} \int_{H(C_i)} q(x_i) m(dx) \quad (8)$$

where $P = \{C_1, C_2, \cdots, C_k\}$ and $\mathcal{P}_k$ is the set of all the partitions of the data points into k clusters. So the maximum likelihood estimates of the unknown convex domains D_i are their convex hulls $H(C_i)$ $(i = 1, \cdots, k)$.

The generalized Hypervolumes clustering criterion W_k^* is defined by

$$W_k^*(P) = \sum_{i=1}^{k} \int_{H(C_i)} q(x) m(dx) \quad (9)$$

where q(.) is the intensity of the non-stationary Poisson process and $H(C_i)$ is the convex hull of the points belonging to C_i $(i = 1, \cdots, k)$.

In the context of a clustering problem, we have to find the partition P^* such that

$$P^* = arg \min_{P \in \mathcal{P}_k} \sum_{i=1}^{k} \int_{H(C_i)} q(x)\, m(dx).$$

3.3 Estimation of the Intensity of the Non-stationary Poisson Process

When the intensity $q(.)$ of the non-stationary Poisson process is not known, it must be estimated. The clustering method described in Sect. 5 in monothetic. So the estimation of the intensity $q(.)$ is to be made in the unidimensional case. We use a non-parametric method: the Kernel method. The Kernel estimate $\hat{q}$ of q is defined by

$$\hat{q}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - x_i}{h}\right) \quad (10)$$

where the Kernel K has the following properties: K is symmetric and continuous, $K \geq 0$ and $\int K(x)dx = 1$. The parameter h is the window width also called "smoothing parameter". The kernel estimate is a sum of "bumps" placed around the observations x_i $(i = 1, 2, \cdots, n)$. The kernel function K generates the shapes of the bumps while the window width h determines their widths. In order to choose h, we distinguish the notions of "bump" and "mode". A mode of a density q is a local maximum of that density. A bump is characterized by an interval such that the density is concave on that interval, but not on a larger interval. Due to its properties, we use a Normal kernel defined by

$$K_N(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}. \tag{11}$$

Silverman [23, 24] proved, in the unidimensional case, that the number of modes for a Normal kernel is a decreasing function of the smoothing parameter h. So for practical purposes we determine h by choosing the first value of h such that $\hat{q}$ remains multimodal.

4 Statistical Tests for the Number of Clusters Based on the Homogeneous Poisson Point Process

The determination of the "true" number of "natural" clusters has often been considered as the central problem of cluster validation. In this section we present two likelihood ratio tests for the number of clusters based on the Hypervolumes clustering criterion: the Hypervolumes test and the Gap test.

4.1 The Hypervolumes Test

The statistical model based on the homogeneous Poisson process allows us to define a likelihood ratio test for the number of clusters (Hardy [8]). Let us denote by $C = \{C_1, C_2, ..., C_\ell\}$ the optimal partition of the sample into ℓ clusters and $B = \{B_1, B_2, ..., B_{\ell-1}\}$ the optimal partition into $\ell - 1$ clusters. We test the hypothesis H_0: $k = \ell$ against the alternative H_1: $k = \ell - 1$, where k represents the number of "natural" clusters ($\ell \geq 2$). The test statistic is deduced from the statistical model for cluster analysis based on the homogeneous Poisson process by applying a likelihood ratio test. The test statistic is defined by

$$S(x_1, x_2, \cdots, x_n) = \frac{W_k(C)}{W_{k-1}(B)} \tag{12}$$

where $W_k(C)$ (respectively, $W_{k-1}(B)$) is the value of the Hypervolumes clustering criterion associated with the best partition into k (respectively, $k - 1$) clusters.

Unfortunately the sampling distribution of the statistic S is not known. The first solution to that problem is to consider an interesting property of the test statistic: $S(x_1, x_2, \cdots, x_n)$ belongs to $[0, 1[$. For practical purposes, we can use the

following decision rule: reject H_0 if S is "close to" 1. We apply the test in a sequential way: if ℓ_0 is the smallest value of $\ell \geq 2$ for which we reject H_0, we choose $\ell_0 - 1$ as the best number of "natural" clusters. More recently Hardy and Blasutig [12] used permutation tests to obtain an approximate distribution for the test statistic S.

4.2 The Gap Test

The Gap test (Kubushishi [15], Rasson, Kubushishi [20]) uses the model for cluster analysis based on the homogeneous Poisson process. We test H_0 : the n observed points are a realization of a Poisson process in D against H_1: n_1 points are a realization of a homogeneous Poisson process in D_1 and n_2 points in D_2 where $D_1 \cap D_2 = \emptyset$ and $n_1 + n_2 = n$. The sets $D,\ D_1,\ D_2$ are convex and unknown. Let us denote by C (respectively, C_1 , C_2) the set of points $\{x_1,\ x_2,\ \cdots,\ x_n\}$ belonging to D (respectively, D_1, D_2). The test statistic is given by (Kubushishi [15])

$$Q(x_1, x_2, \cdots, x_n) = \left(1 - \frac{m(\triangle)}{m(H(C))}\right)^n \tag{13}$$

where $H(C)$ is the convex hull of the points belonging to C, $\triangle = H(C) \setminus (H(C_1) \cup H(C_2))$ is the "gap space" between the clusters and m is the multidimensional Lebesgue measure. So the test statistic is the Lebesgue measure of the gap space between the clusters.

The decision rule is the following: reject H_0, at level α, if (asymptotic distribution)

$$\frac{nm(\triangle)}{m(H(C))} - \log n - (p-1)\log\log n - log\kappa \geq -\log(-\log(1-\alpha)) \tag{14}$$

where the constant κ depends on the shape of the convex domain (Janson [14], Deheuvels et al. [4], Kubushishi [15]).

The Hypervolumes test and the Gap test have been applied to numerous data sets with various data structures (Hardy and Beauthier [11]). These tests were also compared to well-known stopping rules for the number of clusters available in the scientific literature (Milligan and Cooper [16]).

5 Monothetic Divisive Clustering Methods Based on the Poisson Processes

Polythetic partitioning methods are usually computationally complex. Therefore Pirçon [18] proposed five monothetic divisive clustering procedures. One of them is based on the homogeneous Poisson process and the others on the non-stationary Poisson process. We present here the method UNHOPPKI (Unique Non HOmogeneous Poisson Process with Kernel Intensity).

5.1 The Model

We consider the statistical model for cluster analysis based on the non-stationary Poisson process. So we suppose that the n observed data points $x_1, x_2, \cdots, x_n$ are generated by a non-stationary Poisson process of intensity $q(.)$ in a set $D \subset R^p$, where D is the union of k disjoint unknown convex domains $D_1, D_2, ..., D_k$. The estimation of the intensity $q(x)$ of the Poisson process is made by the kernel method with a Normal kernel.

5.2 The Splitting Process

UNHOPPKI is an unsupervised monothetic divisive clustering method. The splitting part of the process is analogue to the classical CART algorithm (Breiman et al. [2]). At each step we split a cluster C into two subclusters C_1 and C_2, which minimize the integrated intensity $WG_2 = \rho(H(C_1)) + \rho(H(C_2))$ on the convex hulls of the two clusters, or equivalently which maximizes the integrated intensity $\rho(\Delta)$ on the gap space of the clusters. The hierarchic divisive procedure is the following

- For each variable Y_i $(i = 1, \cdots, p)$ and each cluster C_j $(j = 1, \cdots, k)$,
  - compute the estimate $\hat{q}(x)$ of $q(x)$ (given in Sect. 3.3);
  - find the best bipartition of C into C_1 and C_2 such that $C = C_1 \cup C_2$ and $\rho(\Delta) = \int_\Delta \hat{q}(x)m(dx)$ is maximal;
- choose the cluster and the variable such that the likelihood function is maximal;
- cut the cluster and repeat the procedure until a stopping rule is fulfilled.

The stopping rule is the number of points in a node. That parameter has to be fixed by the user.

5.3 The Pruning Method

At the end of the splitting process, a complete tree is obtained. No statistical test has been performed as yet in order to test if the splits are statistically significant. So a pruning process is applied in order to obtain a useful tree. The Gap test is applied at each node in order to test

- H_0 : the points are distributed in only one domain D
- H_1 : the points are distributed in two domains D_1 and D_2 $\quad (D_1 \cap D_2 = \emptyset)$.

When the null hypothesis is not rejected, we conclude that the split is not statistically acceptable (bad split). On the other hand, if the null hypothesis is rejected, we consider that the split is statistically acceptable (good split). At the end of the process we adopt the following rule: cut all the branches that contain only bad splits.

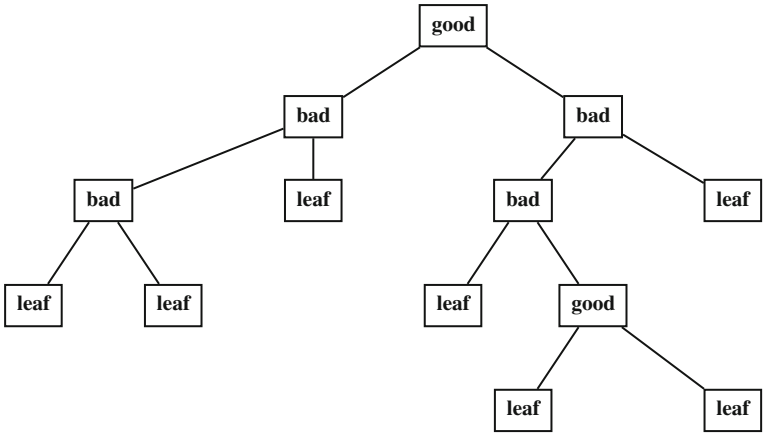
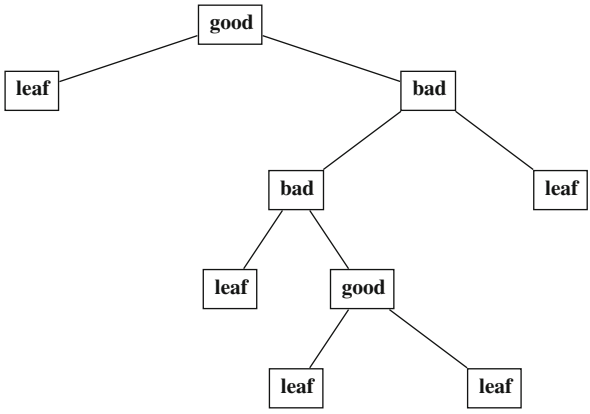


Fig. 1 Tree before pruning

Fig. 2 Tree after pruning



For example, suppose that we obtain the following tree at the end of the splitting process (Fig. 1).

After the pruning process we obtain the following tree (Fig. 2).

5.4 The Merging Process

In some cases, for example when the clusters are not linearly separables, it is not possible for the method to recover the natural structure of the data at the end of the pruning step. That's why we've added a merging step after the pruning step. Additional tests are made on the pairs of clusters not belonging to the same node. Once more, we use the Gap test.

6 Conclusion

The Poisson processes are usually considered as good models of randomness. So the goal of that paper was to review some uses of these processes in order to construct a natural model for cluster analysis. We have presented several clustering methods based on the Poisson processes as well as two hypotheses tests for the number of clusters. Let us mention that some of these methods have been extended to symbolic interval variables. For example SCLASS (Rasson et al. [21]) is an extension to interval variables of the method UNHOPPKI. The clustering criterion used in SCLASS is a symbolic extension of the generalized Hypervolumes clustering criterion. The Hypervolumes test and the Gap test have also been extended to interval variables. The extended Gap test is used in the SCLASS procedure. The method SCLASS is available in the SODAS 2 software.

References

1. Bock, H.-H., Diday, E. (eds.): Analysis of Symbolic Data, Exploratory Methods for Extracting Statistical Information from Complex Data. Studies in Classification, Data Analysis and Knowledge Organisation. Springer, Heidelberg (2000).
2. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. Wadsworth, Monterey, CA (1984).
3. Cox, D.R., Isham, V.: Point Processes. Chapman and Hall, London (1980)
4. Deheuvels, P., Einmahl, J.H.J., Mason, D.M.: The almost sure behavior of maximal and minimal multivariate kn-spacings. J. Multivar. Anal. **24**, 155–176.
5. Diday, E., Noirhomme-Fraiture, M. (eds.): Symbolic Data Analysis and the Sodas Software. Wiley, Chichester (2008)
6. Fisher, L., Van Ness, J.W.: Admissible clustering procedure. Biometrika **58**(1), pp. 91–104 (1971)
7. Hardy, A.: Statistique et classification automatique: un modèle, un nouveau critère, des algorithmes, des applications. PhD thesis, University of Namur, Namur, Belgium (1983).
8. Hardy, A.: A heuristic approach for the hypervolumes method in cluster analysis. Jorbel **36**(1), 43–55 (1996)
9. Hardy, A.: On the number of clusters. Comput. Stat. Data Anal. **23**(1), 83–96 (1996)
10. Hardy, A.: Validation of a clustering structure: determination of the number of clusters. In: Diday, E., Noirhomme-Fraiture, M. (eds.) Symbolic Data Analysis and the Sodas Software, pp. 235–262. Wiley, Chichester (2008)
11. Hardy, A., Beauthier, C.: Comparaison entre le test des Hypervolumes et le Gap test. Research report. University of Namur, Namur, Belgium (2004)
12. Hardy, A., Blasutig, L.: Application des tests de permutation au critère des Hypervolumes en classification automatique. Research report. University of Namur, Namur, Belgium (2007)
13. Hardy, A., Rasson, J.P.: Une nouvelle approche des problèmes de classification automatique. Stat. Anal. Données **7**(2), 41–56 (1982)
14. Janson, S.: Random coverings in several dimensions. Acta Math. **156**, 83–118 (1986)
15. Kubushishi, T.: On some Applications of Point Process Theory in Cluster Analysis and Pattern Recognition. PhD thesis, University of Namur, Namur, Belgium (1996)
16. Milligan, G.W., Cooper, M.C.: An examination of procedures for determining the number of clusters in a data set. Psychometrika **50**(2), pp. 159–179 (1985)
17. Moore, M.: On the estimation of a convex set. Ann. Stat. **12**, 1090–1099 (1984)

18. Pirçon, J.-Y.: La classification et les processus de Poisson pour de nouvelles méthodes monothétiques de partitionnement. PhD thesis, University of Namur, Namur, Belgium (2004)
19. Rasson, J.P., Granville, V.: Geometrical tools in classification Comput. Stat. Data Anal. **23**, 105–123 (1996)
20. Rasson, J.P., Kubushishi, T.: The gap test: an optimal method for determining the number of natural classes in cluster analysis. In: Diday, E. et al. (eds.) New approaches in classification and data analysis, pp. 186–193. Springer, Paris (1994)
21. Rasson, J.P. et al.: Unsupervised divisive classification. In: Diday, E., Noirhomme, M. (eds.) Symbolic Data Analysis and the Sodas Software. Wiley, Chichester (2008)
22. Ripley, B.D., Rasson, J.P.: Finding the edge of a Poisson forest. J. Appl. Probab. **14**, 483–491 (1977)
23. Silverman, B.W.: Using kernel density estimates to investigate multimodality. J. R. Stat. Soc. Ser. B **43**, 97–99 (1981)
24. Silverman, B.W.: Density estimation for statistics and data analysis. Chapman and Hall, London (1986)

TWO-CLASS Trees for Non-Parametric Regression Analysis

Roberta Siciliano and Massimo Aria

Abstract This paper shows that a regression tree problem can be turned into a classification tree problem reducing the computational cost and providing useful interpretation issues. A TWO-CLASS tree methodoloy for non-parametric regression analysis is introduced. Data are as follows: a numerical response variable and a set of predictors (of categorical and/or numerical type) are measured on a sample of objects, with no probability assumption. Thus a non-parametric approach is proposed. The concepts of prospective and retrospective splits are considered. Main idea is to grow a binary partition of the sample of objects such that, at each node of the tree structure, the numerical response is recoded into a dummy or two-class variable (called theoretical response) on the basis of the optimal partition of the objects into two groups within the set of retrospective splits. A two-stage splitting criterion with a fast algorithm is applied: the best split of the objects is found in the set of candidate (prospective) splits of each predictor modalities by maximizing the predictability of the two-class response. Some applications on real world cases and a simulation study allow to demonstrate that the two-class splitting procedure is computationally less intensive than standard regression tree such as CART. Furthermore, the final partitions obtained by the two-class procedure and the standard one are very similar to each other, in terms of percentage of objects belonging together to the same terminal node. Some aids to the interpretation allow to describe the response variable distribution in the terminal nodes.

1 Previous Work

This paper deals with tree-based methods, in particular binary segmentation or exploratory trees [4, 10], namely recursive partitioning of a sample of units on the basis of a set of predictors such to obtain subgroups where a response variable is internally homogeneous and externally heterogeneous. The attention is focalized on regression trees, considering as benchmarking CART [Breiman et al., [2]], as

R. Siciliano (✉)
Department of Mathematics and Statistics, University of Naples Federico II, Naples, Italy,
e-mail: roberta@unina.it

B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-13312-1_6,

extension of AID [9]. Fast splitting algorithm [7, 11] will be also considered within the two-stage methodology [6, 8, 13]. The result is either a *classification tree* when the response variable is categorical or a *regression tree* when the response variable is numerical. Usually binary trees are built up. Any segmentation procedure is characterized by a splitting criterion, a stopping rule and an assignment rule of the response class/value to each terminal node. Exploratory tree allows to describe the dependence relation between the response variable and the predictors. In order to predict the response class/value of new units on which only measurements of predictors are known alternative confirmatory approaches can be considered, such as pruning and selection by test sample or cross-validation estimates as well as ensembles methods [4].

AID [9] was the pioneer work to grow regression trees. The splitting criterion was based on the between group deviation of the response variable, predictors were only of categorical type. CART [2] extended the procedure to any type of variables, introduced the decrease of impurity measure, added pruning and selection for prediction purpose. Two-stage methodology [6, 13] defined the global and the local predictability measures to select at each node the best predictor and the best split respectively. Fast splitting algorithms [7, 11], that are based on suitable mathematical properties of two-stage splitting criterion, allowed to find the best split without trying out all candidate splits and thus saving computational cost. In place of predictability measures, there have been considered statistical models such as factorial method [12], discriminant analysis [8], logistic regression [5].

2 TWO-CLASS Trees Methodology

Let $\mathbf{L} = \{\mathbf{y}, \mathbf{X}\}$ be a learning sample, where the N-vector $\mathbf{y}$ includes either the observations of the numerical response variable Y (in case of regression trees) or the classes of a categorical response variable (in case of classification trees) and the matrix $\mathbf{X}$ includes N row vectors $\mathbf{x}'_n = (x_{n1}, \ldots, x_{nM})'$ of measurements of M predictors $(X_1, \ldots, X_M)$ of a numerical or categorical type, with N the number of observed objects or cases. Let $i_Y(t)$ be the impurity measure of the response variable Y at node t, describing how similar the objects into the node are to each other, the smaller the number of the impurity measure the less impure the group of objects is. For a numerical response variable, the variation measure can be considered: the smaller the variation of Y is the less impure the group of objects is. For a categorical response variable Y, the heterogeneity index of Gini can be considered analogously. In the recursive partitioning, the best split of the objects at any node t is found maximizing the decrease of impurity of Y sending a percentage p_{t_l} of objects from the node t to the left subnode t_l and a percentage p_{t_r} of objects to the right subnode t_r:

$$max_s \Delta_Y(s,t) = i_Y(t) - [i_Y(t_l)p_{t_l} + i_Y(t_r)p_{t_r}] \tag{1}$$

where s is any splitting variable or dummy variable defining the split of the objects into two sub-groups. In case of regression trees the (1) is equivalent to maximizing the between-group deviation of Y in the two subnodes or Pearson correlation coefficient. In case of classification trees the (1) is equivalent to maximizing a dependence measure in two-way cross-classifications, namely the predictability τ index of Goodman and Kruskal.

One point arises: which is the set of splits to be tried out? We distinguish between prospective and retrospective splits of the objects at a given node.

A *prospective split* is any split s of the objects induced by splitting the predictor modalities. As an example, an object goes either to the left sub-node if $X \leq c$ or to the right sub-node if $X > c$. Standard tree-growing procedure adopts prospective splits. Let S denote the set of prospective splits, considering all sets of splits deduced by the predictors.

A *retrospective split* is any split s of the objects induced by splitting the response modalities, without caring for the predictors. If Y is numerical then any cut point of the real interval in which the Y is defined yields a retrospective split. Let K denote the set of retrospective splits. It can be shown that $S \subseteq K$.

We define the *optimal split* of the objects into two sub-groups the split s_{opt} that maximizes the decrease of impurity (1) over all possible *retrospective splits* in the set K. We define the *best split* of the objects into two sub-groups the split s_{best} that maximizes the decrease of impurity (1) over all possible *prospective splits* in the set S.

The optimal split s_{opt} can be theoretical since it can be not necessarily generated by any prospective split of the predictor modalities. The $\Delta_Y(s_{opt}, t)$ is the upper bound of the decrease of impurity that can be reached, as $\Delta_Y(s_{best}, t) \leq \Delta_Y(s_{opt}, t)$, so that the ratio $\Delta_Y(s_{best}, t)/\Delta_Y(s_{opt}, t)$, ranging from zero and one, is an efficiency measure of the best split found at a given node, saying how good is the discrimination between the two sub-groups in terms of the response distribution into the two subnodes.

TWO-CLASS trees for regression defines, at each node, a dummy or two-class response (theoretical) variable Y_{opt} describing the optimal split s_{opt} of the objects; then, two-stage splitting criterion for classification trees using the fast algorithm can be applied in order to find the best (prospective) split s_{best} of the objects.

The partitioning algorithm **TWO-CLASS Tree** is formed by the following steps:

- *Step 1*. The domain K of retrospective splits of Y is generated;
- *Step 2*. The best retrospective split s_{opt} is identified maximizing the decrease of impurity over the set K, where as impurity measure the variation is considered; the split obtained yields to define the theoretical distribution Y_{opt} with two-classes;
- *Step 3*. The domain S of prospective splits is generated considering the predictors X;
- *Step 4*. The best prospective split s_{best} is identified minimizing the decrease of impurity over the set S, where as impurity measure the Gini index of heterogeneity is considered and the fast algorithm is used;

– *Step 5*. The partition of original Y is obtained through the split of objects due to s_{best}.

The algorithm iterates, at each node, up to reaching stopping rules.

The quality of the split can be evaluated by the above mentioned efficiency measure. As a result, an exploratory tree is built up, its terminal nodes define the final partition of the objects. The quality of the overall tree can be evaluated in terms of Relative Mean Square Error with respect to the root node.

3 Comparative Study

Comparative study of TWO-CLASS Trees methodology has been worked out considering real data sets and a simulation plan. The benchmarking methodology is the CART methodology for regression trees. The final exploratory trees have been compared in terms of Relative Root Mean Square Error.

The following real data sets from a well-known archive of UCI repository have been considered: Boston Housing, Automobile, Auto MPG, Forest Fire and Concrete Compressive Strength [1].

The simulation study has been planned such to consider five typologies of relationship between predictors and the response variable. Predictors have been generated from different probability distributions, such as uniform, binomial, normal, chi-square. The response variable has been generated as linear and nonlinear function of the predictors, with an error perturbation generated by either uniform or normal distribution. The dependent links for each simulation are formally defined in Table 1 and graphically represented in Fig. 3.

The last column of Table 2 and the second last of Table 3 show the results of a similarity measure, namely the percentage of objects that fall into the same group considering the final partitions obtained by the TWO-CLASS tree and the CART regression tree. In other words, it counts how many objects fall into terminal nodes characterized by the same splits into both final partitions, although the splits can be not ordered in the same sequence. As an example, in boston housing the final partitions of TWO-CLASS tree (induced by the optimal classification of the objects) and CART tree are similar for the 87.20% of objetcs, as also shown in the scatter plots using different colours for the objects belonging to each group of the partition.

Table 1 Dependence functions for Y response (The letters c, k and h indicate random positive values)

Simulation	Dependent Link
Simulation 1	$Y = \sin(X_1 + k \cdot X_4) + error$
Simulation 2	$Y = \left(k \cdot X_2 + h \cdot \sqrt{X_5}\right)^2 + error$
Simulation 3	$Y = \left(k \cdot X_1 + h \cdot e^{X_3} + h \cdot X_5\right)^2 + error$
Simulation 4	$Y = 1 + \sin\left(k \cdot X_1 + h \cdot e^{X_3} + h \cdot X_5\right)^2 + error$
Simulation 5	$Y = \left[c + \log(X_5 + h \cdot X_3) - h \cdot X_2 + (1 - k) \cdot X_1 - X_4\right] + error$

Table 2 Comparison study: main results about UCI archives

UCI Repository Datasets			TWO-CLASS Tree		CART		
Archive Name	Numeric Variable	Categorical Variable	Number of Term. Nodes	Relative RMSE	Number of Term. Nodes	Relative RMSE	Similarity Measure
Boston Housing	12	1	18	39.01%	19	40.02%	87.20%
Automobile	14	9	19	34.00%	22	30.00%	91.50%
Auto MPG	7	1	14	37.49%	10	44.44%	74.80%
Forest Fire	8	2	30	62.28%	30	64.21%	89.20%
Concrete Compressive Strength	8	0	16	50.74%	16	50.71%	80.30%

Table 3 Comparison study: main results about simulated datasets

			TWO-CLASS Trees		CART Algorithm		Comparison	
Simulated	Numeric	Categorical	Number of	Relative	Number of	Relative	Similar	Comp. Cost
Datasets	Variables	Variables	Term. Nodes	RMSE	Term. Nodes	RMSE	Groups	Reduction
Simulation 1	6	0	21	22.01%	19	22.30%	98.24%	-27.96%
Simulation 2	6	0	23	24.01%	24	24.59%	95.76%	-29.63%
Simulation 3	0	6	18	58.52%	17	59.28%	94.30%	-64.21%
Simulation 4	0	6	16	42.56%	16	42.54%	96.45%	-59.02%
Simulation 5	4	5	17	31.21%	18	34.67%	90.04%	-41.74%

The quality of both the TWO-CLASS tree and the CART Tree can be evaluated in terms of Relative Mean Square Error with respect to the root node. In Table 3 the computational cost reduction of TWO-CLASS Tree over CART Tree in the simulations.

Figure 1 shows for the dataset Boston Housing a comparison of the final partition obtained through the methodologies TWO-CLASS Tree (sub-figure on the left) and CART (sub-figure on the right). In order to graphically visualize the partition and the binary splits made to every step of the algorithm, the classification has been made in both cases, considering only two predictors. It is possible to deduce that the final clusters tend to have an high similarity. You can deduce it also thanks to Fig. 2 which shows a comparison between the two methodologies for the dataset Forest Fire. In the first four levels of the tree the splits coincide. These explain the most important relations between the response and the predictors.

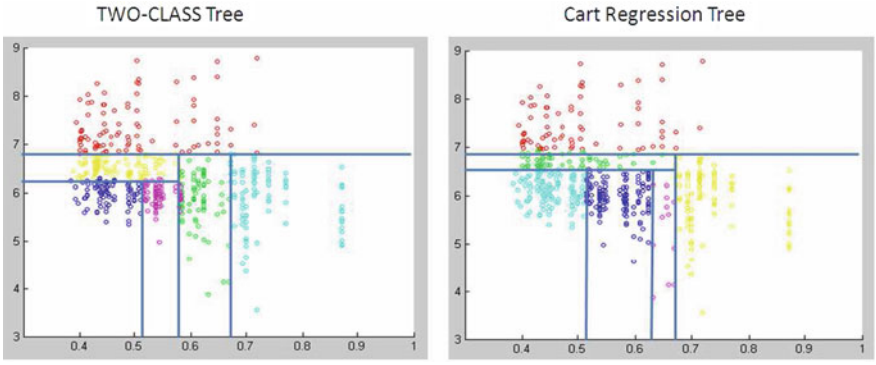


Fig. 1 Final partitions by TWO-CLASS Trees and CART approaches (Boston Housing Dataset)

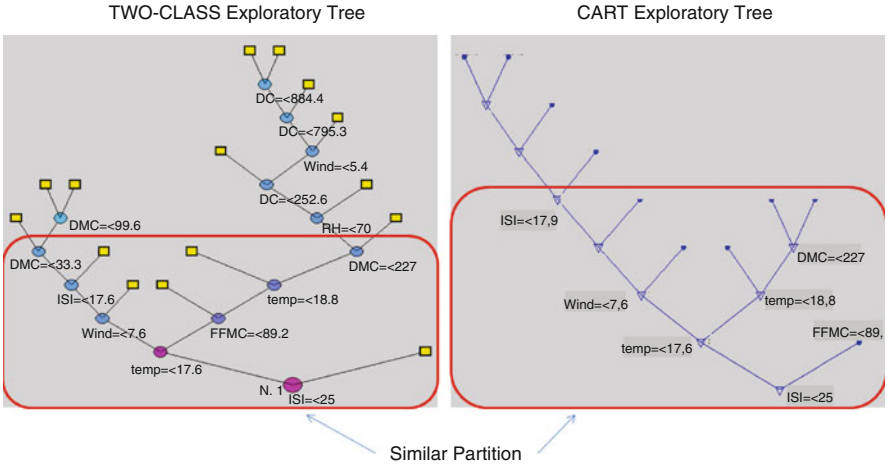


Fig. 2 Trees comparision in terms of similar splits (Forest Fire UCI archive)

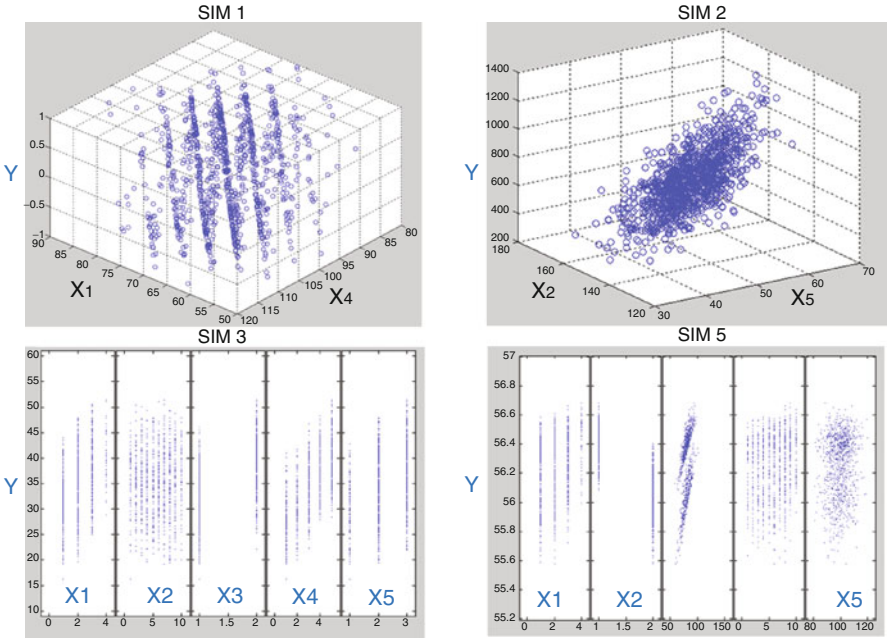


Fig. 3 Simulated datasets: Scatterplots of Sim1, Sim2, Sim3 and Sim5

It can be noticed that although the final partition is obtained through a splitting procedure for classification, in the terminal nodes the numerical response distribution can be described as well, calculating the average and the mean square error.

4 Concluding Remarks

This paper has provided a tree-based methodological framework to non-parametric regression analysis. A regression tree problem has been turned into a two-class partitioning tree procedure. This has been possible through the use of prospective and retrospective splits. Splitting criterion has been based on decrease of impurity, other approaches such as statistical modelling could be considered as well. TWO-CLASS trees have been shown to provide exploratory trees with a very high percentage of objects classified in the same way as in the CART regression trees. Main advantage of the proposed approach is computationally and interpretative: the best split at each node is found using a fast algorithm and, in addition, its quality can be evaluated by an efficiency measure. As a result, it improves the quality of the final partition and decreases the computational cost. TWO-CLASS trees framework can be fruitfully considered for robust tree-based missing data imputation [3] as well as for three-way trees [15]. The basic method has been implemented in MATLAB environment, enriching the Tree Harvest Software [14].

Acknowledgments Financial support from European FP6 Project iWebCare IST-4-02-8055 (Scientific Responsible: Prof. Roberta Siciliano).

References

1. Asuncion, A., Newman, D.J.: UCI Machine Learning Repository, Irvine, CA, University of California, School of Information and Computer Science. http://www.ics.uci.edu/∼mlearn/MLRepository.html (2007)
2. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees, Wadsworth, Belmont, CA (1984)
3. Hastie, T., Friedman, J.H., Tibshirani, R.: The Elements of Statistical Learning. Springer, New York, NY (2001)
4. D'Ambrosio, A., Aria, M., Siciliano, R.: Robust Tree-Based Incremental Imputation Method for Data Fusion. Advances in Intelligent Data Analysis, Springer, Berlin Heidelberg (2007)
5. Mola, F., Klaschka, J., Siciliano, R.: Logistic Classification Trees. In: A. Prat (ed.): Proceedings in Computational Statistics: COMPSTAT '96 (Barcellona), pp. 373–378. Physica-Verlag, Heidelberg (D) (Aug 24–28, 1996)
6. Mola, F., Siciliano, R.: A two-stage predictive splitting algorithm in binary segmentation. In: Dodge, Y., Whittaker, J. (eds.) COMPSTAT'92 Proceedings, pp. 179–184. Physica Verlag, Heidelberg (1992)
7. Mola, F., Siciliano, R.: A Fast Splitting Procedure for Classification and Regression Trees, Statistics and Computing, vol. 7, pp. 208–216. Chapman Hall, New York, NY (1997)
8. Mola, F., Siciliano, R.: Discriminant analysis and factorial multiple splits in recursive partitioning for data mining. In: Roli, F., Kittler, J. (eds.) Proceedings of International Conference on Multiple Classifier Systems, Lecture Notes in Computer Science, pp. 118–126. Springer, Heidelberg (2002)
9. Morgan, J.N., Sonquist, J.A: Problem in the analysis of survey data and a proposal. J. Am. Stat. Assoc. **58** (1963)
10. Siciliano, R.: Exploratory versus decision trees. In: Payne, R., Green, P. (eds.) COMPSTAT '98 Proceedings, pp. 113–124. Physica-Verlag, Heidelberg (1998)
11. Siciliano, R., Mola, F.: A fast regression tree procedure. In: Forcina, A. et al. (eds.) Statistical Modeling. Proceedings of the 11th International Workshop on Statistical Modeling, pp. 332–340. Graphos, Perugia (1996)
12. Siciliano, R., Mola, F.: Ternary classification trees: a Factorial Approach. In: Greenacre, M., Blasius, J. (eds.) Visualization of Categorical Data, pp. 311–323, cap. 22. Academic Press, San Diego, CA (1998)
13. Siciliano, R., Mola, F.: Multivariate Data Analysis through Classification and Regression Thees, Computational Statistics and Data Analysis, vol. 32, pp. 285–301. Elsevier Science, Amsterdam (2000)
14. Siciliano, R., Aria, M., Conversano, C.: Tree Harvest: methods, software and applications. In: Antoch, J. (ed.) COMPSTAT 2004 Proceedings, pp. 1807–1814. Springer, Berlin Heidelberg (2004)
15. Tutore, V.A., Siciliano, R., Aria,M.: Conditional Classification Trees using Instrumental Variables. Advances in Intelligent Data Analysis, pp. 163–173. Springer, Berlin Heidelberg (2007)

Part II
Classification and Discrimination

Efficient Incorporation of Additional Information to Classification Rules

Miguel Fernández, Cristina Rueda, and Bonifacio Salvador

Abstract We propose and discuss improved classification rules when a subset of the predictors is known to be ordered. We compare the performance of the new rules with other standard rules in a restricted normal setting using simulation experiments and real data exposing their good performance.

1 Introduction

Consider the classical discrimination problem with two populations Π_1 and Π_2 and a new observation $z = (z_1, \ldots, z_k)$ that has to be classified in one of the two populations. Let P_1 and P_2 be the distributions of Z in Π_1 and Π_2, and f_1 and f_2 the corresponding density functions. If we consider the 0–1 loss function and equal a priori probabilities for each population the Bayes rule can be written as:

$$\text{Classify } z \text{ in } \Pi_1 \text{iff } f_1(z) \geq f_2(z).$$

In applications it is usual that some additional information is available. It is frequent that this information tells us that the observations from one of the populations, for example Π_1, take higher (or lower) values that those coming from the other, i.e. Π_2. For example, suppose that we are interested in discriminating among populations Π_1 that corresponds to cirrhotic patients, and Π_2 corresponding to non-cirrhotic patients. From the literature it is known (see [2]) that Hyaluronic acid (HA) levels are increased in chronic liver diseases, in particular in patients with cirrhosis, so that HA levels correlate with clinical severity, and that the same happens with the levels of β-Glucuronidase activity and N-acetyl-β-d-glucosaminidase activity.

In [4] and [6] we have proposed two different ways of defining sample rules that take into account this additional information and have lower total misclassification probability (TMP) than the classical rules that do not consider this information. The

M. Fernández (✉)
Departamento de Estadística, Universidad de Valladolid, 47005 Valladolid, Spain
e-mail: miguelaf@eio.uva.es

B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-13312-1_7,

purpose of this work is not only to rejoin the work we have done in those papers, but to take one step further and show how the different ways of incorporating the additional information appearing there can be combined to obtain new rules that outperform the ones previously defined.

In Sect. 2 we describe the different ways of defining restricted rules and present the new ones. In Sects 3 and 4 we show their performance in a simulation study and in real data. Sect. 5 briefly summarizes the final conclusions.

2 Discrimination Rules That Incorporate Additional Information

Although normality is not an essential condition for our rules, let us assume normal distributions $N_k(\mu_i, \Sigma), i = 1, 2$ and equal a priori probabilities for the populations, that is the two populations have equal covariances matrices and all parameters are unknown. In this way the usual Fisher's rule is:

$$\begin{aligned} &\text{Classify } z \text{ in } \Pi_1 \text{iff } f_{N_k(\hat{\mu}_1, S)}(z) \geq f_{N_k(\hat{\mu}_2, S)}(z) \text{ or equivalently,} \\ &\text{Classify } z \text{ in } \Pi_1 \text{iff } (z - (c_1\hat{\mu}_1 + c_2\hat{\mu}_2) + c\hat{\delta})' S^{-1}\hat{\delta} \geq 0, \end{aligned} \tag{1}$$

where $c_i = n_i/(n_1 + n_2)$, $c = (c_1 - c_2)/2$, n_i is the training sample size from population Π_i, S is the sample covariance matrix and $\hat{\mu_i}$, $\hat{\delta}$ are the unrestricted MLEs of μ_i and $\delta = \mu_1 - \mu_2$.

Here we will also assume that the additional information tells us that observations from Π_1 tend to take higher values in each predictor than those coming from Π_2.

We will use a two parameter notation, $R(\lambda, \gamma)$, for the rules considered in this work. First parameter tells us what latent space set (see Sect. 2.2 below) has been used in the definition of the rule with 0 meaning no latent space set used. The second parameter indicates what estimator of the means is being considered (see Sect. 2.1). When this second parameter is equal to "U" it means that an unrestricted estimator of the means is being considered. In this way Fisher's rule is denoted as $R(0, U)$.

2.1 Restricted Parameter Estimation

The most direct approach, developed in [4], is to transform the additional information in restrictions between the parameters. For our case this implies $\mu_1 \geq \mu_2$ (or $\delta \geq 0$) coordinatewise. This line of work appears in [5] under some simplifying assumptions. We could use directly the restricted MLE (RMLE) of $\hat{\delta}$ in Fisher's rule. The RMLE of δ is $\delta^* = p_{S^{-1}}\left(\widehat{\theta}_1 - \widehat{\theta}_2 \,/\, O_k^+\right)$, i.e. the projection of the estimated difference of means on the positive orthant cone $O_k^+ = \left\{x \in \mathbf{R}^k : x_i \geq 0, i = 1, \ldots, k\right\}$ using the metric given by S^{-1}. However, we have found that better results can be obtained using (1) and replacing $\hat{\delta}$ by an estimator of the difference of means δ that belongs to the interior of its restrictions cone (the positive orthant) with probability 1. This may be related to the well known fact that

the estimators belonging with positive probability to the frontier of the parametric space are not admissible (cf. [7]). Our new estimators δ_γ^* are defined as the limit when $i \to \infty$ of the following iterative procedure:

$$\widehat{\delta}_\gamma^{(i)} = p_{s-1}\left(\widehat{\delta}_\gamma^{(i-1)} / O_k^+\right) - \gamma p_{s-1}\left(\widehat{\delta}_\gamma^{(i-1)} \Big/ O_k^{+P}\right), i = 1, 2, \ldots \qquad (2)$$

where $\widehat{\delta}_\gamma^{(0)} = \delta^*$, $O_k^{+P} = \left\{y \in \mathbf{R}^k : y'x \leq 0, x \in O_k^+\right\}$ is the polar cone of O_k^+ and $0 \leq \gamma \leq 1$ is a parameter indicating how much into the cone is the estimator. Notice that $\gamma = 0$ corresponds to the RMLE.

Then the classification rule $R(0, \gamma)$ is

$$\text{Classify } z \text{ in } \Pi_1 \text{iff } (z - (c_1\hat{\mu}_1 + c_2\hat{\mu}_2) + c\delta_\gamma^*)' S^{-1}\delta_\gamma^* \geq 0.$$

The convergence of procedure (2), several good properties of estimators δ_γ^* and a proof for the fact that rules $R(0, \gamma)$ have lower TMP than Fisher's rule under mild conditions may be found in [4].

2.2 Latent Space Rules

A more involved approach that leads to a new theoretical rule, and with a good performance in practice, appears in [6]. The point is to use a $2k$ dimensional latent space of nonobservable values to introduce the additional information directly in the rule formulation. The latent space is derived assuming that for each individual there are two vector values (S_1, S_2) that correspond to observations on the response vector, under Π_1 and Π_2. In particular, here we assume that the marginal densities for S_1 and S_2 are $N_k(\mu_i, \Sigma), i = 1, 2$.

Going back to our cirrhosis example, the latent space represents pairs of HA, B_1 and B_2 values for the same patient that would have been observed under the presence or absence of cirrhosis. The values in this space are non observable because each patient has only a response value and so only S_1 or S_2 is observed. Therefore, an observation: $t = (a, b, c)$ from a cirrhotic patient corresponds in the latent space to $(s_1, s_2) = (a, b, c, \alpha, \beta, \gamma)$ where the last three coordinates are nonobservable and correspond to a potential observation (α, β, γ) which would have been observed if that patient were non-cirrhotic. A data set $n \times k$ is then represented by a data set $n \times 2k$ in the latent space.

The key idea is that in this new setting we can translate the auxiliary information into a presumption of a high probability for the sets,

$$A_m = \left\{s \in \mathbf{R}^{2k} : s_{1i} \geq s_{2i} \text{ for at least } m \text{ indexes } i \subset \{1, \ldots, k\}\right\}.$$

The reason for giving high probability to these sets is clear. If the additional information telling us that observations from Π_1 tend to take higher values coordinate-wise than those coming from Π_2 is true, then these sets must have high probability.

Looking to the example once again and setting $m = 3$, the last statement means that, for a given patient, the probability of increased HA values under cirrhosis, is high; and the same is presumed for B_1 and B_2. The new representation in the latent space allows us to consider an alternative unrestricted model and to include in a different way the auxiliary information, using not only the marginals but also the joint distribution of $S = (S_1, S_2)$. In this way the theoretical rule can be written as:

$$\text{Classify } z \text{ in } \Pi_1 \text{iff } w_1(z, m) f_{N_k(\mu_1, \Sigma)}(z) \geq w_2(z, m) f_{N_k(\mu_2, \Sigma)}(z) \tag{3}$$

where

$$w_1(t, m) = pr(S \in A_m / S_1 = t) \text{ and } w_2(t, m) = pr(S \in A_m / S_2 = t).$$

To use these rules in practice, assumptions about the joint distribution of (S_1, S_2) must be introduced and the choice of m must be considered. In [6] we proved that values of $m \leq [k/2]$ are recommendable, and we proposed the following family of models for (S_1, S_2):

$$M(\lambda): \begin{cases} S_1 = Z_1 + V \quad S_1 \to N_k(\theta_1, \Sigma) \\ S_2 = Z_2 + V \quad S_2 \to N_k(\theta_2, \Sigma) \\ Z_1 \to N_k(\theta_1, \lambda\Sigma),\ Z_2 \to N_k(\theta_2, \lambda\Sigma) \\ V \to N_k(0, (1-\lambda)\Sigma) \end{cases}$$

where λ is a parameter that measures the degree of association between S_1 and S_2. For $\lambda = 0$, S_1 and S_2 are linearly dependent and we obtain the usual Bayes rule (i.e. is equivalent to using no latent space), while for $\lambda = 1$, S_1 and S_2 are independent. Under these assumptions expressions for computing the weights $w_i(t, m)$ are derived in [6]. The value of these probabilities depend on μ_1, μ_2, Σ and λ. A good performance was obtained in that paper using $\mu_1^* = c_1\widehat{\mu}_1 + c_2(\widehat{\mu}_2 + \delta^*)$, $\mu_2^* = c_1(\widehat{\mu}_1 - \delta^*) + c_2\widehat{\mu}_2$ and S as estimators of the parameters, and $\lambda = 1$. This is the rule we denote as $R(1, 0)$ as $\lambda = 1$ and $\gamma = 0$ (recall that $\delta^* = \delta_0^*$). Notice that we use the RMLE as estimator of the difference of the means and from it we recompute the estimators of the means taking into account that $\delta = \mu_1 - \mu_2$ and that $(\mu_1 + \mu_2)/2 = c_1\mu_1 + c_2\mu_2 - c\delta$.

2.3 New Rules Combining Both Approaches

The approach used in [6] does not take advantage of the estimators defined in [4], since the estimator of the difference of the means considered in the former paper is just the RMLE and, therefore, the iterative procedure (2) is not used.

Moreover, it is not difficult to notice that both approaches can be combined if the estimators obtained by the iterative procedure (2) are used when building the sample version of rule (3).

In this line, we will use, in the rules defined in Sect. 2.2, δ_γ^* with $0 \leq \gamma \leq 1$ as estimator of the difference of the means and $\mu_1^\gamma = c_1\widehat{\mu}_1 + c_2\left(\widehat{\mu}_2 + \delta_\gamma^*\right)$ and $\mu_2^\gamma = c_1\left(\widehat{\mu}_1 - \delta_\gamma^*\right) + c_2\widehat{\mu}_2$ as estimators of μ_1 and μ_2. These are the rules we denote as $R(\lambda, \gamma)$.

This way, we use estimators of the difference of the means δ that are in the interior of the cone, unlike the RLME which is not admissible, and we are able to introduce the additional information in the theoretical formulation of the rule, which was one of the main achievements obtained from the latent space approach. In the simulation study and the example that appear in the Sections below we will see that the combination of both approaches give results that in most cases outperform the TMP obtained by the use of only one of them.

3 Simulations

The purpose of this simulation study is to show the good behavior of the new combined rules when compared with the ones already in the literature, Fisher's rule, and the ones appearing in [4] and [6].

For simplicity we concentrate on the case $k = 3$. The simulations have been performed in higher dimensions obtaining similar results. As $m \leq [k/2]$ is recommended, the only value of m considered is $m = 1$. We generate training samples of size $n_1 = n_2 = 5$ from populations Π_1, $N(\delta, \Sigma)$, and Π_2, $N(0, \Sigma)$. The simulations have also been performed with bigger sample sizes ($n_1 = n_2 = 50$) and with unbalanced sample sizes, rescaling the covariance matrices accordingly. Similar results have also been obtained in those cases.

Forty different simulations are conducted to show cases where δ and/or Σ are different. These parameter configurations were already considered in [4] and [6] and are generated by 10 mean vectors and 4 covariance matrices as:

$$\delta_i\,\Sigma_j, i = 1, \ldots, 10,\ \ j = 1, \ldots, 4.$$

The values for the mean vector δ are determined by $\|\delta\|$ and $\cos(\delta, c)$, where c is the central direction of the positive orthant for the metric given by Σ^{-1}. The concept of central direction of a cone can be found in [1].

We consider two kind of values of the difference of means. Some of them are inside the cone such as $\delta_1, \delta_2, \delta_3$ and $\delta_6, \delta_7, \delta_8$. These points cover situations that are more likely to appear in applications. The higher $\|\delta\|$ and $\cos(\delta, c)$, the more inside the cone are the points. The rest of the values are at the frontier of the cone of restrictions. These sort of points will not appear in applications if the variables considered are useful to discriminate among the populations, but they represent limit situations and therefore they are interesting from the theoretical point of view.

Table 1 Mean vector and covariance matrices

$\delta_1 : \|\delta\|^2 = 0.5$ and $cos(\delta, c) = 1$	$\delta_6 : \|\delta\|^2 = 2$ and $cos(\delta, c) = 1$
$\delta_2 : \|\delta\|^2 = 0.5$ and $cos(\delta, c) = 0.9$	$\delta_7 : \|\delta\|^2 = 2$ and $cos(\delta, c) = 0.9$
$\delta_3 : \|\delta\|^2 = 0.5$ and $cos(\delta, c) = 0.7$	$\delta_8 : \|\delta\|^2 = 2$ and $cos(\delta, c) = 0.7$
$\delta_4 : \|\delta\|^2 = 0.5$ and $\delta \in$ dim 2 face of O_3^+	$\delta_9 : \|\delta\|^2 = 2$ and $\delta \in$ dim 2 face of O_3^+
$\delta_5 : \|\delta\|^2 = 0.5$ and $\delta \in$ dim 1 face of O_3^+	$\delta_{10} : \|\delta\|^2 = 2$ and $\delta \in$ dim 1 face of O_3^+

$$\sum_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \sum_2 = \begin{bmatrix} 1 & 0.3 & 0.3 \\ 0.3 & 1 & 0.3 \\ 0.3 & 0.3 & 1 \end{bmatrix}$$

$$\sum_3 = \begin{bmatrix} 1 & 0.3 & -0.3 \\ 0.3 & 1 & -0.3 \\ -0.3 & -0.3 & 1 \end{bmatrix} \qquad \sum_4 = \begin{bmatrix} 1 & 0.7 & 0.7 \\ 0.7 & 1 & 0.7 \\ 0.7 & 0.7 & 1 \end{bmatrix}$$

The values chosen for the covariance matrix are intended to cover usual values in practice for the correlations coefficients. Full details of the configurations are given in Table 1.

For each scenario, we generated 10, 000 training samples for which the rules are determined. For each of these training samples a test observation for each of the two populations has been classified. The values of the TMP of each of the rules considered in this simulation appear in Table 2. The best value for each scenario appears in bold.

In Table 2 we see that there are two rules performing better than the rest, $R(1, 1)$ and $R(1, 0)$. They are the only ones appearing in bold except for the last two scenarios. If we compare among them we find that $R(1, 1)$ is the best one in 24 out of 40 cases while $R(1, 0)$ is the best for 15, 11 of them corresponding to values not likely to appear in applications. Therefore, we recommend $R(1, 1)$, the rule that combines both approaches taking advantage from the iterative estimation procedure and from the new theoretical rule coming from the latent space.

4 Example. Pima Indians Diabetes Database

This dataset is included in the UCI repository of machine learning databases [3]. The diagnostic, binary-value variable investigated is whether the patient shows signs of diabetes or not. The dataset contains 768 instances and eight attributes. There are 268 elements in group 1, who are those who have tested positive for diabetes, and 500 elements in group 0. The sample size will allow us to divide the data into training and test sets for the purpose of evaluating the correct classification probability. We consider the restrictions between population means given by the whole sample. In this way the diabetes population is assumed to have greater mean values than the healthy population for each of the eight variables. In two different trials we split the dataset into training sample and test sample. Each observation belongs to the training sample with probability 0.25 in the first trial (see Table 3) and 0.1 in the

Table 2 TMP for the different scenarios

Scenario/Rule	R(0,U)	R(0,0)	R(0,1)	R(1,U)	R(1,0)	R(1,1)
$\Sigma_1\delta_1$	0.4351	0.4078	0.3947	0.42135	0.39765	**0.39175**
δ_2	0.43475	0.4131	0.40165	0.42175	0.4027	**0.3978**
δ_3	0.43815	0.42005	0.415	0.4295	**0.41305**	0.4143
δ_4	0.44215	0.42365	0.4131	0.4324	0.41235	**0.41125**
δ_5	0.44235	0.4257	0.4266	0.43375	**0.422**	0.42465
δ_6	0.3183	0.30725	0.3011	0.31175	0.30245	**0.298**
δ_7	0.31435	0.30175	0.29395	0.3077	0.29765	**0.29155**
δ_8	0.31685	0.30205	0.30295	0.31005	**0.29845**	0.302
δ_9	0.3194	0.3049	0.30385	0.31215	**0.3014**	0.30165
δ_{10}	0.3177	0.30455	0.3135	0.30965	**0.30155**	0.313
$\Sigma_2\delta_1$	0.43995	0.42025	0.40735	0.41785	0.40445	**0.3988**
δ_2	0.4426	0.42305	0.41055	0.4226	0.40845	**0.4046**
δ_3	0.43935	0.42345	0.4161	0.42395	**0.41135**	**0.41135**
δ_4	0.43735	0.4199	0.4153	0.41915	**0.4092**	0.41125
δ_5	0.44105	0.4273	0.42965	0.43105	**0.42395**	0.42715
δ_6	0.3151	0.30915	0.30405	0.30495	0.30075	**0.29775**
δ_7	0.31395	0.3049	0.30115	0.3018	0.29615	**0.2948**
δ_8	0.3128	0.30185	0.3033	0.3016	**0.29665**	0.29805
δ_9	0.31905	0.30895	0.311	0.3075	**0.3023**	0.30785
δ_{10}	0.3247	0.3087	0.3187	0.31015	**0.30535**	0.31845
$\Sigma_3\delta_1$	0.4394	0.41245	0.3985	0.4292	0.40515	**0.3959**
δ_2	0.44745	0.4188	0.40385	0.43625	0.40995	**0.4005**
δ_3	0.44365	0.42245	0.4151	0.4342	0.41415	**0.41315**
δ_4	0.43765	0.41405	0.4025	0.42825	0.40435	**0.40025**
δ_5	0.439	0.41995	0.4132	0.43065	0.4119	**0.41155**
δ_6	0.32235	0.31175	0.30555	0.3172	0.30805	**0.3033**
δ_7	0.32115	0.3065	0.3003	0.31535	0.30215	**0.29845**
δ_8	0.31915	0.3023	0.3001	0.31205	0.29875	**0.29815**
δ_9	0.32455	0.31175	0.3098	0.3189	**0.30695**	0.30775
δ_{10}	0.3143	0.3008	0.30505	0.30805	**0.29735**	0.30415
$\Sigma_4\delta_1$	0.4342	0.423	0.41315	0.40715	0.4017	**0.39745**
δ_2	0.44155	0.43145	0.42355	0.41485	0.41145	**0.40945**
δ_3	0.45055	0.43775	0.42685	0.42595	0.4215	**0.41775**
δ_4	0.4433	0.42955	0.4287	0.4243	**0.42275**	0.4234
δ_5	0.4419	0.4377	0.44	0.43415	**0.43405**	0.43845
δ_6	0.314	0.3112	0.3086	0.29485	0.2937	**0.2929**
δ_7	0.3234	0.31905	0.31595	0.3062	0.30345	**0.30245**
δ_8	0.32045	0.31305	0.30835	0.3046	0.30225	**0.30025**
δ_9	0.31925	0.3127	0.3139	**0.3055**	0.30605	0.31175
δ_{10}	0.31645	**0.3074**	0.3143	0.3115	0.3124	0.3153

second (see Table 4). In the first sample all restrictions are verified, so changing parameter γ has no effect on the rule, but we can see that choosing $\lambda = 1$ improves the results. For the second trial the order is not verified by the training means of variables 3 and 4, and we can see that the combination of both approaches yields the best rule. Full results are displayed in Tables 3 and 4.

Table 3 Diabetes database. Incorrect Classification table for test sample in trial 1 ($p = 0.25$)

		Fisher	$m = 4$	$m = 3$	$m = 2$	$m = 1$
	Size	$R(0, \gamma)$	$R(1, \gamma)$	$R(1, \gamma)$	$R(1, \gamma)$	$R(1, \gamma)$
Group 0	377	87	87	86	87	87
Group 1	194	62	61	62	60	62
Global %		0.2609	0.2592	0.2592	**0.2574**	0.2609

Table 4 Diabetes database. Incorrect Classification table for test sample in trial 2 ($p = 0.1$)

		Fisher			$m = 4$			$m = 3$		
	Size	$R(0, U)$	$R(0, 0)$	$R(0, 1)$	$R(1, U)$	$R(1, 0)$	$R(1, 1)$	$R(1, U)$	$R(1, 0)$	$R(1, 1)$
Group 0	452	100	100	98	107	108	109	101	101	100
Group 1	240	69	69	68	59	58	59	66	65	64
Global %		0.2442	0.2442	0.2399	0.2399	0.2399	0.2428	0.2413	0.2399	**0.2370**
					$m = 2$			$m = 1$		
					$R(1, U)$	$R(1, 0)$	$R(1, 1)$	$R(1, U)$	$R(1, 0)$	$R(1, 1)$
Group 0					100	100	99	100	101	99
Group 1					70	69	68	70	69	69
Global %					0.2457	0.2442	0.2413	0.2457	0.2457	0.2428

5 Conclusions

We have defined new rules that incorporate efficiently the additional information that may appear in applications, and proved, using simulations and real data, that they perform better than the ones already defined for this purpose. On future works we will try to adapt this technique to other multivariate methods.

Acknowledgments Research partially supported by Spanish DGES grant MTM2009-11161.

References

1. Abelson, R.P., Tukey, J.W.: Efficient utilization of non-numerical information in quantitative analysis: general theory and the case of the simple order. Ann. Math. Stat. **34**, 1347–1369 (1963)
2. Attallah, A., Toson, E., El-Waseef, A., Abo-Seif, M., Omran, M., Shiha, G.: Discriminant function based on hyaluronic acid and its degrading enzymes and degradation products for differentiating cirrhotic from non-cirrhotic liver diseased patients in chronic HCV infection. Clin. Chim. Acta. **369**, 66–72 (2006)
3. Blake, C.L., Merz, C.J.: UCI Repository Learning Databases. University of California, Department of Information and Computer Science, Irvine, CA. http://www.ics.uci.edu/∼mlearn/MLRepository.html (1998)
4. Fernández, M.A., Rueda, C., Salvador, B.: Incorporating additional information to normal linear discriminant rules. J. Am. Stat. Assoc. **101**, 569–577 (2006)

5. Long, T., Gupta, R.D.: Alternative linear classification rules under order restrictions. Commun. Stat. Theory Methods **27**, 559–575 (1998)
6. Rueda,C., Fernández, M.A., Salvador, B.: Bayes discriminant rules with ordered predictors. J. Classification **26**, 201–225 (2009)
7. Van Eeden, C.: Restricted Parameter Space Estimation Problems. Springer, New York, NY (2006)

The Choice of the Parameter Values in a Multivariate Model of a Second Order Surface with Heteroscedastic Error

Umberto Magagnoli and Gabriele Cantaluppi

Abstract The paper describes an experimental procedure to choose the values for a multivariate vector $\mathbf{x}$ under these conditions: average of $Y(\mathbf{x})$ equal to a target value and least variance of $Y(\mathbf{x})$, linked to $\mathbf{x}$ by a second order model with a heteroscedastic error. The procedure consists of two steps. In the first step an experimental design is performed in the feasible space $\mathcal{X}$ of the control factors to estimate, by an iterative method, the parameters characterizing the response surface of the mean. Then a second experimental design is performed on a set $\mathcal{A}$, subset of $\mathcal{X}$ satisfying a condition on the average of $Y(\mathbf{x})$. This second step determines the choice of $\mathbf{x}$ by using a classification criterion based on the ordering of the sample mean squared errors. The research belongs to the theory of optimal design of experiments [2], that is employed in the Taguchi Methods, used in off-line control [6].

1 Introduction

An experimental procedure is presented to assign proper values to the control factors $\mathbf{x}$, which affect the response Y of a system, so that the two following optimal conditions are satisfied: (a) equality of the mean of $Y(\mathbf{x})$ to an assigned value y_0 and (b) minimum dispersion, respectively

$$E\{Y(\mathbf{x})\} = y_0 \quad \text{and} \quad \min_{\mathbf{x}} Var\{Y(\mathbf{x})\}, \tag{1}$$

$\mathbf{x}$ is assumed to belong to a specified space $\mathcal{X}$, that has a practical interest for the control factors, under the assumptions that are given below.

In particular, we consider a model of a response surface $Y(\mathbf{x})$, where $\mathbf{x} = (x_1, x_2)$ is a vector of only two explicative variables, whose optimal values we want to determine. A second-order model is used for estimating the parameters in a non-linear situation. The response Y of the system depends, thus, on the levels of two control variables x_1, x_2 according to the following relationship:

U. Magagnoli (✉)
Dipartimento di Scienze statistiche, Università Cattolica del Sacro Cuore, Milano, Italy,
e-mail: umberto.magagnoli@unicatt.it; umbertomagagnoli@libero.it

B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-13312-1_8,

$$Y(x_1, x_2) = \mu(x_1, x_2) + E(x_1, x_2) = \\ = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + E(x_1, x_2)\ , \quad (2)$$

where: $\beta_0, \beta_1, \beta_2, \beta_{12}, \beta_{11}$ and β_{22} are unknown parameters; the errors $E(x_1, x_2)$ are random variables independently distributed, on varying x_1 and x_2, as normal random variables, $E(x_1, x_2) \sim N(0, \sigma_E^2(x_1, x_2))$, with zero means and variances $\sigma_E^2(x_1, x_2)$, which, we assume, depend upon the levels x_1, x_2 according to an unknown functional relationship. We observe that relationship (2) takes into account the linear effects of x_1 and x_2 on the response level of Y as well as the quadratic and the interaction ones and that the presence of heteroscedasticity characterizes the behaviour of the response Y.

2 The Procedure

To choose the levels of x_1 and x_2 two experimental designs and a selection procedure are performed.

The First Experimental Design. To estimate the β parameters in (2) we assume to observe the response of the random variable Y for each experimental condition in a three-level full factorial design, with n replications for each experimental condition, according to the following experimental design matrix:

$$\mathbf{D} = \begin{bmatrix} x_{11} & x_{21} & x_{11}x_{21} & x_{11}^2 & x_{21}^2 \\ x_{11} & x_{22} & x_{11}x_{22} & x_{11}^2 & x_{22}^2 \\ x_{11} & x_{23} & x_{11}x_{23} & x_{11}^2 & x_{23}^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{13} & x_{21} & x_{13}x_{21} & x_{13}^2 & x_{21}^2 \\ x_{13} & x_{22} & x_{13}x_{22} & x_{13}^2 & x_{22}^2 \\ x_{13} & x_{23} & x_{13}x_{23} & x_{13}^2 & x_{23}^2 \end{bmatrix} = \begin{bmatrix} \mathbf{d}_1' \\ \mathbf{d}_2' \\ \mathbf{d}_3' \\ \vdots \\ \mathbf{d}_7' \\ \mathbf{d}_8' \\ \mathbf{d}_9' \end{bmatrix},$$

where x_{1r}, x_{2s}, $r, s = 1, 2, 3$, denote the levels of the variables x_1 and x_2; while (x_1, x_2) belong, without loss of generality, to the square experimental region $\mathcal{X} = [-1, 1] \times [-1, 1]$, see [4].

Observe that alternative experimental designs could also be adopted to estimate the β parameters, such as the saturated second order or the fractional factorial ones. The three-level full factorial design allows us to estimate the variance of the error too, although this result can be obtained also through the replications of the experimental design.

We will consider also a sampling design where the $9 \times n$ trials are performed on the same points of the support pertaining the three-level full factorial design but with a different number of replications for each experimental condition established according to a D-optimal design for homoscedastic linear models. Should the heteroscedasticity in the errors not be present, both the determinant of the covariance

matrix of the parameter estimates and the variance of the estimated response Y over the region $\mathcal{X}$ would be, in this way, minimized (see [2] and [5]). D-optimal designs for the heteroscedastic situation are considered in [1], [9], [10] and [11], where the structure of a parametric model to describe the heteroscedasticity of the errors is assumed to be known.

Relationships pertaining only the three-level full factorial design are presented. The adaptation to the D-optimal design (for the homoscedastic case) involves taking into account the different number of replications for each experimental condition.

Let $\mathbf{X}$ be the matrix, $9n \times (1+5)$, obtained by concatenating the unitary vector $\mathbf{1}$, $9n \times 1$, with the stacking of the elements $\mathbf{1}_{n\times 1}\mathbf{d}'_j$, $j = 1, \ldots, 9$; in this way elements, which theoretically give the same variance, are grouped. Let $\mathbf{Y}$ be the column vector, $9n \times 1$, whose elements are the responses of the Y variable, defined by (2), for each experimental condition and replication. Let $\boldsymbol{\beta} = [\beta_0\ \beta_1\ \beta_2\ \beta_{12}\ \beta_{11}\ \beta_{22}]'$ be the column vector, $(1+5) \times 1$, containing the unknown parameters, and $\mathbf{E}$ the column vector, $9n \times 1$, whose elements are the error components $E(x_{1ri}, x_{2si}) \sim N(0, \sigma_E^2(x_{1r}, x_{2s}))$, $i = 1, \ldots, n$. Observe that the value $\sigma_E^2(x_{1r}, x_{2s})$ of the variance of E is independent of i. According to (2) we have $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E}$, being $\mathbf{E} \sim N(\mathbf{0}, \boldsymbol{\Omega})$, and $\boldsymbol{\Omega} = \mathbf{I}_n \otimes \boldsymbol{\Sigma}$ the diagonal matrix, $9n \times 9n$, whose non-zero elements are the variances of the errors in the vector $\mathbf{E}$; $\boldsymbol{\Sigma} = \text{diag}[\sigma_E^2(x_{11}, x_{21}), \sigma_E^2(x_{11}, x_{22}), \sigma_E^2(x_{11}, x_{23}), \sigma_E^2(x_{12}, x_{21}), \sigma_E^2(x_{12}, x_{22}), \sigma_E^2(x_{12}, x_{23}), \sigma_E^2(x_{13}, x_{21}), \sigma_E^2(x_{13}, x_{22}), \sigma_E^2(x_{13}, x_{23})]$.

If the elements of $\boldsymbol{\Omega}$ were known we could apply the generalized least squares estimator of the parameters $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{Y};$$

since we cannot assume that the elements of $\boldsymbol{\Omega}$ are known, one can make recourse to the estimation method presented in [8], see also [3]. Referring to this estimation method we suggest a specific iterative procedure, which at each step p provides a provisional estimate $\hat{\boldsymbol{\beta}}_p$ of $\boldsymbol{\beta}$ and a provisional estimate $\hat{\boldsymbol{\Omega}}_p$ of the matrix $\boldsymbol{\Omega}$:

$$\hat{\boldsymbol{\beta}}_p = \left(\mathbf{X}'\hat{\boldsymbol{\Omega}}_{p-1}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\boldsymbol{\Omega}}_{p-1}^{-1}\mathbf{Y};$$

$$\hat{\boldsymbol{\Omega}}_p = \mathbf{I}_n \otimes \mathbf{S}_p^2 = \mathbf{I}_n \otimes \text{diag}[s_p^2(x_{1r}, x_{2s}); r, s = 1, 2, 3];$$

where $s_p^2(x_{1r}, x_{2s}) = \left(\mathbf{Y}_j - \mathbf{X}_j\hat{\boldsymbol{\beta}}_p\right)'\left(\mathbf{Y}_j - \mathbf{X}_j\hat{\boldsymbol{\beta}}_p\right)/(n-1-5)$ with $\mathbf{X}_j = \left[\mathbf{1}\ \mathbf{1}\mathbf{d}'_j\right]$; $\mathbf{1}$ is the unitary vector, $n \times 1$, and the index j denotes also the subset of elements in the vector $\mathbf{Y}$ corresponding to the n replications x_{1ri}, x_{2si}, $i = 1, 2, \ldots, n$, of each experimental condition specified by the levels x_{1r}, x_{2s}.

The generalized least squares estimate $\hat{\boldsymbol{\beta}}_p$ is based on $\hat{\boldsymbol{\Omega}}_{p-1}$ and we assume that, in the first step, $\hat{\boldsymbol{\beta}}_1$ is the ordinary least squares estimate, $\hat{\boldsymbol{\beta}}_1 = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}$.

The final estimate of the parameters $\boldsymbol{\beta}$ is given by the $\hat{\boldsymbol{\beta}}$ values, obtained at the end of the iteration procedure, when $\hat{\boldsymbol{\beta}}_{p-1}$ is sufficiently near to $\hat{\boldsymbol{\beta}}_p$, and the final

estimate of the variances $\sigma_E^2(x_{1r}, x_{2s})$ corresponding to the levels r and s of the control variables is

$$s^2(x_{1r}, x_{2s}) = \frac{1}{n-1-5}\left(\mathbf{Y}_j - \mathbf{X}_j\hat{\boldsymbol{\beta}}\right)'\left(\mathbf{Y}_j - \mathbf{X}_j\hat{\boldsymbol{\beta}}\right), \text{ for } r, s = 1, 2, 3.$$

For all pairs of values x_1, x_2 of the feasible space, $\mathcal{X}$, of the two control factors we may define also the value of the estimated response, $\mu(x_1, x_2)$, of the variable Y:

$$\hat{y}(x_1, x_2) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_{12} x_1 x_2 + \hat{\beta}_{11} x_1^2 + \hat{\beta}_{22} x_2^2. \tag{3}$$

The Second Experimental Design. Let y_0 be a target value of the response Y and let $\varepsilon > 0$ be a value defined by the experimenter. By considering the relationship $\hat{y}(x_1, x_2) - \varepsilon \leq y_0 \leq \hat{y}(x_1, x_2) + \varepsilon$, that defines a set of acceptable y values including the target value y_0, we can define a subset $\mathcal{A} \subset \mathcal{X}$ of x_1, x_2 values, that depends, according to (3), on the estimates of the β parameters obtained in the first experimental design. The set $\mathcal{A}$ may be considered as a set of experimental conditions that ensure the system to satisfy the target value y_0. We observe that all the points in the set $\mathcal{A}$ are characterized approximately by the same mean value of the response variable Y. We extend this set in the new region $\mathcal{A}^+$ defined by adding to each element in the set $\mathcal{A}$ its neighbouring elements.

In the set of the M points on the feasible space of x_1 and x_2 belonging to the set $\mathcal{A}^+$ we consider a subset of K points – that is K pairs (x_{1k}, x_{2k}), $k = 1, 2, \ldots, K$ – to submit to the procedure presented below. The K points may be extracted by making use of a sampling without replacement or with a systematic sampling.

The Selection Procedure. We submit the subset of K points (x_{1k}, x_{2k}), $k = 1, \ldots, K$, sampled in the second experimental design to the following experimental procedure.

1. Obtain m replications of the experiment on every experimental condition belonging to this set of K points.
2. Estimate the mean, $\bar{y}(x_{1k}, x_{2k})$, and the variance, $s_Y^2(x_{1k}, x_{2k})$, with the m replications of the response Y for every experimental condition in the set of K points.
3. To identify the most desirable experimental conditions re-order the K points (x_{1k}, x_{2k}) according, see [7], to their estimated mean squared error

$$MSE_Y(x_{1k}, x_{2k}) = \left[\bar{y}(x_{1k}, x_{2k}) - y_0\right]^2 + s_Y^2(x_{1k}, x_{2k}),$$

   where y_0 is the target value for the response Y.
4. Choose the points characterized by the least mean squared error levels.

We observe that the final selection of the ideal experimental condition from this set of points may be performed by the experimenter by having also recourse to economical arguments; say, the ideal experimental condition could be the one minimizing a cost function over this final set of points.

3 Some Results Obtained by Simulation

A simulation example is proposed to evaluate the experimental procedure with regard to a system whose response is defined, according to (2), as follows:

$$Y(x_1, x_2) = 2.6 - 5.2x_1 - 5.2x_2 - 2.4x_1x_2 + 4x_1^2 + 4x_2^2 + E(x_1, x_2). \quad (4)$$

The errors $E(x_1, x_2)$ are independently distributed as normal random variables $E(x_1, x_2) \sim N(0, \sigma_E^2(x_1, x_2))$ with variances depending upon the control factors according to the relationship, assumed to be unknown to the experimenter

$$\sigma_E^2(x_1, x_2) = 0.03125 - 0.0625x_1 + 0.0625x_2 + 0.375x_1x_2 + 0.25x_1^2 + 0.25x_2^2 . \quad (5)$$

The target value $y_0 = 6$ is considered for Y: the corresponding "optimal" experimental condition ($x_{10} = -0.26$, $x_{20} = 0.26$) is characterized by least variance.

The first experimental design is performed on $\mathcal{X} = [-1, 1] \times [-1, 1]$ by making recourse to both a three-level full factorial design, with $n = 10$ trials for each experimental condition $x_{11} = -1, x_{12} = 0, x_{13} = 1$, and $x_{21} = -1, x_{22} = 0, x_{23} = 1$, and a D-optimal experimental design for homoscedastic models.

The set $\mathcal{A}^+$ is defined, in the second experimental design, by the experimental conditions that ensure, according to the parameter estimates obtained in the first experimental design, values of the response Y far from y_0 no more than $\varepsilon = 0.15$.

The feasible space $\mathcal{X}$ is assumed to consist of a mesh of $81 \times 81 = 6{,}561$ possible experimental conditions. Sample of different sizes, K, extracted with the simple random sampling without replacement and the systematic sampling techniques, with various numbers, m, of replications for each experimental condition are considered. Observe that the value of m is independent of n.

The interest has been focused, for different replication sizes N of the procedure, on the following properties: (a) The fraction of simulations with some points, among those sampled in the set $\mathcal{A}^+$, belonging also to a set $\mathcal{C}$ defined by experimental conditions satisfying $\sigma_E^2(x_{1r}, x_{2s}) \leq \sigma_0^2 = 1.5 \times \sigma_E^2(x_{10}, x_{20})$, where $\sigma_E^2(x_{10}, x_{20}) = 0.072$ is the least variance level for the "optimal" experimental condition (x_{10}, x_{20}). (b) The average number of the sampled points belonging to $\mathcal{C} \cap \mathcal{A}^+$, see Fig. 1.

With regard to the properties (a) and (b) both experimental designs, considered in the first step, give similar results, see Table 1. We observe that they are both suboptimal, since an explicit relationship to model the presence of heteroscedasticity has not been considered. Table 1 shows that, for almost every N and K, over 90% of the replications of the procedure present at least a value in $\mathcal{C} \cap \mathcal{A}^+$. With regard to the average number of the sampled points in $\mathcal{C} \cap \mathcal{A}^+$ the systematic sampling works generally better than the simple random sampling without replacement.

For each replication, $i = 1, \ldots, N$, of the simulation, let W_i be the random variable describing the number of pairs (x_{1k}, x_{2k}) in $\mathcal{C} \cap \mathcal{A}^+$ among the K pairs sampled in $\mathcal{A}^+$. If p denotes the probability to sample a pair in $\mathcal{C} \cap \mathcal{A}^+$, then

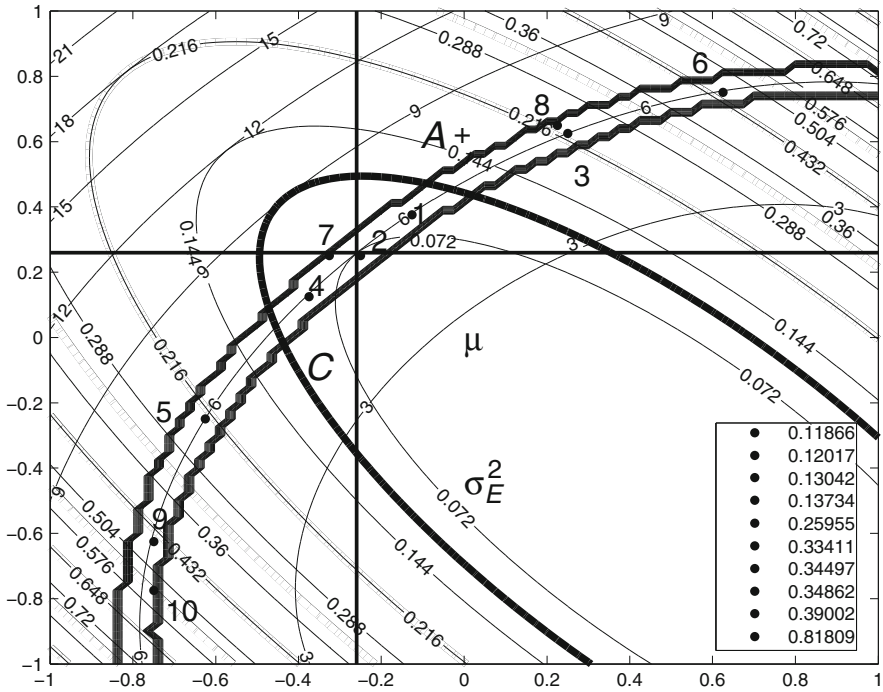


Fig. 1 Response surfaces for the mean and the variance, see (4) and (5). $\mathcal{A}^+$ is the set between the thicker contour levels of the mean. $\mathcal{C}$ is the convex set defined by the thicker contour level of the variance. The points in $\mathcal{A}^+$ denote the experimental conditions that have been sampled in the second experimental design; they are numbered according to the ranking of the mean squared errors of Y; the best one, identified by the number 1, has a value of 0.11866

Table 1 The fraction of simulations with some of the points in the set $\mathcal{A}^+$ belonging also to the region $\mathcal{C}$ and the average number of the sampled points in $\mathcal{C} \cap \mathcal{A}^+$

$y_0 = 6$		Simple random sampling			Systematic sampling		
N	K	10	15	20	10	15	20
50	[a]	0.840, 1.960	0.980, 3.300	1.000, 3.980	1.000, 2.840	1.000, 3.940	1.000, 5.460
100	[a]	0.900, 2.160	0.980, 3.380	0.990, 4.240	1.000, 2.780	1.000, 4.230	1.000, 5.470
200	[a]	0.920, 2.320	0.965, 3.375	0.995, 4.345	0.985, 2.835	1.000, 4.280	1.000, 5.430
500	[a]	0.916, 2.186	0.968, 3.420	0.998, 4.414	0.984, 2.856	1.000, 4.288	1.000, 5.512
50	[b]	0.940, 2.400	1.000, 3.820	1.000, 4.800	0.980, 2.860	1.000, 4.340	0.980, 5.520
100	[b]	0.900, 2.250	0.990, 3.510	1.000, 4.650	0.990, 2.750	1.000, 4.350	0.990, 5.500
200	[b]	0.940, 2.340	0.985, 3.405	1.000, 4.795	0.995, 2.730	1.000, 4.245	0.995, 5.410
500	[b]	0.932, 2.318	0.988, 3.356	1.000, 4.530	0.988, 2.818	1.000, 4.254	0.998, 5.376

[a] 1^{st} experimental design: 3-level full factorial with 10 trials for each experimental condition
[b] 1^{st} experimental design: D-optimal with 90 trials on the same set of the 3-level full factorial
K: number of sampled points in the region $\mathcal{A}^+$,
N: number of replications of the whole procedure

$W_i \sim W \sim Bin(K, p)$. We remind that in each replication the empirical value p and the set $\mathcal{A}^+$ depend upon the estimates of the parameters in (4) obtained in the first step.

With reference to the following system of hypotheses

$$H_0 : p \geq p_0 \quad \text{versus} \quad H_1 : p < p_0 \tag{6}$$

we can decide to accept the null hypothesis $H_0 : p \geq p_0$, in presence of N replications of the simulation, depending on the value of one of the statistics $T = \sum_{i=1}^{N} W_i$ or $R = \sum_{i=1}^{N} W_i/N$, where $T \sim Bin(NK, p)$. Table 3 reports the critical values for R at the significance level $\alpha = 0.05$ for each pair N, K. If we consider $p_0 = 0.20$ for the simple random sampling and $p_0 = 0.25$ for the systematic sampling we may observe that every average number of experimental conditions in $\mathcal{C} \cap \mathcal{A}^+$, see Table 1, is greater than the pertaining critical value. For $K = 10, 15, 20$ we may expect to sample respectively at least 2, 3 and 4 values close to the theoretical value (x_{10}, x_{20}). The best experimental conditions can be chosen by sorting the mean squared errors of the response Y for each pair (x_{1k}, x_{2k}). Figure 1 shows an example of the points, sampled in the second experimental design, ordered according to their estimated mean squared errors for $K = 10$: four of the sampled points are in $\mathcal{C} \cap \mathcal{A}^+$.

To study the attitude of the procedure to select experimental conditions with low variance of the error E, Table 2 reports the average number of experimental conditions with rank of the mean squared error of Y lower than r_0, where r_0 is 4, 6, 8 for $K = 10, 15, 20$. A test similar to (6) can be performed to check if the probability that "the best sampled experimental conditions are in the set $\mathcal{C} \cap \mathcal{A}^+$" is greater than a value p_0. The critical values corresponding to the different N, K for the statistic R are reported for $\alpha = 0.05$ in Table 3. If we consider $p_0 = 0.15$ for the

Table 2 Average number of experimental conditions, among the four with least MSE, in $\mathcal{C} \cap \mathcal{A}^+$

$y_0 = 6$		Simple random sampling						Systematic sampling					
	m	10	10	10	20	20	20	10	10	10	20	20	20
N	K	10	15	20	10	15	20	10	15	20	10	15	20
50	[a]	1.360	2.520	3.140	1.440	2.540	3.260	2.460	3.400	4.460	2.540	3.420	4.620
100	[a]	1.560	2.610	3.330	1.640	2.610	3.460	2.440	3.620	4.460	2.530	3.630	4.630
200	[a]	1.705	2.555	3.325	1.795	2.605	3.505	2.505	3.625	4.455	2.565	3.685	4.620
500	[a]	1.668	2.592	3.406	1.742	2.678	3.544	2.496	3.612	4.460	2.568	3.710	4.594
50	[b]	1.780	2.920	3.600	1.900	2.960	3.800	2.480	3.680	4.440	2.540	3.740	4.520
100	[b]	1.680	2.680	3.570	1.740	2.740	3.720	2.390	3.710	4.390	2.460	3.800	4.540
200	[b]	1.730	2.630	3.710	1.785	2.725	3.830	2.370	3.600	4.425	2.445	3.685	4.560
500	[b]	1.738	2.576	3.504	1.836	2.692	3.658	2.446	3.636	4.386	2.522	3.728	4.528

[a] 1^{st} experimental design: 3-level full factorial with 10 trials for each experimental condition
[b] 1^{st} experimental design: D-optimal with 90 trials on the same set of the 3-level full factorial
K: number of sampled points in the region $\mathcal{A}^+$,
m: number of replications for each sampled point,
N: number of replications of the whole procedure

Table 3 Critical values for the statistic $R = \sum_{i=1}^{N} W_i/N$ at the significance level $\alpha = 0.05$ to test the hypotheses $H_0 : p \geq p_0$, $H_1 : p < p_0$ with reference to a binomial random variable $W \sim Bin(K, p)$

		$p_0 = 0.25$			$p_0 = 0.20$			$p_0 = 0.15$		
N	K	10	15	20	10	15	20	10	15	20
50		2.180	3.360	4.560	1.700	2.640	3.580	1.240	1.940	2.640
100		2.280	3.480	4.680	1.790	2.750	3.710	1.320	2.020	2.740
200		2.340	3.555	4.775	1.855	2.820	3.795	1.370	2.090	2.815
500		2.400	3.626	4.858	1.908	2.886	3.868	1.418	2.148	2.882

simple random sampling and $p_0 = 0.20$ for the systematic sampling, which seems to have a better performance, we may observe that every average number of "best" experimental conditions in $\mathcal{C} \cap \mathcal{A}^+$ is greater than the pertaining critical value. We may then expect to have respectively for $K = 10, 15, 20$ at least 1.5, 2.25 and 3 values in $\mathcal{C} \cap \mathcal{A}^+$ characterized by a rank of the mean squared error lower than r_0 among the sampled experimental conditions.

4 Conclusions

An experimental procedure is proposed to choose the values of control factors to realize a target value for the response of a system described by a second order model with heteroscedastic error. To ensure a good definition of the response surface for the mean, the space $\mathcal{X}$ pertaining a first experimental design is defined over a wide range for all control factors. To search the experimental condition with least variance the attention is then focused, by a second experimental design, on a restricted set $\mathcal{A}^+$, whose elements satisfy approximately a condition on the average of Y.

The procedure may help the operator, by simulation, to choose the parameters – like the experimental design, the number of experimental trials, the extension of the set $\mathcal{A}^+$ and the size of the sample in $\mathcal{A}^+$ – which can affect a particular type of experimentation. An example of simulation is reported to study the behaviour of the procedure. Future developments will regard the analysis of the behaviour of the procedure in presence of various theoretical relationships for the variance, of various specifications of the experimental parameters n, K, m and of a possible definition of the set $\mathcal{A}^+$ as a function of the precision of the β parameter estimates obtained through the proposed algorithm and the study of the performance of the estimation algorithm proposed in the first step.

References

1. Atkinson, A.C., Cook, R.D.: D-optimum designs for heteroscedastic linear models. J. Am. Stat. Assoc. **90**(429), 204–212 (1995)
2. Giovagnoli, A.: La teoria dei piani di esperimento ottimi. In: Atti della XXXVI Riunione della Società Italiana di Statistica, Pescara, pp. 163–186 (1992)

3. Goldstein, H.: Multilevel mixed linear model analysis using iterative generalized least squares. Biometrika **73**(1), 43–56 (1986)
4. Khuri, A.I., Cornell, J.A.: Response Surfaces. Design and Analyses. Marcel Dekker, New York, NY (1987)
5. Kiefer, J.C.: Collected Papers Volume 3: Design of Experiments. (Brown, L.D., Olkin, I., Sacks, J., Wynn, H.P. (eds.)). Springer, New York, NY (1985)
6. Logothetis, N., Wynn, H.P.: Quality Through Design: Experimental Design, Off-line Quality Control and Taguchi's Contributions. Clarendon Press, Oxford (1989)
7. Magagnoli, U.: Il controllo della qualità "off-line": problemi statistici relativi a strategie decisionali ottimali. Quaderni di Statistica e Matematica applicata alle Scienze Economico-Sociali **XIV**(5), 193–210 (1992)
8. Magagnoli, U., Chiodini, P.: Su una procedura iterativa per la stima di una funzione di regressione mediante il criterio dei minimi quadrati ponderati. In: Frosini, B.V., Magagnoli, U., Boari, G. (eds.) Studi in onore di Angelo Zanella, pp. 423–438. Vita e Pensiero, Milano (2002)
9. Montepiedra, G., Wong, W.K.: A new design criterion when heteroscedasticity is ignored. Ann. Inst. Stat. Math. **53**(2), 418–426 (2001)
10. Rodríguez, C., Ortiz, I.: D-optimum designs in multi-factor models with heteroscedastic errors. J. Stat. Plan. Inference **128**, 623–631 (2005)
11. Tack, L., Goos, P., Vandebroek, M.: Efficient Bayesian designs under heteroscedasticity. J. Stat. Plan. Inference **104**, 469–483 (2002)

Mixed Mode Data Clustering: An Approach Based on Tetrachoric Correlations

Isabella Morlini

Abstract In this paper we face the problem of clustering mixed mode data by assuming that the observed binary variables are generated from latent continuous variables. We perform a principal components analysis on the matrix of tetrachoric correlations and we then estimate the scores of each latent variable and construct a data matrix with continuous variables to be used in fully Guassian mixture models or in the k-means cluster analysis. The calculation of the expected a posteriori (EAP) estimates may proceed by simply considering a limited number of quadrature points. Main results on a simulation study and on a real data set are reported.

1 Introduction

One possible approach to cluster analysis is the mixture maximum likelihood method, in which the data to be clustered are assumed to come from a finite mixture of populations. The method has been well developed and much used for the case of normal populations. A main advantage in using Gaussian distributions is that a number of possible restrictions on the covariance matrices has been proposed in literature (e.g., [1, 3]) to deal with different local dependencies and, at the same time, to alleviate the problem of the rapidly growing of the parameters with the data dimension and with the number of clusters. A large range of Gaussian models are available, from the simple spherical one to the least parsimonious where all elements of the covariance matrix are allowed to vary across clusters. Practical applications, however, often involve mixture of categorical and continuous variables. Everitt [4] and Everitt and Merette [5] extended the normal model to deal with mixed mode data but the computation involved in their model is so extensive that is only feasible for data with very few categorical variables. Lawrence and Krzanowski [7] and Vermunt & Magidson [12] propose conditional Gaussian models with local

I. Morlini (✉)
Dipartimento di Scienze Sociali, Cognitive e Quantitative, Università di Modena e Reggio Emilia, 42100 Reggio Emilia, Italy,
email: isabella.morlini@unimore.it

B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-13312-1_9,

independence structure. Local dependencies are specified only between pairs of categorical variables and between pairs of continuous variables and are dealt via joint multinomial and multivariate normal distributions. In the "Latent Gold" package [11] the dependence between a categorical and a continuous variable may be dealt with a sort of "trick", by doubling the categorical variable and treating the variable also as a covariate. The estimated dependence, however, may not vary between groups. The mixture model for large data sets implemented in the package SPSS is also based on joint multinomial and gaussian distributions and postulate the hypothesis of local independence between a categorical and a continuous variable.

Here we face the problem of clustering data with different scales and allowing local dependencies also between a categorical and a continuous variable by assuming that each observed categorical variable is generated from a latent continuous variable and by estimating the scores of these latent variables. In economics, these variables are called utility functions and the assumption is that the response (which may be, for example, the presence or the absence of a public service or a public utility) are determined by the crossing of certain thresholds in these functions (see, among others, [8]). Heckman [6] models whether or not American states have introduced fair-employment legislation and describes the corresponding latent response as the "sentiment" favoring fair-employment legislation. In genetics, the latent response is interpreted as the "liability" to develop a qualitative trait or phenotype. There are also examples of continuous variables which are sampled as binary (among others, bit data which are originated by electric voltages). Skrondal and Rabe-Hesketh [10], pp. 16–17, report various interpretations of these latent variables and also state that assuming a latent continuous variable may be useful regardless of whether the latent response can be given a real meaning.

This work represents the first step in the construction of fully Gaussian models for classification, in which correlations among variables may vary across groups and also variable selection may be faced differently in each group. Here we estimate the scores of each latent variable and reach a data matrix with all continuous variables to be used in these models. An application shows that some benefits of using a data matrix with all continuous variables instead of a mixed mode data matrix may be reached in the k-means cluster analysis.

2 From Binary Variables to Continuous Variables

The essential feature of the method to be described in this section is that the observed categorical variables are generated from underlying latent continuous variables according to the values of a set of thresholds. Here we formalize results regarding binary variables but the theory may be extended to multinomial variables by estimating the matrix of polychoric correlations. Given p vectors of binary variables observed for a sample of size n, a contingency table for each couple of variables X_k and X_j is constructed, with the following cell frequencies:

	$x_k = 0$	$x_k = 1$
$x_j = 0$	e_{jk}	b_{jk}
$x_j = 1$	c_{jk}	d_{jk}

The estimated value for the threshold generating the variable X_k is the value h_k satisfying $\Phi(h_k) = (e_{jk} + c_{jk})/n$. For variable X_j it is the value h_j satisfying $\Phi(h_j) = (e_{jk} + b_{jk})/n$, where Φ is the standard normal cumulative distribution function. We then estimate the tetrachoric correlation coefficient r_{jk} conditional on these thresholds, via maximum likelihood. The solution may be found iteratively or by using the following approximate analytic solution:

$$r_{jk} = \sin\left(\frac{\pi}{2}\left(1 + \frac{4e_{jk}b_{jk}c_{jk}d_{jk}n^2}{(e_{jk}d_{jk} - b_{jk}c_{jk})^2(e_{jk} + d_{jk})(b_{jk} + c_{jk})}\right)^{-1/2}\right) \quad (1)$$

In tables with zero frequencies, zero values are set to 0.5. In a simulation study with 5000 different data sets of size (100×6) generated from 10 multivariate normal populations, the estimator (1) has been shown to give better results than the other ones based on approximate analytic solutions of the likelihood function. The $(n \times p)$ matrix of the scores of the p latent continuous variables is reached with expected a posteriori (EAP) estimates. In order to reach semi parametric estimates, we consider a model based on principal components rather than on factors (see, for example, [2] and [9], for EAP estimates reached by considering a fully parametric model where also thresholds, eigenvalues and eigenvectors associated with each factor are estimated by maximizing the likelihood function). We perform a principal component analysis on the matrix of tetrachoric correlations (which does not require previous smoothing if the matrix is not positive definite) and consider the following model:

$$t_{ij} = a_{j1}y_{i1} + a_{j2}y_{i2} + \ldots + a_{jk}y_{ik} + \ldots + a_{jp}y_{ip} \quad (2)$$

where t_{ij} is the score of principal component j for case i, a_{jk} are the loadings (eigenvectors) and y_{ik} is the score for case i relative to the k latent variable associated with the observed categorical variable x_k as follows: $x_{ik} = 1$ if $y_{ik} \geq h_k$ and $x_{ik} = 0$ if $y_{ik} < h_k$. As assumed for the thresholds estimates, $\mathbf{y} \sim N(\mathbf{0}, I)$ and $\mathbf{t} \sim N(\mathbf{0}, \Lambda)$ where Λ is a diagonal matrix with elements $\lambda_j^2 = \sum_{k=1}^{p} a_{jk}^2$ equal to the eigenvalues. The EAP estimator of the jth principal component score is the mean of the posterior distribution of t_j, which is expressed by:

$$\tilde{t}_{ij} = E(t_{ij}|\mathbf{x}_i; \mathbf{w}) = \int t_j f(t_j|\mathbf{x}_i; \mathbf{w}) dt_j = \int \frac{t_j f(\mathbf{x}_i|t_j; \mathbf{w}) g(t_j|\mathbf{w})}{\int f(\mathbf{x}_i|t_j; \mathbf{w}) g(t_j|\mathbf{w}) dt_j} dt_j \quad (3)$$

where **w** is the vector of known parameters (the thresholds and the eigenvectors). In the following equations, for economy of space, **w** will be omitted. Given $\sigma_{jk}^2 = \lambda_j^2 - a_{jk}^2 = \sum_{h \neq k} a_{jh}^2$, then

$$P(x_{ik} = 1|t_j) = \frac{1}{\sigma_{jk}\sqrt{2\pi}} \int_{h_k}^{\infty} e^{\left(-(t_{ij} - a_{jk} y_{ik})^2 / 2\sigma_{jk}^2\right)} dy_{ik} \tag{4}$$

Introducing the change in the variable:

$$P(x_{ik} = 1|t_j) = \frac{1}{a_{jk}\sqrt{2\pi}} \int_{-\infty}^{(t_{ij} - a_{jk} h_k)/\sigma_{jk}} e^{(-z^2/2)} dz \qquad (a_{jk} > 0) \tag{5}$$

$$P(x_{ik} = 1|t_j) = \frac{1}{-a_{jk}\sqrt{2\pi}} \int_{(t_{ij} - a_{jk} h_k)/\sigma_{jk}}^{\infty} e^{(-z^2/2)} dz \qquad (a_{jk} < 0) \tag{6}$$

Letting $z_{jk} = (t_{ij} - a_{jk} h_k)/\sigma_{jk}$ and $F_{jk}(t_j) = (a_{jk})^{-1} \Phi(z_{jk})$ when $a_{jk} > 0$, $F_{jk}(t_j) = |a_{jk}|^{-1}(1 - \Phi(z_{jk}))$ when $a_{jk} < 0$, assuming the independence of the binary variables x_k conditional on each component t_j, it results

$$f(\mathbf{x}_i|t_j) = \prod_{k=1}^{p} F_{jk}(t_j)^{x_{ik}} [1 - F_{jk}(t_j)]^{1 - x_{ik}} \tag{7}$$

We consider S quadrature points and estimate the scores as follows:

$$\tilde{t}_{ij} = \sum_{s=1}^{S} t_{sj} \frac{\phi(t_{sj}) \prod_{k=1}^{p} F_{jk}(t_j)^{x_{ik}} [1 - F_{jk}(t_j)]^{1 - x_{ik}}}{\sum_{s=1}^{S} \phi(t_{sj}) \prod_{k=1}^{p} F_{jk}(t_j)^{x_{ik}} [1 - F_{jk}(t_j)]^{1 - x_{ik}}} \tag{8}$$

where t_{sj} are equally spaced points in $[-z_j, z_j]$ with $\Phi(-z_j/\lambda_j) = 0.001$, $\phi(t_{sj})$ are the density functions of these points in the $N(0, \lambda_j^2)$ curve times the interval size.

Given the estimates $\tilde{t}_{ij}$, the EAP estimates $\tilde{y}_{ik}$ of the latent variables may be then reached through analogous steps. The EAP estimator of the kth variable scores is the mean of the posterior distribution of y_k, which is expressed by:

$$\tilde{y}_{ik} = E(y_{ik}|x_{ik}; \mathbf{t}_i) = \int y_k f(y_k|x_{ik}; \mathbf{t}_i) dy_k = \int \frac{y_k f(x_{ik}|y_k; \mathbf{t}_i) g(y_k)}{\int f(x_{ik}|y_k; \mathbf{t}_i) g(y_k) dy_k} dy_k \tag{9}$$

Let y_{ik}^+ be the values $y_{ik} \geq h_k$ and y_{ik}^- be the values $y_{ik} < h_k$, then

$$P(x_{ik}=1|y_k;\tilde{t}_{ij}) = \frac{1}{a_{jk}\sqrt{2\pi}} \int_{-\infty}^{\frac{\tilde{t}_{ij}-a_{jk}y_{ik}^{+}}{\sigma_{jk}}} e^{-z^2/2}dz \qquad (a_{jk}>0) \tag{10}$$

$$P(x_{ik}=1|y_k;\tilde{t}_{ij}) = \frac{1}{|a_{jk}|\sqrt{2\pi}} \int_{\frac{\tilde{t}_{ij}-a_{jk}y_{ik}^{+}}{\sigma_{jk}}}^{\infty} e^{-z^2/2}dz \qquad (a_{jk}<0) \tag{11}$$

$$P(x_{ik}=0|y_k;\tilde{t}_{ij}) = \frac{1}{|a_{jk}|\sqrt{2\pi}} \int_{-\infty}^{\frac{\tilde{t}_{ij}-a_{jk}y_{ik}^{-}}{\sigma_{jk}}} e^{-z^2/2}dz \qquad (a_{jk}<0) \tag{12}$$

$$P(x_{ik}=0|y_k;\tilde{t}_{ij}) = \frac{1}{a_{jk}\sqrt{2\pi}} \int_{\frac{\tilde{t}_{ij}-a_{jk}y_{ik}^{-}}{\sigma_{jk}}}^{\infty} e^{-z^2/2}dz \qquad (a_{jk}>0) \tag{13}$$

Let

$$z_{jk}^{+} = \frac{\tilde{t}_{ij}-a_{jk}y_{ik}^{+}}{\sigma_{jk}} \qquad z_{jk}^{-} = \frac{\tilde{t}_{ij}-a_{jk}y_{ik}^{-}}{\sigma_{jk}} \tag{14}$$

and $F_{jk}^{+}(y_k) = (a_{jk})^{-1}\Phi(z_{jk}^{+})$ when $a_{jk}>0$, $F_{jk}^{+}(y_k) = |a_{jk}|^{-1}(1-\Phi(z_{jk}^{+}))$ when $a_{jk}<0$, $F_{jk}^{-}(y_k) = |a_{jk}|^{-1}\Phi(z_{jk}^{-})$ when $a_{jk}<0$, $F_{jk}^{+}(y_k) = (a_{jk})^{-1}(1-\Phi(z_{jk}^{-}))$ when $a_{jk}>0$. Then $f(x_{ik}|y_k;\mathbf{t}_i) = \sum_{j=1}^{p} F_{jk}^{+}(y_k)^{x_{ik}}F_{jk}^{-}(y_k)^{1-x_{ik}} \times \phi(\tilde{t}_{ij})$. Considering S quadrature points we estimate the scores as follows:

$$\tilde{y}_{ik} = \sum_{s=1}^{S} y_{sk} \frac{\phi(y_{sk})\sum_{j=1}^{p}(F_{jk}^{+}(y_s)^{x_{ik}}F_{jk}^{-}(y_s)^{1-x_{ik}} \times \phi(\tilde{t}_{ij}))}{\sum_{s=1}^{S}\phi(y_{sk})(\sum_{j=1}^{p}F_{jk}^{+}(y_s)^{x_{ik}}F_{jk}^{-}(y_s)^{1-x_{ik}} \times \phi(\tilde{t}_{ij}))} \tag{15}$$

where y_{sk} are equally spaced points in $[-z_j \quad h_k]$ when $x_{ik}=0$, in $[h_k \quad z_j]$ when $x_{ik}=1$, with $\Phi(-z_j)=0.001$, $\phi(y_{sk})$ being the density functions of these points in the $N(0,1)$ curve times the interval size.

3 Main Results on a Simulation Study and on a Real Data Set

A simulation study is used to evaluate the accuracy of the tetrachoric correlations and the scores estimates. From 10 standard multivariate normal populations with correlation matrices $\mathbf{P}$ with equal elements $\rho_{rc}, r \neq c$, out of the main diagonal, ranging form 0.0 to 0.95, we generate 5,000 data sets (500 from each population) of size (100×6). We then dichotomize the 6 variables by imposing random thresholds from a uniform distribution in the interval $[-2+2]$. The mean absolute errors (MAEs) for the thresholds estimates for each variables (averaged over the 5,000 data sets and the 100 observations of each set) are always less than 0.06. Considering "difficult variables", originated by thresholds outside the interval $[-1+1]$, the MAEs increase to 0.11. These less accurate estimates also lead to larger errors for the scores estimates. Using (1), the mean absolute errors (MAEs) obtained for the

different ρ, averaged over the 500 data sets generated with each correlation matrix, the 100 observations of each set and the 15 correlation coefficients, are:

$\rho = 0$	$\rho = 0.2$	$\rho = 0.3$	$\rho = 0.4$	$\rho = 0.5$	$\rho = 0.6$	$\rho = 0.7$	$\rho = 0.8$	$\rho = 0.9$	$\rho = 0.95$
0.07	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.05	0.03

Results seem particularly accurate for all values of ρ. Mean errors also decrease as long as the real correlations among variables increase. Boxplots of the MAEs for the eigenvalues of the principal components, calculated between the eigenvalues of the correlation matrix **P** used to generate the data and the correlation matrix **R** of the generated data, are reported in Fig. 1. For values of ρ_{rc} not exceeding 0.8, estimates

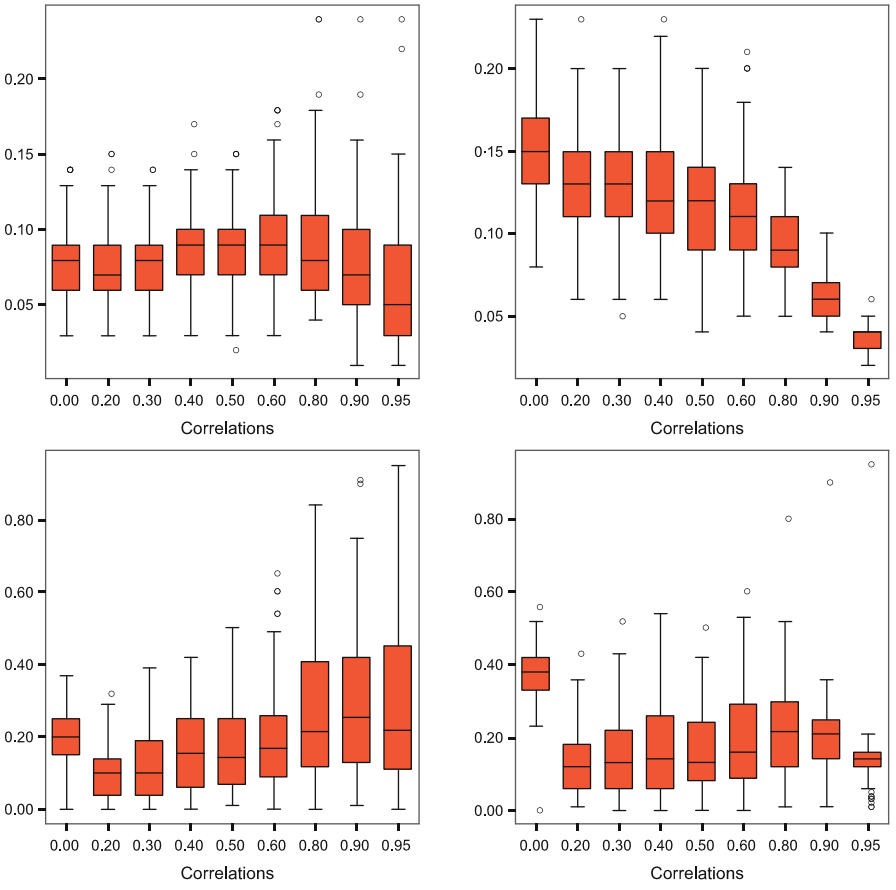


Fig. 1 Boxplots of the mean absolute errors of the eigenvalues plotted along the original correlations ρ_{rc}. In the left-hand boxes, errors are calculated between eigenvalues of the tetrachoric correlation matrix and eigenvalues of the matrix **R** of the generated data. In the right-hand boxes, errors are calculated between eigenvalues of the tetrachoric correlation matrix and eigenvalues of the matrix **P** used to generate the data. In the upper boxes errors are averaged over the six eigenvalues. In the lower boxes, errors are calculated only for the first eigenvalues

of all the eigenvalues better recover the computed correlation matrix, rather than the matrix used to generate the data. This is not true for the first eigenvalue: when this one is large (and the correlations are larger than 0.8) the estimates better recover the first eigenvalues of the matrix **P**. We then estimate the scores of each latent variable and of the principal components. The MAEs, averaged over the 500 data sets generated for each correlation matrix and over the six variables and the 100 observations of each set, are:

	$\rho = 0$	$\rho = 0.2$	$\rho = 0.3$	$\rho = 0.4$	$\rho = 0.5$	$\rho = 0.6$	$\rho = 0.7$	$\rho = 0.8$	$\rho = 0.9$	$\rho = 0.95$
MAE $(\tilde{t}_{ij})$	0.87	0.70	0.69	0.65	0.64	0.60	0.57	0.51	0.45	0.42
MAE $(\tilde{y}_{ij})$	0.59	0.59	0.58	0.58	0.58	0.59	0.58	0.58	0.58	0.59

As long as the correlation among variables increases, there is an improvement in the principal components estimates. On the contrary, results regarding the latent variables do not seem to depend on ρ. Estimates of the scores of the latent variables show improvements in average accuracy when the generated thresholds are close to zero, that is are close to the mean (and the median) of the latent variables. When the thresholds are beyond the range $[-1 + 1]$, average errors are significantly greater. Average errors, however, are always less than the variance of each variable and results seem enough accurate. Table 1 reports the MAEs of the latent variables scores obtained in a further study. Here the 6 binary variables are obtained by generating a (5000×6) data set from the same zero-mean multivariate normal populations as before, but with fixed thresholds: -2, -0.5, 0, 0.2, 0.5, 2. Average errors (in the last row) show that the accuracy of the EAP estimates increases as long as the threshold approaches zero. On the other hand, considering the errors computed for different values of the true scores, we note that minima average errors (reported in bold) are obtained for values near the thresholds. The worst fittings are obtained for large positive values when the threshold is -2 and for large negative values when the threshold is $+2$. For variables with thresholds -0.2, 0, 0.2 and 0.5, the correlations between real and estimated scores are 0.74, 0.78, 0.78 and 0.75, respectively.

Table 1 Mean Absolute Errors for the estimates of the 6 latent variables scores, divided into 9 groups. Groups are based on the magnitude of the true score values

	thresh.$= -2$	thresh.$= -0.5$	thresh.$= 0.0$	thresh.$= 0.2$	thresh.$= 0.5$	thresh.$= 2$
scores	MAEs	MAEs	MAEs	MAEs	MAEs	MAEs
< -1.3	0.62	1.02	1.46	1.62	1.89	2.74
$[-1.3 - 0.8)$	**0.20**	0.32	0.77	0.92	1.21	1.99
$[-0.8 - 0.5)$	**0.17**	**0.11**	0.40	0.56	0.84	1.60
$[-0.5 - 0.3)$	0.48	**0.23**	**0.11**	0.27	0.54	1.28
$[-0.3 + 0.0)$	0.75	**0.08**	**0.17**	**0.08**	0.29	1.00
$[+0.0 + 0.3)$	1.02	0.29	**0.15**	**0.23**	**0.08**	0.73
$[+0.3 + 0.5)$	1.30	0.55	**0.13**	**0.09**	**0.19**	0.47
$[+0.5 + 0.8)$	1.61	0.83	0.41	**0.22**	**0.13**	**0.18**
$\geq +0.8$	2.40	1.54	1.14	0.96	0.64	**0.41**
average	0.77	0.53	0.48	0.48	0.50	0.75

We then consider the internet advertisement data set from the UCI machine learning depository (http://archive.ics.uci.edu/ml/). The features encode the geometry of the image as well as phrases occurring in the URL, the image's URL and alt text, the anchor text, and words occurring near the anchor text. The cluster membership of each image is known (clusters are: advertisement or not advertisement). After removing instances with missing values and selecting binary variables with relative frequencies higher than 0.1, we reach a data set with 2,359 instances, 3 continuous variables and 10 binary variables. We perform a k-means cluster analysis and we run the mixture model implemented in SPSS first with mixed mode variables (normalizing continuous variables in the interval [0 1]) and then with all continuous variables (with the estimated scores of the binary ones). The classification error rate decrease from 33 to 30% with k-means and from 35 to 32% with the mixture model.

4 Concluding Remarks

Although it is clearly impossible to generalize from the results presented, it does appear that estimating the scores of the latent continuous variables generating the binary values may improve the clustering results and, above all, it allows fully Gaussian models with different correlations among the variables in each group to be used for classification. This paper describes an initial investigation into the feasibility of estimating the scores of each latent continuous variable. In literature, only EAP estimates of the most relevant factors have been presented, for the different aims of estimating composed items that are assumed to represent a particular set of constructs and for data reduction. Here the aim is to reach a continuous data matrix, of the same dimension of the original one. Possible variations and improvements to the method proposed are relevant topics for future research. Future simulations involve data generated from distributions rather than the normal, to explore whether the EAP estimates work well also in these cases. Indeed, although the threshold estimates are based on the normal distribution and the t_{ij} and the y_{ij} are supposed to be Gaussian, EAP estimates are little affected by the choice of this distribution since loadings and eigenvalues are not estimated by maximum likelihood.

References

1. Banfield, J.D., Raftery, A.E.: Model based Gaussian and non Gaussian clustering. Biometrics **48**, 803–821 (1993)
2. Bock, R.D., Mislevy, R.J.: Adaptive EAP estimation of ability in a microcomputer environment. Appl. Psychol. Meas. **6**(4), 431–434 (1982)
3. Celeux, G., Govaert, G.: Gaussian parsimonious clustering models. Pattern Recognit. **28**: 781–793 (1995)
4. Everitt, B.S.: A finite mixture model for the clustering of mixed mode data. Stat. Probab. Lett. **6**, 305–309 (1988)
5. Everitt, B.S., Merette, C.: The clustering of mixed-mode data: a comparison of possible approaches. J. Appl. Stat. **17**(3), 284–297 (1990)

6. Heckman, J.J.: Dummy endogenous variables in a simultaneous equation system. Econometrica **47**, 153–161 (1978)
7. Lawrence, C.J., Krzanowski, W.J.: Mixture separation for mixed-mode data. Stat. Comput. **6**, 85–92 (1996)
8. Manski, C.: Identification of binary response models. J. Am. Stat. Assoc. **83**, 729–738 (1988)
9. Muraki, E., Engelhard, G.: Full-information item factor analysis: application of EAP scores. Appl. Psychol. Meas. **9**(4), 417–430 (1985)
10. Skrondal, A., Rabe-Hesketh, S.: Generalized Latent Variable Modeling. Chapman & Hall, London (2004)
11. Vermunt, J.K., Magidson, J.: Latent Gold User's Guide. Statistical Innovation, Belmont, MA (2000)
12. Vermunt, J.K., Magidson, J.: Latent class cluster analysis. In: Hagenaars, J.A., McCutcheon, A.L. (eds.) Applied Latent Class Analysis, pp. 89–106. Cambridge University Press, Cambridge (2002)

Optimal Scaling Trees for Three-Way Data

Valerio A. Tutore

Abstract The framework of this paper is developed on tree-based models for three-way data. Three-way data are measurements of variables on a sample of objects in different occasions (i.e. space, time, factor categories) and they are obtained when prior information play a role in the analysis.

Three way data can be analyzed by exploratory methods, i.e., the factorial approach (TUCKER, PARAFAC, CANDECOMP, etc.) as well as confirmatory methods, i.e., the modelling approach (log-trilinear association models, simultaneous latent budget models, etc.).

Recently, we have introduced a methodology for classification and regression trees in order to deal specifically with three-way data. Main idea is to use a stratifying variable or instrumental variable to distinguish either groups of variables or groups of objects. As a result, prior information plays a role in the analysis providing a new framework of classification and regression trees for three-way data.

In this paper we introduce a tree-based method based on optimal scaling in order to account of the presence of non-linear correlated groups of variables. The results of a real world application on Tourist Satisfaction Analysis in Naples will be also presented.

1 Introduction

Three-way data are data classified in three ways. Longitudinal data, i.e., are three way, because of repeated observation of the same variables on the same objects.

So far segmentation methods for classification and regression trees have been proposed as supervised approach to analyze data sets where a response variable and a set of predictors are measured on a sample of objects or cases. Classification and regression trees are a fundamental approach to data mining and prediction [1, 6]. In particular, they can be fruitfully used for either exploratory or confirmatory

V.A. Tutore (✉)
Department of Mathematics and Statistics, University of Naples Federico II, Naples, Italy
e-mail: v.tutore@unina.it

B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-13312-1_10,

analysis. A further extension is two-stage discriminant trees [11] based on multiple factorial splits.

From the exploratory point of view, in binary segmentation, the aim is to find the best split of a predictor to split the cases into two sub-groups in order to reduce the impurity of the response within each sub-group. The recursive splitting of the cases yields a tree structure. As an example, partitioning procedures such as two-stage segmentation and fast algorithm use the concept of global prediction of each explanatory variable and the local prediction due to each split of predictor categories [8].

Following the pioneer work [15] and further developments [16, 14], this paper provides the methodology for the analysis of three-way data characterized at the same time by sets of objects and sets of predictors. Therefore, data sets can be described by a cube, namely a set of variables (including both predictors and responses) is measured on a sample of objects in a number of distinct situations, also called occasions. Each slide of the cube is a two-way data matrix, i.e. units times variables. Typically, the occasions are associated to modalities of a categorical variable. Alternatively, a time variable could be also considered as well.

As an alternative to other methods for either exploratory analysis [7] or confirmatory analysis [9, 12, 13], main idea is to analyse this type of data with suitable methods for classification and regression trees. In the following we propose a partitioning procedure for exploratory trees dealing with three-way data.

2 The Data and the Two-Stage Splitting Criterion

The three ways of the data set are cases, attributes and situations, respectively. Let **D** be the three-way data matrix of dimensions N, V, Q, where N is the number of cases, objects or units, V is the number of variables, Q is the number of situations. Assume that the V variables can be distinguished into two groups, namely there are M predictor variables $X_1, \ldots, X_m, \ldots, X_M$ and C response variables $Y_1, \ldots, Y_c, \ldots, Y_C$ where $M + C = V$. The Q situations refer to modalities of a stratifying variable, which is called *instrumental variable*. Alternatively a time variable can be also considered for longitudinal data analysis.

Predictors can be of categorical and/or numerical type whereas responses can be either categorical or numerical, thus a distinction can be made between a classification problem and a regression problem respectively.

The two-stage splitting criterion for $C = 1$ can be defined as follows:

$$\max_m \sum_q \gamma_Y(t|_q X_m)\, p_Y(t|q) \tag{1}$$

$$\max_s \sum_q \gamma_Y(t|s)\, p_Y(t|q) \tag{2}$$

for $q = 1, \ldots, Q$ (i.e. subsamples), $m = 1, \ldots, M$ (i.e. predictors), $s = 1, \ldots, S$ (i.e. splitting variables), with $\sum_q p_Y(t|q) = 1$, where $\gamma_Y(t|_q X_m)$ is the global impurity proportional reduction measure of Y due to each predictors X_m and $\gamma_Y(t|s)$

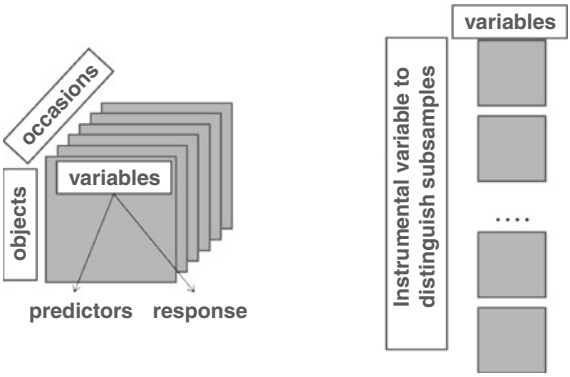


Fig. 1 The structure of three-way data

the local impurity proportional reduction measure of Y due to each splitting variable s. The first one is a weighted average of the measures calculated across the Q occasions. A suitable weighting system $p_Y(t|q)$ can be given by the percentage of the total impurity of the response in each subsample.

Analogously, it can be defined the local impurity proportional reduction measure due to each splitting variable. On the basis of the type of response variables, we can choose a suitable impurity measure for classification trees as well as for regression trees. In particular, impurity measures for classification can be measured by entropy, Gini index, misclassfication-ratio, whereas the ones for regression can be measured by variation or deviation.

In the following, we consider a special constrained version of the three-way data matrix **D**, as described in Fig. 1. In particular, the instrumental variable allows to distinguish subsamples such as groups of objects.

3 The Method

Let Y be the output, namely the response variable, and let $\mathbf{X} = \{X_1, \ldots, X_M\}$ be the set of M inputs, namely the predictor variables. In addition, let Z_O be the stratifying object variable with Q categories. The response variable is a nominal variable with J classes and the M predictors are all categorical variables (or categorized numerical variables). The sample is stratified according to the Q categories of the instrumental variable Z_O.

We assume that the M predictors are structured into K non-linear internally correlated groups of variables. To deal with non-linear correlated groups of variables we consider Nonlinear Canonical Correlation Analysis and the approach of Gifi [3].

This allows to summarize the information within each group through a latent factor. Nonlinear Canonical Correlation Analysis is applied before the segmentation procedure. Optimal scaling means that for each categorical variable a nonlinear transformation is permitted, so that it maximizes the analysis criterion [2, 17]. We find the NLCCA's object scores minimizing

$$\begin{gathered} \sum_{t=1}^{K} tr\,(\mathbf{X}-\mathbf{Q}_t\mathbf{A}_t)'\,(\mathbf{X}-\mathbf{Q}_t\mathbf{A}_t) \\ \mathbf{X}'\mathbf{X}=n\mathbf{I},\,\mathbf{u}'\mathbf{X}=0,\,\mathbf{q}=\mathbf{f}(\mathbf{h}),\,\mathbf{f}\in C(\mathbf{h}) \end{gathered} \tag{3}$$

where $\mathbf{X}$ are object scores, $\mathbf{Q_t}$ are the transformed variables from the original variable matrix $\mathbf{H}$ and $\mathbf{A_t}$ are the collection of multiple and single category quantifications across variables and sets, $C(\mathbf{h})$ is the set of possible transformations of $\mathbf{h}$, $\mathbf{f}$ refers to a transformation. In this way we found category quantifications of the sets of variables, our new predictors, i.e. the latent factors υ_k for $k = 1, \ldots, K$, to be considered in the recursive partitioning.

The above approach allowed to reduce the dimensionality of the analysis, shifting the attention toward a set of latent predictors synthesis of the original variables. Therefore, the latent predictors allowed to define the set of all possible splits as candidates for the best split of the objects. In this field Tutore et al. [15] introduced Partial Predictability Trees based on the use of predictability indexes for three-way cross-classifications. The idea is to extend the nonlinear transformation to data structured into three ways with different groups of individuals.

We consider the two-stage splitting criterion based on the predictability τ index of Goodman and Kruskal [4] for two-way cross-classifications: in the first stage, the best group of predictors is found maximizing the global prediction with respect to the response variable; in the second stage, the best split of the best group of predictors is found maximizing the local prediction.

In the following, we extend this criterion in order to consider the predictability power explained by each group/split with respect to the response variable conditioned by the instrumental variable Z_O.

For this purpose, we consider the predictability indexes used for three-way cross-classifications, namely the multiple τ_m and the partial τ_p predictability index of Gray and Williams [5], that are extensions of the Goodman and Kruskal τ_s index.

At each node, in the first stage, among all available groups of predictors υ_k for $k = 1, \ldots, K$, we maximize the partial index $\tau_p(Y|\upsilon_k, Z_O)$ to find the best predictor υ^* conditioned by the instrumental variable Z_O:

$$\tau_p(Y|\upsilon_k, Z_O) = \frac{\tau_m(Y|\upsilon_k Z_O) - \tau_s(Y|Z_O)}{1 - \tau_s(Y|Z_O)} \tag{4}$$

where $\tau_m(Y|X_m Z_O)$ and $\tau_s(Y|Z_O)$ are the multiple and the simple predictability measures and υ_g are the object scores of the groups constructed by correlated original variables. In the second stage, we find the best split s^* of the best predictor υ^* maximizing the simple index $\tau_s(Y|s, Z_O)$.

4 The Analysis

In this section, we present an application about tourism satisfaction survey in Naples. This survey of $N = 1{,}878$ tourists has been collected measuring the level

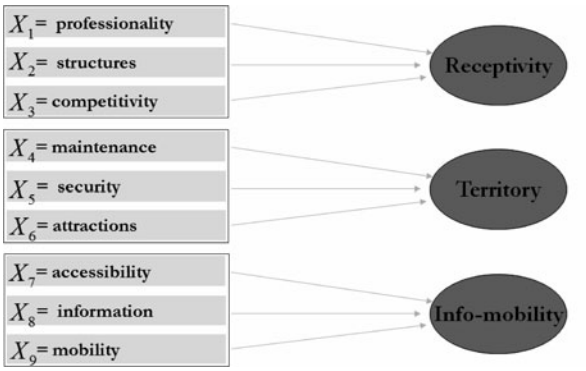


Fig. 2 Original variables and latent variables

of global satisfaction and the level of satisfaction with respect to three dimensions of the service, each considering three aspects.

The ordinal predictors (professionality, structures, competitivity, maintenance, security, attractions, accessibility, information, mobility) have 5 levels of satisfaction. Fig. 2 shows the original variables and the new latent variables obtained by correlated original variables.

The response variable has two classes distinguishing the satisfied and the unsatisfied tourists. The three dimensions are accommodation, territory, info-mobility. We choose as instrumental variable Z_O the nationality of the tourists with three categories (Italian, European, Extra-European). Fig. 3 shows the final binary tree

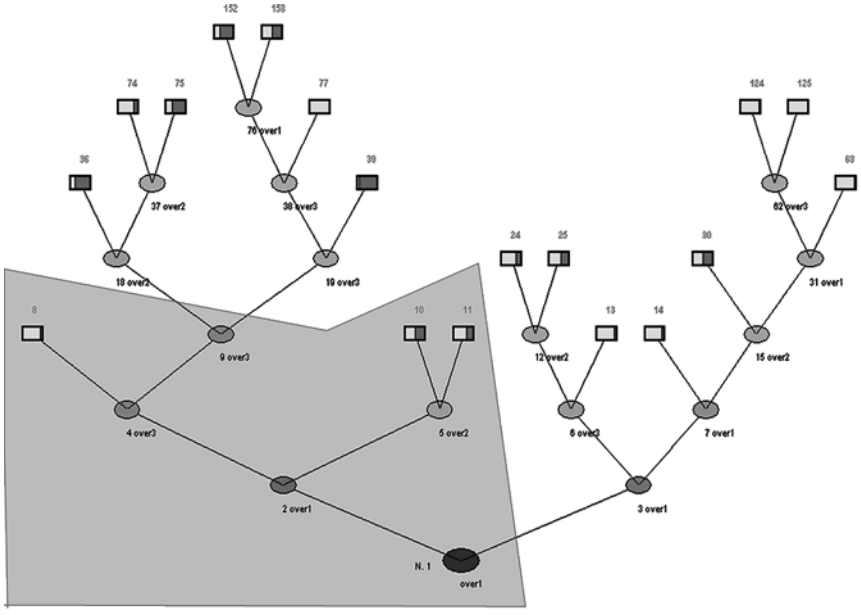


Fig. 3 Tree Graph - Tourism Satisfaction Survey

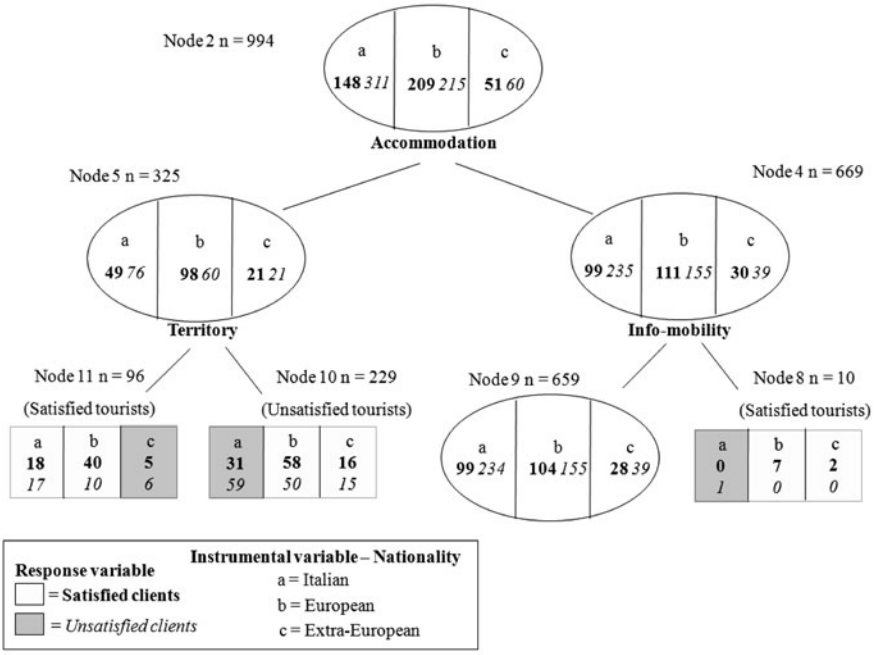


Fig. 4 Path 1-11 - Tourism Satisfaction Survey

with 18 terminal nodes in which we marked the predictor and the split at each nonterminal node.

In the end, Fig. 4 shows the path from root node to node $n = 11$. In particular, in Fig. 4 we report the response classes distribution of the objects within the three strata of Z_O, for the predictor selected in each nonterminal node.

In each node we indicate the total number of subjects, the number of subjects divided into the three categories of the instrumental variable and for each category we report in bold the number of satisfied tourists and in italics the number of unsatisfied tourists. For the terminal nodes we indicate with two different colours the presence of satisfied or unsatisfied tourists.

As an example, we can see that in node 10 there is a bigger presence of unsatisfied than satisfied tourists, but relatively to z_2 and z_3 there are more satisfied than unsatisfied tourists. Then it's possible to interpret in a different way that terminal node respect to the presence of nationality of the tourists.

5 Conclusions

In a previous work [15], we presented a partitioning procedure for a data mining problem, consisting in find a tree-based model for the analysis of a large set of within group correlated predictors. The goal was to define a suitable variable to

summarize each group of predictors. Two results were found: a multiple split was considered at each node of the tree and it was possible to understand the relevance of each group and not of the single original variable in the splitting procedure. This procedure was called optimal scaling tree.

In another work [16], we introduced a stratifying variable in the analysis accounting of prior information through specific constraints upon objects. With this method, called Partial Predictability Tree, the stratifying variable allowed to distinguish groups of objects, thus a suitable splitting criterion has been defined to find the best simultaneous partition of the objects. Main issue was to provide distinct response distributions in each node sample and the best split was found as a compromise of the impurity reduction of all response class distributions in the distinct subsamples.

This paper combines the two main ideas of the previous works providing optimal scaling trees using an instrumental variable for the analysis of three-way data. The goal is to understand the structure of the data in presence of complex data with blocks of objects as well as blocks of predictors. An application on a real data set has been briefly described in order to show the advantages of our approach. The results of several applications have been very promising, showing that our methodology:

1. allows to understand the partial dependence structure on subsamples within terminal as well as nonterminal nodes;
2. works with a few number of latent factors playing the role of predictors and, then, it decreases redundant information;
3. overcomes the limits of classical approach in the analysis of data structured in complex way.

The procedure has been implemented in MATLAB environment enriching the Tree Harvest Software [10] developed by the research unit in Naples). We are developing it as an alternative to standard tree-based methods for special structures of data.

References

1. Breiman, L., Friedman, J.H., Olshen, R.A., Stone. C.J.: Classification and Regression Trees. Wadsworth, Belmont, CA (1984)
2. De Leeuw, J., Young, F.W., Takane, Y.: Additive structure in qualitative data: an alternating least square method with optimal scaling features. Psychometrika **31**, 33–42 (1976)
3. Gifi, A.: Nonlinear Multivariate Analysis. Department of Data Theory, University of Leiden, Leiden (1981)
4. Goodman, L.A., Kruskal, W.H.: Measures of association for cross-classification. J. Am. Stat. Assoc. **48**, 732–762 (1954)
5. Gray, L.N., Williams, J.S.: Goodman and Kruskal's tau b: multiple and partial analogs. In: Proceedings of the American Statistical Association, pp. 444–448 (1975)
6. Hastie, T., Friedman, J.H., Tibshirani, R.: The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer, New York, NY (2001)
7. Kiers, H.A.L.: Hierarchical relations among three-way methods. Psychometrika **56**(3), 449–470 (1991)
8. Mola, F., Siciliano, R.: A two-stage predictive splitting algorithm in binary segmentation. In: Dodge, Y., Whittaker, J. (eds.) Computational Statistics: COMPSTAT 92, 1, pp. 179–184. Physica Verlag, Heidelberg (1992)

9. Siciliano, R.: Latent budget trees for multiple classification. In: Vichi, M., Optitz, P. (eds.) Classification and Data Analysis: Theory and Application, Springer, Heidelberg (1999)
10. Siciliano, R., Aria, M., Conversano, C.: Harvesting trees: methods, software and applications. In Proceedings in Computational Statistics: 16th Symposium of IASC, held Prague, August 23–27, 2004 (COMPSTAT2004), Eletronical Edition (CD) Physica-Verlag, Heidelberg (2004)
11. Siciliano, R., Mola, F.: Discriminant Analysis and Factorial Multiple Splits in Recursive Partitioning for Data Mining. In: Roli, F., Kittler, J. (eds.) Proceedings of International Conference on Multiple Classifier Systems, Chia, June 24–26, 2002, Lecture Notes in Computer Science, pp. 118–126. Springer, Heidelberg (2002)
12. Siciliano, R., Mooijaart, A.: Three-factor association models for three-way contingency tables. Comput. Stat. Data Anal. **24**(3), 337–356. Elsevier North Holland, Amsterdam (1997)
13. Siciliano, R., Van Der Heijden, P.G.M. Simultaneous Latent Budget Analysis of a Set of Multidimensional Contingency Tables. Metron **LII**(1–2), 155–180 (1994)
14. Siciliano, R., Tutore, V.A., Aria, M.: 3Way trees. In: Classification and Data Analysis 2007, Macerata, 12–14 Sept 2007. Book of Short Papers, pp. 231–234. EUM, Macerata (2007)
15. Tutore V.A., Mooijaart, A.: Optimal scaling trees. In: Classification and Data Analysis 2007, Macerata, 12–14 Sept 2007. Book of Short Papers, pp. 359–362. EUM, Macerata (2007)
16. Tutore, V.A., Siciliano, R., Aria, M.: Conditional Classification Trees using Instrumental Variables. Advances in Intelligent Data Analysis, pp. 163–173. Springer, Berlin Heidelberg (2007)
17. Van de Burg, E.: Nonlinear Canonical Correlation and Some Related Techniques. DSWO Press, Leiden (1988)

Part III
Data Mining

A Study on Text Modelling via Dirichlet Compound Multinomial

Concetto Elvio Bonafede and Paola Cerchiello

Abstract This contributions deals with a generative approach for the analysis of textual data. Instead of creating heuristic rules for the representation of documents and word counts, we employ a distribution able to model words along text considering different topics. In this regard, following Minka proposal [5], we implement a Dirichlet compound Multinomial distribution that is a mixture of random variables over words and topics. On the basis of this model we evaluate the predictive performance of the distribution by using seven different classifiers and taking into account the count of words in common between text document and reference class.

1 Introduction

With the rapid growth of on-line information, text categorization has become one of the key techniques for handling and organizing data in textual format. Text categorization techniques are an essential part of text mining and are used to classify new documents and to find interesting information contained within several on-line web sites. Since building text classifiers by hand is difficult, time-consuming and often not efficient, it is worthy to learn classifiers from examples.

In this proposal we employ a generative approach for the analysis of textual data. Thus, following Minka [5] and Madsen et al. [3] proposals, in Sec. 2 we develop a "Dirichlet Compound Multinomial" (DCM) distribution that is a mixture over words and topics, and we show how to estimate the parameters of the models.

Then, in Sec. 3, we have the application and the predictive performance of the distribution by using seven different classifiers. Conclusions are Sec. 4.

C.E. Bonafede (✉)
Univesity of Pavia, Pavia, Italy
e-mail: ingc.bonafede@gmail.com

B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-13312-1_11,

2 Background: The Dirichlet Compound Multinomial

The DCM distribution introduced by Minka is an hierarchical model: on one hand, the Dirichlet random variable is devoted to model the Multinomial parameters θ; on the other hand, the Multinomial variable models the words count vectors comprising the document. The distribution function of the DCM mixture model is:

$$p(\bar{x}|\alpha) = \int_{\theta} p(\bar{x}|\theta)p(\theta|\alpha)d\theta. \tag{1}$$

where $p(\theta|\alpha)$ is the Dirichlet distribution:

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{w=1}^{W}\alpha_w)}{\prod_{w=1}^{W}\Gamma(\alpha_w)}\prod_{w=1}^{W}\theta_w^{\alpha_w-1} \tag{2}$$

with θ_w the probability of emitting a word w and α_w the Dirichlet parameters for each word; thereby the whole set of words (bag-of-words) is modelled. The expression "bag of words" is typical in text modelling context and refers to the words present in a corpora considered as an unordered vector, disregarding grammar. Instead $p(\bar{x}|\theta)$ is the Multinomial distribution:

$$p(\bar{x}|\theta) = \frac{n!}{\prod_{w=1}^{W}x_w}\prod_{w=1}^{W}\theta_w^{x_w} \tag{3}$$

in which $\bar{x}$ is the words' count vector and x_w is the count for each word.

Thus a text (a document in a set) is modelled as a "bag-of-bags-of-words", (see [3, 5]) and developing the previous integral we obtain:

$$p(\bar{x}|\alpha) = \frac{n!}{\prod_{w=1}^{W}x_w}\frac{\Gamma(\sum_{w=1}^{W}\alpha_w)}{\Gamma(\sum_{w=1}^{W}(x_w+\alpha_w))}\prod_{w=1}^{W}\frac{\Gamma(x_w+\alpha_w)}{\Gamma(\alpha_w)}; \tag{4}$$

From another point of view we can state that in the DCM model the Dirichlet represents a general topic that compound the set of documents and each Multinomial, linked to specific sub-topics, make the emission of some words more likely than other for a specific document. Thus the DCM could be also described as "bag-of-scaled-documents".

Moreover the added value of the DCM approach consists in the ability to handle the "burstiness" of a rare word without introducing heuristics [6]. Burstiness is the tendency of rare word appearing many times in a single document; if a word does appear once, it is much more likely to appear again, i.e. words appear in bursts.

When we consider the entire set of documents (D) where each document is independent and identified by its count vector, ($D = \{\bar{x}_1, \bar{x}_2, ..., \bar{x}_N\}$), the likelihood of the whole documents set (D) is:

$$p(D|\alpha) = \prod_{d=1}^{N} p(\bar{x}_d|\alpha)$$
$$= \prod_{d=1}^{N} \left(\frac{\Gamma(\sum_{w=1}^{W} \alpha_w)}{\Gamma(x_d + \sum_{w=1}^{W} \alpha_w)} \prod_{w=1}^{W} \frac{\Gamma(x_{dw} + \alpha_w)}{\Gamma(\alpha_w)} \right); \quad (5)$$

where x_d is the sum of the counts of every word in the document d-th ($\sum_{w=1}^{W} x_{dw}$) and x_{dw} the count of word w-th for the document d-th. The log-likelihood and the gradient necessary to find the parameters are respectively:

$$log(p(D|\alpha)) = \sum_{d}^{N} \left(log\Gamma \left(\sum_{w=1}^{W} \alpha_w \right) - log\Gamma \left(x_d + \sum_{w}^{W} \alpha_w \right) \right) +$$
$$+ \sum_{d}^{N} \sum_{w}^{W} (log\Gamma(x_{dw} + \alpha_w) - log\Gamma(\alpha_w)) \quad (6)$$

$$g_w = \frac{\partial log(p(D|\alpha)}{\alpha_w} =$$
$$= \sum_{d}^{N} \left(\Psi \left(\sum_{w}^{W} \alpha_w \right) - \Psi \left(x_d + \sum_{w}^{W} \alpha_w \right) + \Psi(x_{dw} + \alpha_w) - \Psi(\alpha_w) \right) \quad (7)$$
$$with \ \Psi(z) = \frac{d\Gamma(z)}{dz} = digamma \ function$$

Now we have the task of maximizing the log-likelihood and finding the parameters. Among different methods we have chosen the fixed-point iteration. Such a method has its roots in the the Expected Maximization (EM) algorithm (see [2]) which can be built up in different ways.

One possibility is to see the EM as a lower bound maximization where we alternate the E-step to calculate an approximation of the lower bound for the log-likelihood and maximize it in the M-step until a stationary point (zero gradient) is reached (see [4]).

However if we are able to find a lower bound for the log-likelihood we can maximize it via a fixed-point iteration in fact it is the same principle of considering the EM as a lower bound maximization (see [4, 5]).

So for the DCM the lower bound with $log(p(D|\alpha))$ is the following quantity:

$$log(p(D|\alpha)) \geq -(\textstyle\sum_{w}^{W} \alpha_w - 1) \sum_{d}^{N} [b_d + \sum_{w}^{W} a_{dw} log\alpha_w] + (const.) \quad (8)$$
$$where \ b_d = \Psi(x_d + \textstyle\sum_{w}^{W} \alpha_w) - \Psi(\sum_{w}^{W} \alpha_w)$$
$$and \ a_{dw} = (\Psi(x_{dw} + \alpha_w) - \Psi(\alpha_w))\alpha_w$$

this allows us to use a fixed point iteration which steps are:

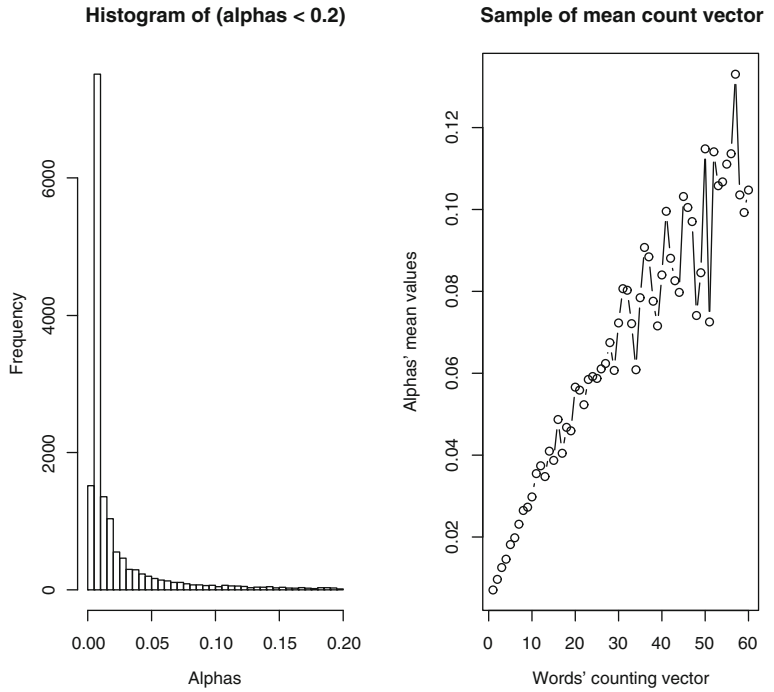


Fig. 1 Average alpha versus count vector and alphas' histogram

$$\alpha_w^{k+1} = \alpha_w^k \frac{\sum_d^N \left[\Psi(x_{dw} + \alpha_w^k) - \Psi(\alpha_w^k)\right]}{\sum_d^N \left[\Psi(x_d + \sum_w^W \alpha_w^k) - \Psi(\sum_w^W \alpha_w^k)\right]} \tag{9}$$

where x_d is the sum of the counts of each word in the document d-th ($\sum_{w=1}^W x_{dw}$), x_{dw} the count of word w-th for the document d-th and α_w^k the Dirichlet coefficient for word w at the k-th step. The algorithm is stopped when a degree of approximation ε is reached; in our case we have used an ε equal to 10^{-10}.

The parameters found out, as said before, have an important characteristic: they follow the "burstiness" phenomenon of words. In fact the smaller an α_w is, the more "burstinesss" effect is contained within a word, as revealed in Fig. 1.

3 Application

With the scope of analyzing the performance of the classification procedure, we have used the Reuter-21,578[1] data set which contains 21,578 documents identified by the following attributes: topics, places, peoples, organizations, exchanges and companies.

[1] http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html

For classification purposes, we consider topics as document classes to predict. The original topics are 135 and most of the documents are unknown, thus we employ 46 classes (topics) characterizing 8,045 documents. We use this data set according two different approaches:

1. First of all we evaluate the best classifier, based on the parameters extracted from the DCM distributions, by considering the whole documents set divided into training (with 80% of documents) and test set distributed over 46 classes. Thus there are 6,436 training documents with a vocabulary (already filtered and stemmed) of 15,655 words.
2. Secondly we compare DCM to sbDCM by using a training data set with only 2,051 documents containing a vocabulary of 4,096 words.

In the first we have used three kind of classifiers developed in Rennie et al. [6] and their four compositions so to have seven different classifiers to be tested and to evaluate their performance. These classifiers select the document class with the highest posterior probability:

$$l(d) = argmax_c \left[log\ p(\theta_c) + \sum_{w=1}^{N} f_w\ log\ \theta_{cw} \right] \tag{10}$$

where f_w is the frequency count of word w in a document, $p(\theta_c)$ is a prior distribution over the set of topics (that we consider uniformly distributed) and $log(\theta_{cw})$ is the weight for word w.

The weight for each class is estimated as a function of alpha coefficients:

$$\hat{\theta_{cw}} = \frac{N_{cw} + \alpha_w}{N_c + \sum_{w=1}^{N_c} \alpha_w} \tag{11}$$

where N_{cw} is the number of times word w appears in the documents of class c, N_c the total number of words occurrences in class c.

Rennie et al. [6] propose three main classifiers which are the normal (N), the complement (C) and the mixed (M) ones:

1. Normal:

$$l(d) = argmax_c \left[log\ p(\theta_c) + \sum_{w=1}^{N} f_w\ log\ \frac{N_{cw} + \alpha_w}{N_c + \sum_{w=1}^{N_c} \alpha_w} \right]; \tag{12}$$

2. Complement Version (COMP):

$$l(d) = argmax_c \left[log\ p(\theta_c) - \sum_{w=1}^{N} f_w\ log\ \frac{N_{\tilde{c}w} + \alpha_w}{N_{\tilde{c}} + \sum_{w=1}^{N_{\tilde{c}}} \alpha_w} \right]; \tag{13}$$

where $N_{\tilde{c}w}$ is the number of times word w occured in documents in all classes excepted c and $N_{\tilde{c}}$ is the total number of word occurrences in classes other than c.

3. Mixed:

$$l(d) = argmax_c \left[log\ p(\theta_c) + \sum_{w=1}^{N} f_w\ log\ \frac{N_{cw} + \alpha_w}{N_c + \sum_{w=1}^{N_c} \alpha_w} - \sum_{w=1}^{N} f_w\ log\ \frac{N_{\tilde{c}w} + \alpha_w}{N_{\tilde{c}} + \sum_{w=1}^{N_{\tilde{c}}} \alpha_w} \right]; \tag{14}$$

3.1 Performance of the Dirichlet Compound Multinomial

In this section we describe the evaluation performed on different classifiers by using the parameters estimated from the DCM distribution. Thus, our training data set is compound of 6,436 documents with a vocabulary (already filtered and stemmed) of 15,655 words so we have to estimate 15,655 $\alpha's$. The alpha is able to model the "burstinesss" of a word in fact the smaller the α parameters are, the more bursty the emission of words is. This phenomenon is characteristic of rare words, therefore $\alpha's$ coefficients are, on average, smaller for less counted words. This is showed in Fig. 1 where there are displayed the mean values of an α for each word count appearing in the document collection and the relative histogram. The average value of overall $\alpha's$ is 0.0,342, the standard deviation is 0.1,087 and maximum and minimum values are respectively 6.6,074 and 0.0,025. As we can see from the histogram of $\alpha's$ the document collection is characterized by bursty words.
Once obtained coefficient $\alpha's$ we employ seven different classifiers, three of which are described in Rennie et al. [6] (normal (N), complement (C) and mixed (M)). The remaining ones are proposed as the appropriate combination of the previous ones, in order to improve their characteristics.

In fact the new four classifiers are set in function of the number of words that a test-document has in common with the set of documents that compound a class; in this way we create a classifier in function of the number of words in common. Thus we analyze the following additional classifiers: Complement + Mixed + Normal (CMN), Complement + Normal (CN), Complement + Mixed (CM), Mixed + Normal (MN).

In order to evaluate the classification performance we employ three performance indexes:

1. *(Ind1)* The proportion of true positive over the total number of test-documents.

$$\left(\sum_{d=1}^{D} \frac{TP_d}{D} \right) \times 100;$$

2. *(Ind2)* The proportion of classes that we are able to classify.

$$\left(\sum_{c=1}^{C} \frac{I_c}{C} \right) \times 100;$$

where I_c is an indicator that we set 1 if at least one document of the class is classified correctly, otherwise we set 0.

3. *(Ind3)* The proportion of true positive within each class over the number of test documents present in the class.

$$\left(\frac{1}{C}\sum_{c=1}^{C}\frac{TP_c}{M_c}\right)\times 100;$$

where M_c is the number of test-documents in each class, TP_c is the number of true positive in the class and C the number of classes (46).

With regards to the last four composite classifiers, such indexes have been calculated by varying the number of words in common between the test document and the class. In particular for our test we have used three different thresholds for the number of words (n): 15, 10 and 5. For example, we indicate with the initials CM_n the classification rule that employs classifier C when the number of common words are less or equal to n and classifier M when the number of words in common is more than n. Instead, the initial $CMN\_n.m$ identifies the using of classifier C until n, the classifier N over m and the classifier M between n and m.

For the data at hand the number of words in common between the two sets (training and evaluation set) varies between 1 and 268.

The above mentioned combination is based on the following idea: if the number of words in common between the bag of words and the correct class is low, then the most informative content is in the complement set. Otherwise the needed information is contained either in the normal set or in the complement one. Taking into account such consideration we have set up different combination and we concluded that the useful trade-off among classifiers is equal to 10 (see Table 1).

The results are reported in Table 1. As we can see the best classifiers are the mixed and the CM_10 ones.

Table 1 Comparison among classifiers

Classifier	*Ind1*	*Ind2*	*Ind3*
Normal	73.46%	100%	66.74%
Comp.	66.93%	39%	10.26%
Mixed	76.88%	100%	66.79%
CM_5	76.88%	100%	66.79%
CM_10	76.94%	100%	66.84%
CM_15	76.13%	100%	65.15%
CMN_10.50	75.14%	100%	67.18%
CMN_10.152	76.69%	100%	67.11%
CMN_10.200	76.81%	100%	66.80%
CN_5	73.65%	100%	66.74%
CN_10	73.65%	100%	66.83%
CN_15	72.90%	100%	65.16%
NM_5	73.46%	100%	66.74%
NM_10	73.58%	100%	66.78%
NM_15	73.64%	100%	66.82%

These are able to classify respectively 1,237 and 1,238 over 1,609 documents that are distributed not uniformly among classes (46). These classifiers are able to classify at least a document per class even if there are classes containing only two documents. Between them the CM_10 classifier has index three slightly better than mixed one. The worse classifier, in this case, is the complement version alone. From the reported results we can conclude that the DCM distribution is a valid approach for modelling textual data and it is worthy to further investigate on its characteristics.

Moreover the Log-Likelihhod (LL) and the corrected Akaike Information Criterion (AICc)[2] before and after the optimization procedure are respectively: LL from $-222,385$ to $-205,286$ and AICc from 454,264 to 420,066.

4 Conclusion

In this contribution we show the characteristics of the DCM distribution employable in the context of text analysis with the purpose of document classification. DCM main feature is the capability to take into account the "burstiness" phenomenon of rare words. With the α coefficients coming from the DCM distribution we implement a classification procedure which uses the Naive Bayes classifier. Among all the proposed classification rules, we have shown that the best performances are obtained by composition of complement classifier and mixed (i.e. we use the complement when we have less then 10 words in common and mixed for more than 10) and by the mixed alone.

Moreover the DCM distribution models each document as a "bag-of-scaled-document" where the Dirichlet random variable generates the general topic and the Multinomial one the specific sub-topics that compound the document. In Cerchiello and Bonafede [1] the DCM is developed and modified in order to insert directly unknown or known topics within the model by means of a new vector of parameters. Such information will be useful to expand and improve the application of this model.

Acknowledgments The paper is the result of the close collaboration among the authors, however Sec. 3 has been written by Elvio Bonafede, Secs. 1, 2 and 4 by Paola Cerchiello. This work has been supported by MUSING 2006 contract number 027097, 2006–2010 and FIRB, 2006–2009.

References

1. Cerchiello, P., Bonafede, E.C.: Dirichlet compound multinomials for text data analysis, Technical Report. http://www-3.unipv.it/dipstea/workingpapers.php (2010)
2. Dempster, A.P., Laird, M.N., Rubin, D.B.: Maximum likelihood from incomplete data via the EM Algorithm. J. R. Stat. Soc. Ser. B. (Methodological) **39**(1), 1–38 (1977)

[2] The AICc has the following formula: $AICc = -2log(LL)+2k+\frac{2k(1+k)}{(n-k-1)}$ where k is the number of parameters and n the number of observations.

3. Madsen, R. E., Kauchak, D., Elkan, C.: Modeling word burstiness using the Dirichlet distribution. In: Proceeding of the 22nd International Conference on Machine Learning, Bonn, Germany (2005)
4. Minka, T., Expectation-maximization as lower bound maximization, Technical Report. http://research.microsoft.com/minka/papers/em.html (1998)
5. Minka, T., Estimating a Dirichlet distribution, Technical Report. http://research.microsoft.com/en-us/um/people/minka/papers/dirichlet/ (2003)
6. Rennie, J.D.M., Shih, L., Teevan, J., Karger, D.R.: Tackling the Poor Assumptions of Naive Bayes Text Classifier. In: Proceeding of the 20th International Conference on Machine Learning, Washington, DC (2003)

Automatic Multilevel Thresholding Based on a Fuzzy Entropy Measure

D. Bruzzese and U. Giani

Abstract Histogram Thresholding is an image processing technique whose aim is that of separating the objects and the background of the image into non overlapping regions. In gray scale images this task is obtained by properly detecting, on the corresponding gray levels histogram, the valleys that space out the concentration of the pixels around the characteristic gray levels of the different image structures. In this paper, a novel procedure will be discussed exploiting fuzzy set theory and fuzzy entropy to find automatically the optimal number of thresholds and their location in the image histograms.

1 Introduction

Typical computer vision applications often require an image segmentation preprocessing step in order to extract the distinct objects enclosed in the image foreground. For intensity images, several approaches to image segmentation have been proposed in last years. According to the image features exploited, they can be broadly classified into three main classes: edge-based, region-based and cluster-based techniques [1].

Edge-based methods disclose objects by highlighting their contours, usually characterized by a sharp change in the intensity levels of neighboring pixels. The main advantage of such approach is that the edge representation of an image efficiently reduces the amount of data to be processed. However, the accuracy of the results can be seriously compromised if broken boundaries are present; in this case edge-linking techniques become necessary for contour filling (e.g. [7]). Furthermore, noise may result in erroneous edges and thus expensive preprocessing work has to be done in order to filter-out noisy edges.

Region-based methods, on the contrary, proceed by grouping adjacent pixels with uniform properties like grayscale, texture, and so forth. Thereafter a coarse-grained segmentation is obtained by merging adjacent regions according to the similarities

D. Bruzzese (✉)
Department of Preventive Medical Sciences, Federico II University, Naples, Italy
e-mail: dario.bruzzese@unina.it

B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-13312-1_12,

among the properties in these regions. Region-based segmentation usually produces coherent regions with no gaps due to missing edge pixels. However, while edge definition is quite simple to set, both from a logical and a formal point of view, the criteria for region memberships are more difficult to assess.

In histogram-based techniques, the image structures are elicited only by looking at the shape properties of the intensity level histogram. The main *assumption* on which these methods rely is that no spatial information is required to segment the image because the different objects can be uncovered by looking at the shape properties of the histogram. Actually, in a well-defined image, the corresponding histogram presents a deep valley between two peaks. Around these peaks the object and background gray levels are concentrated and the optimum threshold value must be located in the valley region. According to the number of objects that have to be recognized, bi-level or multi-level thresholding techniques are employed.

In this paper a novel procedure, exploiting Fuzzy Set theory in the context of multilevel histogram thresholding, will be discussed.

The paper is structured as follows. After a brief recall of Fuzzy Set theory and Fuzzy Histogram Thresholding techniques (Sec. 2), the proposed algorithm (Sect. 3) will be discussed and its application on several real and synthetic images will be shown (Sect. 4). The issues for discussion and for future research work will set out at the last section.

2 Fuzzy Set Theory and Histogram Thresholding

Let X be a universe of elements; a Fuzzy Set A is defined as in [11],

$$A = \{x, \mu_A(x) | x \in X, \mu_A(x) \in [0, 1]\} \tag{1}$$

where μ is called the membership function or grade of membership and measures the coherence of each $x \in X$ with the properties that characterize the fuzzy set A. In computer vision applications an image can be considered as an array of fuzzy singletons, the pixels, each having a value of membership denoting the grade of possessing some specific property (for example brightness) that depends on the problem to be solved.

After the pioneristic work of Murthy and Pal [5], several thresholding algorithms based on fuzzy set theory are reported in the literature (see [8] for exhaustive and up-to-date survey of image thresholding methods). In these approaches, usually, Fuzzy Set Theory intervenes at a double level. First, for each candidate threshold, the gray values have to be mapped to the fuzzy domain by using an appropriate membership function, and, in a second step, a measure of fuzziness for the whole image have to be computed. The optimum threshold is found by minimizing (maximizing) the index of fuzziness over the gray-level range.

Our proposal represents an extension of the procedure that was originally discussed in Huang and Wang [2] and afterwards refined in a subsequent work [3]. Here the authors propose a multilevel thresholding technique based on the optimization of a Fuzzy Classification Entropy which describes the fitness of the histogram to a

multimodal distribution. According to our opinion, the main drawback of the procedure relies on the fact that the number of thresholds is required as input parameter of the algorithm; moreover, the computational complexity of the searching phase, when the number of classes increases, requires to abandon an exhaustive search in favour of some sub-optimal search strategy.

The proposed algorithm, on the contrary, by using a divide et impera paradigm, iteratively applies the general scheme of Fuzzy Histogram Thresholding until a stopping rule is met. The stopping rule adopts a sensitivity criterion that allows to control the granularity of the resulting segmentation and gets automatically to find the optimal number of thresholds and their location in the image histograms. In the following section, the details of the proposed procedure will be outlined.

3 Automatic Multilevel Fuzzy Histogram Thresholding

In the description of the algorithm, the following notation will be adopted. Let $\Gamma = \{x_i, i = 1, .., L\}$ be a set of consecutive bin values, (e.g. in Image Analysis L=256 and $x_1 = 0, x_2 = 1, .., x_L = 255$), and let n_i be the corresponding bin frequency. Let $c = \{x_s, ..., x_r;\ s, r \in \{1, .., L\};\ s < r\}$ be a generic range (segment) of bin values. After an initialization step in which the whole set Γ is added to the set $\Im$ of candidate segments, the recursion begins.

Let $c^* = \{x_{s^*}, ..., x_{r^*}, s^*, r^* \in \{1, .., L\}, s^* < r^*\}$ be the range being processed; for each bin $t \in c^*$ the corresponding value is set as candidate threshold and the fuzzy membership function is computed by using the m-function, proposed in [2]

$$\mu(x_i) = \frac{C}{C + |x_i - \mu_r|}, \quad r = \begin{cases} 0 \text{ if } x_{s^*} \le x_i \le t \\ 1 \text{ if } t < x_i \le x_{r^*} \end{cases} \tag{2}$$

where μ_0 and μ_1 denote the averages of those bins, respectively, up to and from the t-th x value with $t \in \{1, ..., L\}$ and C is a normalization factor. Once the grades of membership have been obtained, a fuzzy entropy measure is computed [6].

$$H_t(\underline{\mu}) = k \sum_i \{\mu(x_i) \ln [\mu(x_i)] - [1 - \mu(x_i)] \ln [1 - \mu(x_i)]\} \times n_i \tag{3}$$

where the factor k constraints the fuzzy entropy in the closed interval [0,1].

The bin value that achieves the minimum of the measure of fuzziness, t^*, is added to the set of potential thresholds. In order to become effective, at least one of the two adjacent intervals branched off from t^* should present a relative increment of the fuzzy entropy larger than the sensitivity criterion ε. If this condition is met, the threshold t^* become effective and that (those) interval(s) with a relative increment of the fuzzy entropy larger than ε is (are) added to $\Im$. The segment in $\Im$ with the maximum entropy value is thus processed in the next cycle of the recursion. The procedure stops when the set of candidate segments become empty.

The pseudo-code of the algorithm is shown in the following box.

Input: Histogram-like structure $I = \{x_i, n_i\}$, Sensitivity Criterion ε
Output: Array of thresholds locations $\underline{t}$
Initialization: Set $c^* = \{x_1, ..., x_L\}$, add c^* to $\Im$
Step 1: For each $x_i \in c^*$, set $t = x_i$ and do
  Step 1.1: Compute the fuzzy membership values $\mu(x_i)$, $(x_i \in c^*)$, using eq. 2
  Step 1.2: Compute the fuzzy Entropy $H_t(\underline{\mu})$, using eq. 3
Step 2: Find $t^* = \arg\min_{t \in c^*} \left[H_t(\underline{\mu}) \right]$ and remove c^* from $\Im$
Step 3: Set $c_l = \{\min(c^*), ..., t^*\}$, $c_r = \{t^* + 1, ..., \max(c^*)\}$
Step 4: For each $k \in \{l, r\}$ set $c_k = c^*$ and compute t_k^* by applying **Step 1** and **Step 2**
Step 5: if $\exists\, k \in \{l, r\} : \mathit{StoppingRule}\ (c_k) = \mathit{False}$, then add c_k to $\Im$, add t_k^* to $\underline{t}$
Step 6: if $\sharp(\Im) > 0$ then find $c^* = \arg\max_{c \in \Im} \left[H_{t*}(\underline{\mu}) \right]$ and go to **Step 1** else return
StoppingRule: If $\frac{[H_{t_k^*}(\underline{\mu}) - H_{t*}(\underline{\mu})]}{H_{t*}(\underline{\mu})\}} > \varepsilon$ then $\mathit{StoppingRule}\ (c_k) = \mathit{False}$

With the intent of clarify the working mechanism of the proposed procedure, its application to a fluorescent microscope image of *Staphylococcus Aureus* bacteria is shown in Fig. 1. In this example and in the following applications the sensitivity criterion was set equal to 0.05.

Fig. 1 **(a)** original *Staphylococcus Aureus* image; **(b)** gray level histogram of image **(a)**; **(c-f)** steps of the proposed procedure.
In the Initialization phase, the entire gray level range is set as candidate segment and for each $t \in [0, 255]$ the fuzzy entropy is computed. The gray level achieving the minimum value of the fuzzy entropy is $t^* = 198$ (c). The two contiguous intervals $c_l = [0, 198]$ and $c_r = [199, 255]$ are thus processed by computing $H_t(\underline{\mu})$, $t \in c_l$ and $H_t(\underline{\mu})$, $t \in c_r$ (d). Since c_l shows a relative increment of the fuzzy entropy greater than ε, the threshold $t^* = 198$ become effective, c_l is added to $\Im$, while c_r is discarded because of the Stopping Rule.
In the next step, the only eligible candidate in $\Im$, $c^* = [0, 198]$, is processed. The potential threshold $t^* = 87$ (that was obtained in the previous step) splits c^* into the two sets $c_l = [0, 87]$ and $c_r = [88, 198]$. Applying steps 1 and 2 to both of them, it turns out that $c_r = [88, 198]$ should be discarded, while $c_r = [88, 198]$ could be added to $\Im$ because of a relative increment of the Fuzzy Entropy greater than ε; $t^* = 87$ is thus added to $\underline{t}$ (e).
Finally the segment $c^* = [88, 198]$ is processed. Because both the two sub-branches springing from the potential threshold $t^* = 128$ encounter the Stopping Rule (f) the procedure stops. The image is thus segmented using the threshold set $\underline{t} = \{87, 198\}$

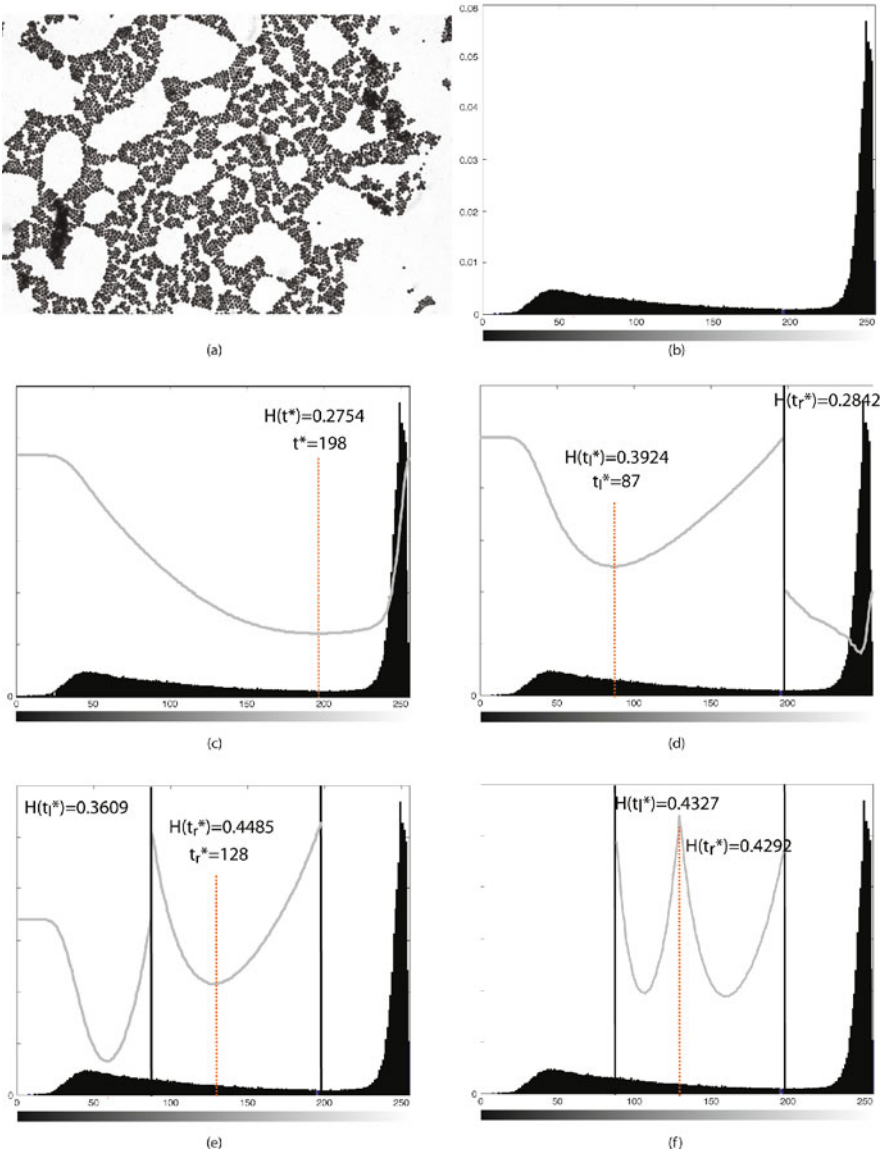


Fig. 1 (continued)

4 Experimental Results

The algorithm has been applied on different images (both real and artificially generated); the results have been compared with those obtained with a different automatic multilevel thresholding technique [9] where the optimum number of classes is

identified by minimizing a cost function that takes into account both the discrepancy between the thresholded and the original images and a penalty term proportional to the number of bits used to represent the thresholded image. The individual thresholds are instead chosen by optimizing the Total Correlation function of the image histogram [10].

Different quantitative evaluation criteria will be used to test the performance of the two methods. In case of synthetic image, where ground-truth information is available, the misclassification error (ME), calculated by comparing the thresholding results with the ideal results, will be computed. In case of real images, the Non Uniformity measure (NU) proposed by Levin and Nazif [4] will be adopted. This measure is defined as a weighted average of the within-class variance normalized by the variance of the overall distribution. NU ranges from 0 to 1 where values close to 0 correspond to a well segmented image.

In Fig. 2a, a synthetic image (400 × 800 pixels, 256 gray levels) with two rectangles of different size and different intensity levels (level 26 for the darker rectangle and level 128 for the brighter one) on a white background (gray level 255) is shown. Fig. 2b shows the same image after adding Gaussian noise with 0 mean and standard deviation equal to 15. The histogram of the noisy image is reported in Fig. 2c. Both the algorithms correctly identify the number of classes, but they differ in the location of the corresponding thresholds getting to very different results.

In Fig. 3 the thresholded images of the *Staphylococcus Aureus* bacteria obtained with the proposed algorithm (a) and with the procedure in [9] (b) are shown. As can be noticed the *fuzzy* segmentation, characterized by the threshold set $\underline{t} = \{87, 198\}$,

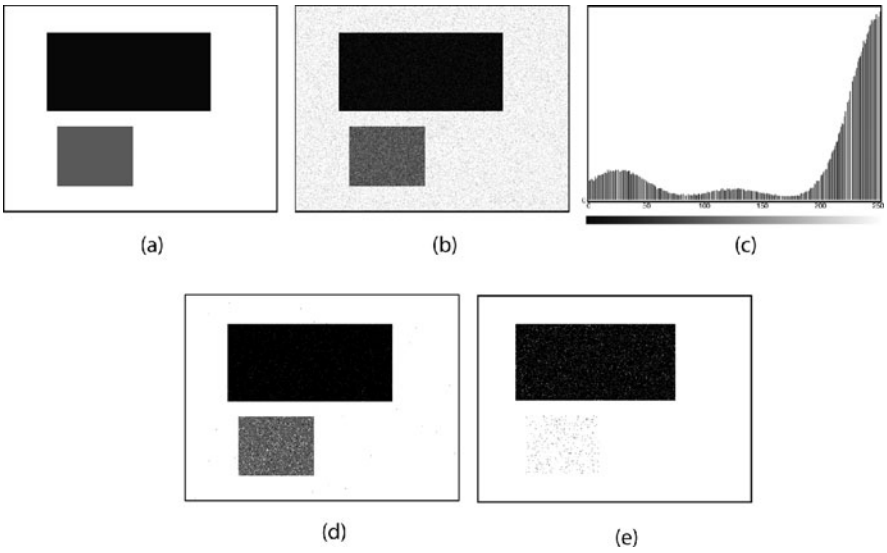


Fig. 2 (a) Original image: two rectangles of different intensity and different size on a white background, (b) noisy image, (c) histogram of the noisy image, (d)–(e) thresholding results of the noisy image by the proposed method ($\underline{t} = \{76, 161\}$, $\varepsilon = 0.05$, $ME = 0.0130$) and Sezgin and Taşaltín algorithm ($\underline{t} = \{49, 78\}$, $ME = 0.1135$)

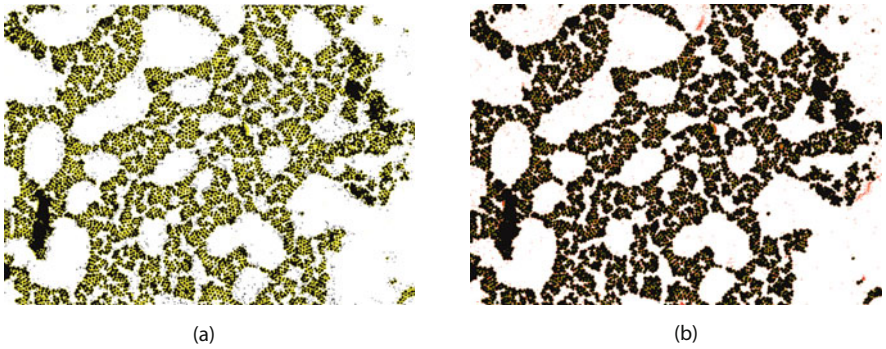


Fig. 3 (a) Thresholded image using the proposed approach: $\underline{t} = \{87, 198\}$, NU=0.0325. (b) Thresholded image using Sezgin and Taşaltín algorithm: $\underline{t} = \{133, 179, 228\}$, NU=0.0425

discloses both the bacteria and the fluid structure linking them; both these *objects* are covered by the darker gray levels, but they were not recognizable by looking at the original image histogram (Fig. 1b). The algorithm proposed in [9], applied on the same image, spawns three different thresholds: {133, 179, 228}. The first threshold is shifted to the right thus causing the bacteria to appear more coarse-grained; furthermore the fluid structure is split up into two different objects (sprung from the two thresholds $t_2 = 179$ and $t_3 = 198$) that actually do not seem to match with patterns really enclosed in the original image.

In Fig. 4 three different microscope images together with their corresponding gray-level histograms are shown. The number of obtained thresholds, their position and the NU value associated to the segmented images are show in Table 1.

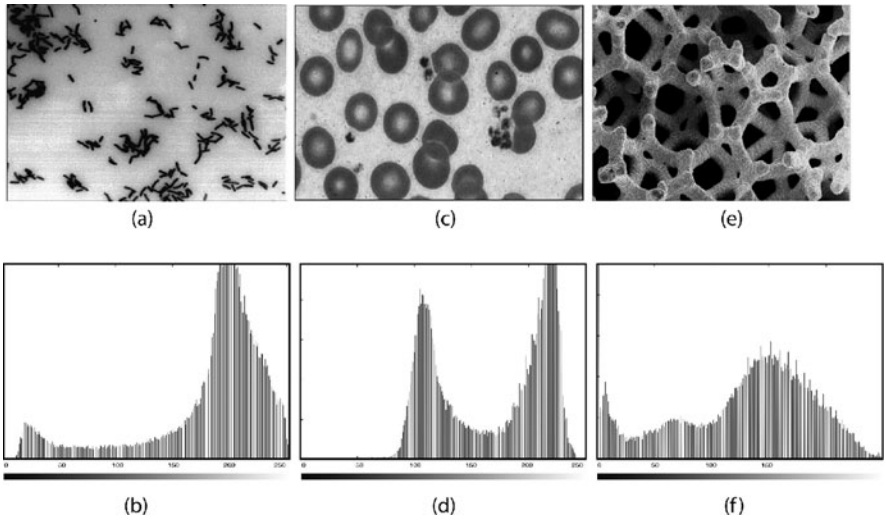


Fig. 4 (a) *Legionella* image; (b) Histogram of image (a); (c) *Red Blood Cells* image; (d) Histogram of image (c); (e) *Foam* image; (d) Histogram of image (e)

Table 1 Threshold values $\underline{t}$ and NU values for images in Fig 4. Method-1 refers to the proposed algorithm ($\varepsilon = 0.05$) while Method-2 refers to Sezgin and Taşaltín algorithm

Image	Method-1		Method-2	
	$\underline{t}$	NU	$\underline{t}$	NU
Legionella image	{59, 92, 127, 211}	**0.0856**	{59, 83, 150}	0.1436
Red Blood Cells image	{160, 193, 208}	**0.0635**	{81, 165}	0.1138
Foam image	{60}	0.2991	{38, 55, 81, 118, 160}	**0.0439**

As can be noticed, in all the cases the two algorithms differ in the number of discovered thresholds. *Foam* image is the only one in which the proposed approach underperforms, according to the NU measure, Sezgin and Taşaltín method. For this image our approach identifies only one threshold which allows to separate the foam structure from the background (segmented images are not shown) whereas Sezgin and Tasaltín algorithm splits the structure into several regions that actually seem to reflect more illumination conditions than different morphological elements.

5 Concluding Remarks

Image thresholding is a low-level image analysis task that can be considered as the bottle-neck of the development of image processing technology because all the subsequent tasks rely heavily on the quality of the image segmentation process. It is quite a simple task in well-defined images where, usually, the histogram has a deep valley between two peaks around which the object and background's gray levels are condensed but in case of multilevel thresholding, methodological and computational hindrances stand out.

The algorithm presented in this paper tries to overcome all these drawbacks and appears as a multipurpose technique suitable to segment histogram-like structures by looking at the compactness of the *signal* around characteristic values.

However, in order to provide a methodological validation of the algorithm, the analytical properties of the stopping rules have to be investigated. Actually this rule exploits a sensitivity criterion ε that depends on the subjective choice of the user. Although simulation results have shown that the number of classes do not change significantly by varying ε in the range $[0.05 - 0.15]$ (data not shown), future work will be devoted to modify the stopping rule by dynamically tune the sensitivity criterion according to the characteristics of the specific image being processed.

References

1. Gonzalez, R.C., Wood, R.E.: Digital Image Processing, 3rd edn. Prentice Hall, Upper Saddle River, NJ (2008)
2. Huang, L.K., Wang, M.J.J.: Image thresholding by minimizing the measures of fuzziness. Pattern Recognit. **28**(1), 41–51 (1995)

3. Huang, L.K., Wang, M.J.J.: Image thresholding by minimizing the measures of fuzziness. Pattern Recognit. **28**(1), 41–51 (2000)
4. Levine, M.D., Nazif, A.M.: Dynamic measurement of computer generated image segmentations. IEEE Trans. Pattern Anal. Mach. Intell. **PAMI-7**, 155–164 (1985)
5. Murthy, C.A., Pal, S.K.: Fuzzy thresholding: mathematical framework, bound functions and weighted moving average technique. Pattern Recognit. Lett. **11**(3), 197–206 (1990)
6. Pal, S.K., Dutta Majumder, D.K.: Fuzzy Mathematical Approach to Pattern Recognition. Wiley, New York, NY (1986)
7. Pathegama, M.P., Göl, Ö.: An artificial neural process to create continuous object boundaries in medical image analysis. Int. Sci. J. Comput. **8**(1) (2004)
8. Sezgin, M., Sankur, B.: Survey over image thresholding techniques and quantitative performance evaluation. J. Electronic Imaging **13**(1), 146–165 (2004)
9. Sezgin, M., Taşaltín, R.: A new dichotomization technique to multilevel thresholding devoted to inspection application. Pattern Recognit. Lett. **21**, 151–161 (2000)
10. Yen, J.C., Chang, F.J., Chang, S.: A new criterion for automatic multilevel thresholding. IEEE Trans. Image Process. **4**(3), 370–378 (1995)
11. Zadeh, L.: Probability measures of fuzzy events. J. Math. Anal. Appl. **23**, 421–427 (1968)

Some Developments in Forward Search Clustering

Daniela G. Calò

Abstract The Forward Search (FS) represents a useful tool for clustering data that include outlying observations, because it provides a robust clustering method in conjunction with graphical tools for outlier identification. In this paper, we show that recasting FS clustering in the framework of normal mixture models can introduce some improvements: the problem of choosing a metric for clustering is avoided; membership degree is assessed by posterior probability; a testing procedure for outlier detection can be devised.

1 Introduction

The Forward Search (FS) is a method of revealing the structure of multivariate data. It provides a robust approach to clustering and allows to detect observations that do not belong to any cluster.

Given a set of n independent d-dimensional observations $\{x_1, \ldots, x_n\}$, the FS clustering procedure described in Chap. 7 of [2] consists of three main phases: (1) a preliminary analysis, providing information on cluster existence and definition, which ends with the specification of a number, say K, of "tentative" clusters: $T_k, k = 1, \ldots, K$; (2) an exploratory analysis of the tentative clusters, in order to ensure that most of their units are correctly classified; (3) a confirmatory analysis, where units not yet classified are allocated by running a Forward Search with K populations.

Atkinson et al. [3] have suggested to augment the forward plots of the monitoring statistic with bootstrap envelopes, to be used in both *phase 1* (in order to provide reliable inference on cluster identification) and *phase 3* (in order to confirm each of the clusters finally found). In addition, Atkinson and Riani [1] have improved *phase 1* by using many searches from randomly selected starting points; when monitored through the Random Start Forward Plot (RSFP), these searches provide information about the number of clusters and cluster membership. The present paper focuses on *phase 3*, with the aim of proposing possible improvements. In its original

D.G. Calò (✉)
Department of Statistics, University of Bologna, 40126 Bologna, Italy
e-mail: danielagiovanna.calo@unibo.it

B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-13312-1_13,

formulation given in [2], *phase 3* starts by robustly selecting a subset C_k from each tentative cluster ($C_k \subset T_k$, $k = 1, \ldots, K$) and progresses as summarized in the following steps:

- Set the starting basic subset, B_m (where m denotes its size) :

$$B_m \leftarrow \bigcup_k C_k$$

- While $m < n$
  - Estimate the Mahalanobis distance of x_j from each cluster centroid (for $j = 1, \ldots, n$):

$$D_{j,m,k} = [(x_j - \bar{x}_{k,m})^T S_{k,m}^{-1} (x_j - \bar{x}_{k,m})]^{1/2}$$

    where $\bar{x}_{k,m}$ and $S_{k,m}$ respectively denote the mean vector and (unbiased) covariance matrix computed on the set of units in B_m assigned to cluster k
  - Allocate units to the nearest cluster, *i.e.* according to:

$$\arg\min_k (D_{j,m,k})$$

  - Order all n observations by increasing "overall outlyingness", measured by:

$$D_{j,m} = \begin{cases} D_{j,m,k} & \text{if } x_j \in T_k \\ \min_k (D_{j,m,k}) & \text{if } x_j \text{ is unassigned} \end{cases} \quad (1)$$

  - Update the basic subset as the set of the first $m + 1$ units in the ranking[1]:

$$B_{m+1} = \{x_{[1],m}, x_{[2],m}, \ldots, x_{[m+1],m}\};\ m \leftarrow m + 1$$

In the above sketched procedure, two issues could deserve further attention and probably have some room for improvement.

Firstly, at each step of the search, units are allocated to clusters by comparing Mahalanobis distances from the groups, *i.e.* according to the elliptical K-means allocation criterion. When some clusters are more dispersed than others, it may happen that loose clusters overwhelm the tighter ones. Atkinson et al. [2] suggest to use distances standardized by the determinant of the estimated covariance matrix. Alternatively, Authors working in elliptical K-means clustering propose to circumvent the problem by adopting the so-called normalized Mahalanobis distance instead of the standard one (e.g. [11]). However, different distances can lead to searches with different behavior.

Moreover, as far as unassigned units are concerned, a "soft" assignment could be preferable in case of partially overlapping clusters, where it may be impossible

[1] The user can specify several options about the way the basic subset should grow (see [2], p. 371)

to assign these units unambiguously to a cluster. For this reason, Atkinson et al. [2] suggest to monitor cluster membership of unassigned units, so as to gain some insights on the uncertainty associated with their classification. In particular, they propose to assess the membership degree of an observation to a cluster by the proportion of steps in which that observation is allocated to that cluster. A simpler alternative to this heuristic solution would be to describe membership degrees by the posterior probabilities of belonging to the parental populations of the clusters.

Both these issues could be tackled by rephrasing the FS clustering procedure in the framework of model-based clustering. We explore this possibility in the case of a finite mixture of normal distributions, as illustrated in the following.

2 A Model-Based Formulation of FS Clustering

The Forward Search with K populations presented in [2] can be naturally recast in the framework of model-based clustering.

In fact, it is designed so that K Forward Searches are carried out simultaneously on the whole data, each one inside a distinct cluster. Hence, it relies on the underlying assumption that each cluster comes from a normally distributed population (or, more generally, an elliptically distributed one). It is worth noting that the same assumption is also involved when, following [1], envelopes derived for multivariate normal distribution are used to finally confirm the clusters found.

If the normality assumption is made explicit, then a mixture model with K d-dimensional Gaussian components turns out:

$$p(x) = \sum_{k=1}^{K} \pi_k \phi(x|\mu_k, \Sigma_k), \tag{2}$$

where each component $\phi(\cdot)$ is parameterized by its mean vector μ_k and covariance matrix Σ_k, and is weighted by the corresponding mixing proportion π_k.

In this section, we will devise a mixture-based version of the procedure sketched in Sect. 1, starting from the same robust initialization of the basic subset.

We observe that the first two operations carried out iteratively by the search are nothing but the estimation and the allocation step in elliptical K-means algorithm, respectively. Since this algorithm is a "hard" assignment version of the normal mixture model (see [4]), it is natural to replace these operations with the individual M-step and E-step of the EM (Expectation-Maximization) iteration, respectively[2].

Therefore, when dealing with a mixture of K normal populations, FS iterations can be reformulated as follows:

[2] We remind that an E-step computes the estimate z_{jk} of the conditional probability that observation x_j belongs to cluster k given the current parameter estimates (for $j = 1, \ldots, n$ and $k = 1, \ldots, K$), and an M-step computes parameter estimates given the conditional probabilities.

- While $m < n$
  - (M-step) Compute $\hat{\pi}_{k,m}$, $\hat{\mu}_{k,m}$, $\hat{\Sigma}_{k,m}$, given the current membership degrees
  - (E-step) "Soft" assignment, given the current parameter estimates:

$$z = \frac{\hat{\pi}_{k,m}\phi(x|\hat{\mu}_{k,m}, \hat{\Sigma}_{k,m})}{\sum_{k=1}^{K}\hat{\pi}_{k,m}\phi(x|\hat{\mu}_{k,m}, \hat{\Sigma}_{k,m})}$$

  - Order all n observations by increasing "overall outlyingness"
  - Update: $B_{m+1} = \{x_{[1],m}, x_{[2],m}, \ldots, x_{[m+1],m}\}$; $m \leftarrow m + 1$

The most tricky issue is that a coherent ranking criterion is needed in order to progress in the search. Any statistic measuring how typical an observation x is of a mixture can serve the purpose: in the following, two possible choices are proposed.

An immediate measure of typicality is the value that the estimated mixture density function takes in x, that is $\hat{p}(x) = \sum_{k=1}^{K}\hat{\pi}_k\phi(x|\hat{\mu}_k, \hat{\Sigma}_k)$. If we take its negative natural logarithm as a measure of outlyingness, the basic subset B_m represents (or is a close approximation of) the subset of observations minimizing $\sum_i^m -logp(x_i)$. Therefore, the estimates computed on B_m can be interpreted as *trimmed likelihood* estimates of mixture parameters, and the whole FS procedure can be viewed as a strategy for robustly fitting mixtures, alternative to the one recently proposed in [9]. In addition, plotting the values of $-log\,\hat{p}(x_{[m+1,m]})$ against m can serve for effectively monitoring the inclusion of outlying points during the search (as shown in [5]), thus helping in the choice of the trimming percentage.

In our experiments we used the typicality measure proposed by McLachlan and Basford in [8]. They assume that the mixture is fitted to $B_m = \{x_{hk}, h = 1, \ldots, m_k$ and $k = \ldots K\}$ consisting of $m = \sum_k m_k$ random observations from (2) that have been previously classified, and compare the new observation x to each of the fitted components of the mixture in turn, forming a typicality measure with respect to each component mean. This measure, say $a_k(x)$, is the p-value of the following test-statistic:

$$\frac{m_k(\nu_k + 1)}{(m_k + 1)d(\nu_k + d)}D^2(x; \hat{\mu}_k, \hat{\Sigma}_k) \tag{3}$$

where the null distribution of (3) is known to be F_{d,ν_k+1}, with $\nu_k = m_k - d - 1$ and m_k denoting the number of observations assigned to cluster k. If the p-value is sufficiently small for all the components, then x is supposed to be an outlier. Therefore, the Authors suggest to assess how typical x is of the mixture by:

$$a(x) = \begin{cases} a_k(x) & \text{if } x \in B_m \\ \max_k[a_k(x)] & \text{if } x \notin B_m \end{cases} \tag{4}$$

Note that definition (4) is in strict analogy with (1) as far as the treatment of assigned/unassigned units is concerned.

When used as a ranking criterion, this typicality measure has an appealing feature: observations assigned to the same cluster are sorted according to their Mahalanobis distance. This implies that observations belonging to a given cluster are included in the same order as if a search on that single cluster was carried out.

It is worth noting that using this measure in the search implies that at each step statistic (3) is computed on subset B_m, which can be taken as a random sample from the mixture only when m is close to the sum of the cluster sizes; generally, it is a truncated subset of such a sample. This remark invalidates its use for inferential purposes, but does not invalidate its use as a mere ranking criterion.

2.1 An Illustrative Example

The performance of the proposed strategy is illustrated on a simulated example, with two not completely separated clusters having different dispersion, and compared with that of FS clustering in [2][3]. The data set is a random sample of size $n = 300$ from the mixture given by: $p(x) = (2/3)\phi(x; (0, 0), 6I_2) + (1/3)\phi(x; (5, 5), I_2)$.

Two "tentative" clusters can be identified via the RSFP (with 100 random starts). The starting subsets, C_1 and C_2, are selected by applying the method of robustly centered ellipses to each tentative cluster, trying different amounts of trimming. The compared strategies are run starting from this common initialization of the basic subset, and the final "hard" clustering is compared with actual classification. Due to the arbitrariness in the choice of the metric, function `fwdmv` was run both with the standardized Mahalanobis distance and the usual one, and the best result (in terms of classification performance) was registered.

Table 1 Misclassification rates (in percentage) estimated on 10 independent replicates

Technique	Fraction= 0.5	Fraction= 0.7	Fraction= 0.9
FS	3.13	3.00	2.77
Mixture-based FS	1.90	1.93	1.90

Table 1 shows the percentage (averaged over 10 simulations) of observations misclassified by the original procedure (FS) and the mixture-based one (MFS) for three trimming fractions. On this simulated example, MFS achieves lower error rates that the best performance attained by the default setting of FS. Moreover, it proves to be less sensitive to the amount of trimming in the tentative clusters, which can be of particular interest form the user's perspective: in fact, the smaller the size of the starting subset the more confident one is that the initial inclusion of outliers is avoided.

[3] The latter procedure was carried out in `R` using package `Rfwdmv`; *phase 3* was performed by running the default version of function `fwdmv`. The proposed procedure has been implemented using package `mclust` (see [6] and [7]) for mixture modelling. In its present implementation there is the constraint that the value m_k in (3) cannot decrease when passing from a step to the next one (for $k = 1, \ldots, K$), which seems reasonable in the Forward Search setting.

3 An Additional Advantage of Mixture-Based Forward Search

In case of contaminated data, both the FS original procedure and its mixture-based version may lead to a biased final classification. However, the search is designed so that outlying observations remain the last ones to be included. Thus, a possible solution could be to stop the search when signalled by suitable diagnostic plots (as we briefly mentioned in Sect. 2). Besides yielding a robust classification, this can identify possibly outlying observations which may be of interest in their own right.

The mixture-based approach offers an alternative way to deal with contaminated data. In fact, as a probabilistic model is assumed, the task of outlier detection can be defined and managed rigourously, not only through exploratory techniques. Thus, provided that strong clustering information is attained during *phase 1* and *phase 2*, the idea is to finally confirm it by means of an outlier testing procedure.

The task of testing for outliers from a mixture can be accomplished by the typicality measure (4). According to [8], an observation x is assessed as being atypical if its typicality value, $a(x)$, is smaller than a specified significance level. This means that x will be labelled as outlying only if it is outlying w.r.t. all of the mixture components, coherently with the definition of "outlier" in the multiple cluster setting.

As pointed out in Sect. 2, this test requires a random sample from the mixture. The basic subsets corresponding to the peaks in the RSFP can be taken as random samples, each one from a distinct component of the mixture, provided that the peaks are clear enough. For each peak, it could be generally wise to take an "earlier" subset (that is, to stop the trajectory showing the peak a number of steps before it), so that the risk of including moderately outlying observations is reduced. This subset is obtained by elliptical truncation, since it consists of the observations having the smallest distances; hence, the estimates computed on it should be corrected according to Tallis' result (see [10]), so that consistency to the normal model is achieved.

To get some insight about the effectiveness of this outlier detection strategy, a limited simulation study was conducted, considering both clean and contaminated data. Clean data consist of two clusters ($n_1 = n_2 = 300$) from standard normal distributions located at $\mu_1 = 0\mathbf{1}_d$ and $\mu_2 = 2\chi_{d,.99}\mathbf{1}_d$, respectively, with $d = 4$. Separate contamination consists of 60 additional data from a standard normal density centered in $4\chi_{d,.99}\mathbf{1}_d$. Diffuse contamination consists of 60 additional points drawn from a normal distribution (having sample location and covariances equal to five times the sample ones) and falling outside both the $\chi_{d,.99}$ ellipses of clean data. For better understanding, an instance of the last two settings for $d = 2$ is plotted in Fig. 1.

For each setting, 500 replicates were generated. For each instance, the RSFP was built and the trajectories showing the two most relevant peaks were identified; two "earlier subsets" were selected accordingly, so that their size is 80% the size indicated by the corresponding peak; statistic (3) was computed on the pool of these subsets and performed on all observations in the sample, at the 5% and 1% level. The mean percentage of misclassified observations in reported in Table 2.

We can see that on clean data type I error (that is the percentage of null points identified as outlying by the test) is slightly above the target significance level. This

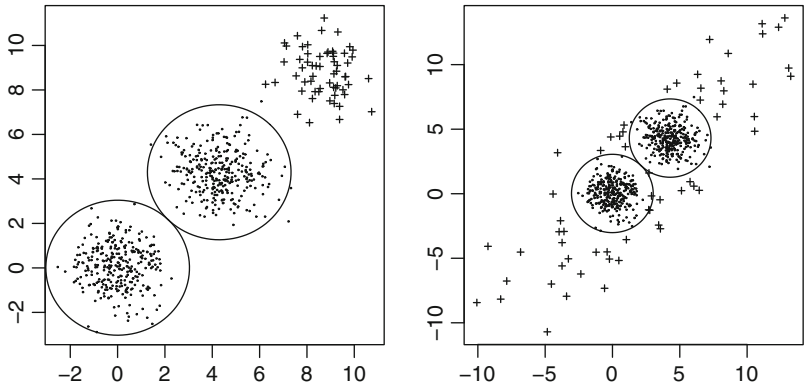


Fig. 1 A bivariate analogue of the separate and diffuse contamination situations. Planted outliers are marked by a cross. The 99 % level ellipses of the true mixture components are shown

Table 2 The average percent of type I and type II misclassification errors

Significance level	Error type	Clean data	Separate contamination	Diffuse contamination
$\alpha = 0.01$	Type I	1.11	1.17	0.92
$\alpha = 0.01$	Type II		0.07	3.48
$\alpha = 0.05$	Type I	5.38	5.34	4.65
$\alpha = 0.05$	Type II		0.04	0

is not surprising because McLachlan and Basford's test performs multiple testing, thus it is expected to have poor control over the overall significance level.

When outliers are concentrated in a cluster, nearly all true outlying points are identified as outlying, for both the levels of significance. Diffuse contamination is more challenging: given $\alpha = 0.01$, type II error is equal to 3.48%. A possible explanation is that moderate outliers may still introduce some bias: when some outliers happen to be included before the peak, the guess for the cluster size is inflated, thus Tallis' correction factor overinflates the covariance matrix estimate and this causes some outliers to be missed. If we set $\alpha = 0.05$, all outliers are correctly identified, but a larger type I error must be tolerated.

To better appreciate the method capabilities a real data experiment is finally reported. The data consist of three variables observed on 103 investment funds, and have been extensively analyzed in [2]. The proposed outlier detection procedure started by the inspection of the RSFP, which reveals two clear peaks/clusters: they were elliptically truncated so that the size of the resulting cluster is about 0.8 times the size of the original one. At the 1% level, a set of 9 observations are nominated as outliers (units 14, 21, 39, 50, 52, 54, 77, 80, 89, highlighted in Fig. 2), which includes the 6 observations (21, 50, 52, 54, 77, 89) identified by [2] using the traditional FS graphical exploratory tools.

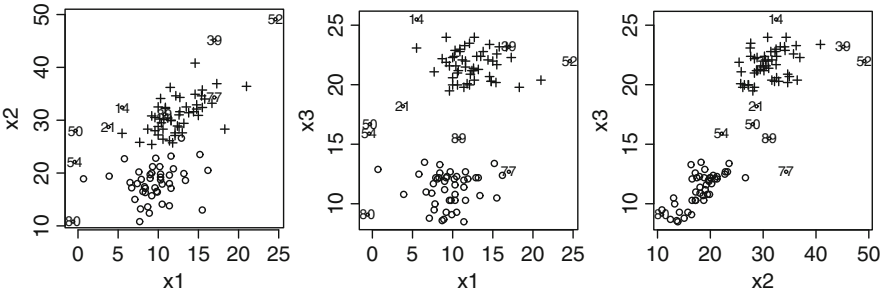


Fig. 2 Financial data: scatterplot matrix. Points declared as outliers by McLachlan and Basford's test are labelled. Clusters are denoted by dots and crosses

4 Concluding Remarks

Forward Search (FS) clustering is naturally recast in the framework of normal mixture models, thus simplifying some difficulties and augmenting the traditional exploratory tools provided by the FS. The proposals presented in this paper can be considered also according to a reversed perspective, where one wants to investigate whether normal mixture models can benefit from the FS as well. In particular, they could help in handling the problem of outliers: by using a trimming strategy, based on the mixture-based FS clustering procedure, or by an outlier testing procedure, provided that enough information is available from the Random Start Forward Plot. These conjectures are going to be investigated and are the matter of future research.

References

1. Atkinson, A.C., Riani, M.: Exploratory tools for clustering multivariate data. Comput. Stat. Data Anal. **52**, 272–285 (2007)
2. Atkinson, A.C., Riani, M., Cerioli, A.: Exploring Multivariate Data with the Forward Search. Springer, New York, NY (2004)
3. Atkinson, A.C., Riani, M., Cerioli, A.: Random start forward searches with envelopes for detecting clusters in multivariate data. In: Zani, S. et al. (eds.) Data Analysis, Classification and the Forward Search, pp. 163–171. Springer, Berlin (2006)
4. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, Heidelberg (2006)
5. Calò, D.G.: Mixture models in forward search methods for outlier detection. In: Preisach, C. et al. (eds.) Data Analysis, Machine Learning and Applications, pp. 103–110. Springer, Berlin (2008)
6. Fraley, C., Raftery, A.E.: Model-based clustering, discriminant analysis and density estimation. J. Amer. Stat. Assoc. **97**, 611–631 (2002)
7. Fraley, C., Raftery, A.E.: MCLUST version 3 for R: normal mixture modeling and model-based clustering. Technical report no. 504, Department of Statistics, University of Washington (Sept 2002)
8. McLachlan, G.J., Basford, K.E.: Mixture Models: Inference and Applications to Clustering. Marcel Dekker, New York, NY (1988)
9. Neykov, N., Filzmoser, P., Dimova R., Neytchev, P.: Robust fitting of mixtures using trimmed likelihood estimation. Comput. Stat. Data Anal. **52**, 299–308 (2007)

10. Riani, M., Atkinson, A.C., Cerioli, A.: Finding an unknown number of multivariate outliers. J. R. Stat. Soc. Ser. B. **71**, 447–466 (2009)
11. Sung, K., Poggio, T.: Example-based learning for view-based human face detection. IEEE Trans. Pattern Anal. Mach. Intell. **20**, 39–51 (1998)

Spectral Graph Theory Tools for Social Network Comparison

Domenico De Stefano

Abstract The problem faced in this paper is related to the comparison between two undirected networks on n actors. Actors are in two different configurations G_k ($k = 1, 2$). Comparison is based on the evaluation of the how the relational node distances evolve in the passage from the first net (G_1) to the second net (G_2). The procedure consists of two steps: (*i*) define an appropriate relational distance among nodes of the two networks; (*ii*) compare the corresponding distance matrices. The first step is based on the so-called Euclidean Commute-Time Distance among the n nodes computed from a random walk on the graph and Laplacian matrix. The second step concerns the comparison between the obtained distance matrices by using Multidimensional Scaling techniques. The procedure has a wide range of application, especially for experimental purposes in social network applications where this issue has not been treated systematically.

1 Introduction

Network comparison represents a widely explored topic, especially within Pattern Recognition (PR) field. In many PR applications, a crucial operation is the comparison among objects or between an object and a model to which the object could be related [6]. When we deal with graphs this comparison is made by using some kind of *graph matching* algorithm.

A natural way to deal with graph matching problems is represented by *kernel methods* [9]. Roughly speaking a kernel $k(x, x')$ is a measure of similarity between objects x and x' that satisfies the conditions of symmetry, that is $k(x, x') = k(x', x)$ and positive semi-definiteness [9]. When objects are graphs, the kernels allow to compute the dot product of a pair of graphs in a vector space without explicitly define a mapping between them.

D. De Stefano (✉)
Dipartimento di Scienze Economiche e Statistiche, Università di Salerno, 84084 Fisciano Salerno, Italy
e-mail: d.destefano@unina.it

B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-13312-1_14,

Alternatively, it is possible to use another general family of procedures known as *graph embedding methods*. Graph embedding permits to explicitly find a mapping between graphs and real vectors that allows to operate in the associated space [1], [14].

Recently, new approaches are based on merging both graph kernel based methods and graph embedding in order to extend the statistical tools for graph comparison tasks [4]. In this paper we use graph embedding procedures in order to compare social networks. A *social network* can be represented by a graph, but with respect to later it presents peculiar aspects related to the ties formation among actors. In SNA the comparison issue has not yet been systematically addressed. A first step to face it could be to work on the graph representation and to use the results above outlined for graph comparison.

In particular, we deal with network comparison problem by using a special graph embedding technique related to the spectral properties of the network. We follow the work of Bai et al. [1] where graph comparison is based on applying multidimensional scaling (MDS) to a matrix of shortest geodesic distances and then the embedding is used for graph matching. However, the use of geodesic distances among nodes does not capture the complexity of the structure typical of a social network. Indeed, generally in social network analysis (SNA, from now on) is necessary to take into account both the dependence structure and the node neighborhood characteristics. For these reason, here we embed the matrix of the Euclidean commute-time distance (ECTD) among nodes, rather than the geodesic distance matrix. Indeed, ECTD has very interesting properties related to well-known quantities used in SNA.

The paper is organized in two parts: in the first one we introduce the two steps of the comparison procedure for social networks, i.e. the development of the node relational distance and its use in exploratory network comparison; the second part concerns an example of a possible application, that shows how the procedure works on real networks.

2 Network Comparison in Social Network Analysis

In the present paper, we will answer to the following research question: given two observed configurations of a single social network is it possible to define a comparison (in a static fashion) between these states in such a way one can measure the difference between them?

Different methods are scattered in SNA literature to handle network comparison problems. Basically two approaches are used: (*i*) classic approach [3]; (*ii*) local structures censuses approach [7]. In the following, we indicate these methods, respectively, with MA and MB.

A network is a mathematical object $G = (V, E)$ composed of two sets: a set $V \equiv \{v_1, v_2, \ldots, v_n\}$ of cardinality $|V| = n$, containing nodes (actors) and a set $E \equiv \{e_1, e_2, \ldots, e_m\}$, of cardinality $|E| = m$, containing edges (ties) that are pairs of connected actors. The network configurations are totally represented by its edgeset E. A network $G = (V, E)$ can be fully described by its adjacency matrix $\mathbf{A}$, a $n \times n$

square matrix whose entries a_{ij} (with $i, j = 1, \ldots, n$) are equal to 1 when the edge e_{ij} belongs to the edgeset E, 0 otherwise.

Briefly, the MA class of methods is based on the comparison between the matrices **A** related to the two nets by using well-known statistical indices. The comparison is made by measuring the observed differences between two adjacency matrices by mean of an index α (for example the correlation coefficient or the Goodman-Kruskal Gamma) [3]. The MB approach is a procedure used to compare two or more networks that consists in comparing the distributions of the *local structures*, e.g. triads, observed in each of the k configurations [7]. In MB is supposed that two networks are similar if they share an high number of equals triad isomorphism classes. This method is developed by constructing a $t \times k$ contingency table **C** whose rows are indexed by the t local structure isomorphism classes [15] ($t = 16$ in the case of triad censuses) and columns are indexed by the k observed networks. **C** is projected in a factorial space by means of Correspondence Analysis [7].

Alternatively, we propose to use a method based on a particular relational node distance. Differently from MA and MB (that are both based on adjacency matrix analysis) we also provide some evidence that other graph related matrices could be more useful, than matrix **A**, for certain SNA exploratory issues.

3 The Graph Embedding Approach for Social Networks Comparison

In the present approach we take into account that any network comparison process may be based exclusively on the underlying relational structure embedded in the nets. It means that a meaningful network comparison approach may pass through the detection of the differences that occurs in actors interrelations. In particular we distinguish between two classes of actor interrelations: *direct interrelations*, with which we refer to the set of links in E among the actors; *indirect interrelations*, with which we mean the indirect connectivity, i.e. the sharing of a common neighborhood. Indeed, it is reasonable to suppose that network distance may depend on changes in both direct and indirect connectivity modifications. In other terms, these two kinds of connections jointly influence network topological structure. Therefore two networks presenting notable differences in these connectivity characteristics are very distant from each other.

Our first purpose is to develop a quantity that describes the actor relational position in a given configuration. We require that this quantity must capture the actors distance in terms of their direct and indirect interrelations. The procedure to build up this distance is based on spectral graph theory concepts [5]. In particular we use the laplacian matrix of the network, its pseudo-inverse and the specification of a random walk on the network. Afterward we will compare network configurations by detecting the differences in actors relational distances observed in the network states. As we will see, this distance has the nice property of being an euclidean distance among the nodes. This allows us to compare configurations by

using statistical exploratory approach for metric data. To summarize, the proposed procedure consists of two steps: *i*) definition of the node relational distance and of the distance matrices Δ_k (with $k = 1, 2$) for each compared configuration; (*ii*) comparison between the obtained Δ_k matrices.

Step 1: Definition of node relational distance matrix. In order to define the node distance matrix on a network we require that the compared networks are on the same actor set $V \equiv \{v_1, v_2, \ldots, v_n\}$ of cardinality n. At least we require that exists a one-to-one correspondence between the two graphs nodesets. Furthermore, for the computation of the node relational distance, we need to introduce: (*i*) the laplacian matrix $\mathbf{L}$; (*ii*) its pseudo-inverse $\mathbf{L}^+$; (*iii*) a random walk on the network (RW); (*iv*) transition related quantities.

The *laplacian matrix* $\mathbf{L}$ is formally the discrete analogue of the *Laplace operator*. In the spectral graph theory literature, there are several definitions of $\mathbf{L}$ [5]. We will use the non normalized version of $\mathbf{L}$: $\mathbf{L} = \mathbf{D} - \mathbf{A}$, where $\mathbf{D}$ is the diagonal node-degree matrix and $\mathbf{A}$ is the adjacency matrix of the network G. The laplacian matrix, its pseudo-inverse $\mathbf{L}^+$ and their spectra have several important properties related to the network connectivity [5]. In particular the matrix $\mathbf{L}^+$ is a valid kernel [13]. Here we are interested in the properties connected to the RW transition related quantities. A RW on a graph is a first-order reversible Markov chain (MC) on the nodes [10]. The transition probability to pass from a starting node i to an adjacent node j is: $Pr_{i,j} = Pr[s(t+1) = j | s(t) = i] = \frac{a_{ij}}{\sum_{j=1}^{n} a_{\cdot j}} = \frac{a_{ij}}{d_i}$, where a_{ij} is the (i, j)-th entry of $\mathbf{A}$ of G_k, d_i is the degree of i-th actor, $s(t) = i$ and $s(t+1) = j$ are the states of the MC at the time t and $t + 1$, respectively. The transition probability matrix is then defined as: $\mathbf{P} = \mathbf{D}^{-1}\mathbf{A}$. Briefly, the MC evolution is characterized by: $x(0) = x_0$; $x(t + 1) = \mathbf{P}^t x(t)$.

For a random walker, on a RW, it is possible to define two *transition related quantities*. These are directly connected to the elements of $\mathbf{L}^+$. The first one is the *average first passage time*, $m(j|i)$ which represents the mean number of steps, a random walker needs, to reach for the first time the node j starting from the node i. This quantity measures the minimum time until hitting state j is reached, as $T_{ij} = min(t \geq 0 | s(t) = j, s(0) = i)$. It is possible to prove that this quantity is directly obtained by the (i, j)-th element l_{ij}^+ of $\mathbf{L}^+$ [8]: $m(j|i) = \sum_{h=1}^{n} l_{ih}^+ - l_{ij}^+ - l_{jh}^+ + l_{jj}^+$. Where h indicates the intermediate nodes between i and j.

The second related quantity is the *average commute time* ($\overline{CT}$), which is a transition related quantity defined from $m(j|i)$: $\overline{CT}(i, j) = m(j|i) + m(i|j)$. $\overline{CT}$ represents the mean number of steps, a random walker needs, to reach for the first time the node j starting from i and coming back again to the node i. It is easy to show that also this quantity is computed by the $\mathbf{L}^+$ elements [13]: $\overline{CT}(i, j) = V_G(l_{ii}^+ + l_{jj}^+ - 2l_{ij}^+)$, where V_G is the volume of the graph (i.e. the sum of all the node degrees). It can be showed that this quadratic form is a distance [11]. In particular the CT distance (not in average) between two nodes is expressed as: $CT(i, j) = l_{ii}^+ + l_{jj}^+ - 2l_{ij}^+$.

The principal property of the CT is that its square root is an Euclidean distance in the space $\Re^n$ of the nodes called *Euclidean commute time distance* (ECTD) [13]. Briefly, ECTD is a Mahalanobis distance with weight matrix $\mathbf{L}^+$. The useful property of $ECTD(i, j)$ is that it decreases when the number of walks between i and j

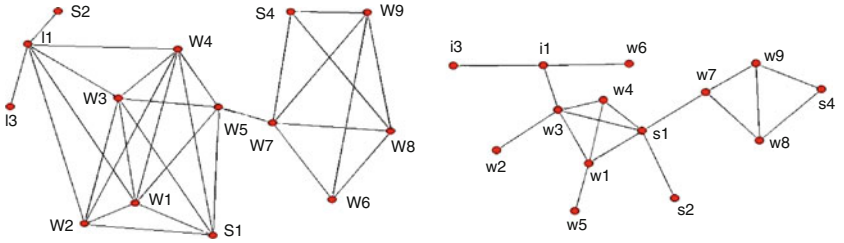
increases and/or the length of these walks decreases. In short, if between two actors there is a low ECTD they can be considered close to each other, given that they have many short paths connecting them. From a SNA viewpoint, ECTD captures both the closeness and the degree centrality of actors [15]. The first step of our proposed method ends with the computation of the two ECTD distance matrices Δ_k for the two networks G_k.

Step 2: Definition of node relational distance matrix. The second step of the procedure consists in the comparison between the ECTD distance matrices. In this phase we have to capture the change in CT, or equivalently in ECTD, registered for the i_k-th node in the passage through the k network configurations. The change in CT between i and j does not necessarily corresponds to the appearance/disappearance of a link between them (i.e it is not only expressed by their direct interrelations). However, it could equally mean that their neighborhood has now changed (i.e. a modification of their indirect interrelations is occurred). In particular, $CT(i, j)$ decreases if i and j share more common neighbors, viceversa it increases if the number of common neighbors decreases. We distinguish two ways of comparing the networks G_k equipped with ECTD: (*i*) adopting an exploratory comparison tools in order to measure and/or to visualize the elementwise distance registered through the two configurations; (*ii*) implementing an analytical comparison based on a synthetic index representing the difference between the Δ_k (e.g. by using the euclidean matrix norm). We focus on the first comparison approach because the second can be easily obtained by the $n \times n$ distance matrix whenever an appropriate matrix distance has been selected. Our exploratory comparison solution allows us to visualize the occurred networks differences by applying a metric MDS in combination with a non-orthogonal procrustes transformation[1] [2]. This approach seems to be the most appropriate in social networks comparison because it allows a node by node comparison, which is desirable because a single distance measure does not capture network complexity. Moreover, the procedure works well especially when we are able to identify the single nodes, either because of the small network size either because of availability of labels or actors attributes. This is not a lack of generality because in SNA, the network size is relatively small and the actors are univocally labelled [15]. Briefly with MDS, it is possible to identify the relational differences emerged in the passage through the configurations.

4 A Simulated Applicative Example

In order to show how the procedure works we run an example on simulated data. In this section we will explain how the presented method has a wide range of application especially when additional nodes information are available. We simulate data

[1] Non-orthogonal procrustes transformation operates on a selected configuration object and it allows to match it to another configuration object as closely as possible, i.e. without lose their metric properties.



(a) Network G_1 (b) Network G_2

Fig. 1 **(a)** Initial network configuration G_1; **(b)** successive network configuration G_2

concerning the relational behavior of a set of 14 firm employers[2] [12]. Let suppose we know, for each actor, the role covered in the firm, in other words we suppose that additional node attributes are available. In particular we have: 3 supervisors (S1, S2, S4); 2 instructors (I1, I2); 9 workers (W1, W2,..., W9). We indicate with capital letters (i.e. I1, ..., S1, ..., W9) the actors observed in G_1 and with small letters (i.e. i1, ..., s1, ..., w9) the ones observed in G_2. The purpose is to compare the cooperation/relationship status among them over two distinct time points, i.e. over the passage through the two distinct network configurations G_1 and G_2 (Fig. 1). The first step consists in obtaining the corresponding ECTD matrices Δ_1 and Δ_2 for G_1 and G_2 by means of the computations in Sect. 3. Once we have the Δ_k we can carry out an MDS for each of them (see Fig. 2). As it is clear from Fig. 2 and from the networks (Fig. 1), the most central actors are closer to the origin because their mean relative ECTD (with respect to all the other nodes) is small. The comparison between the two configurations is based on the loss and on the gain in centrality of nodes (in terms of ECTD). Indeed, for example actor W5 (Fig. 2, part (a)) is the most central actor in the first configuration whereas in the second network its position becomes a little more peripheral (Fig. 2, part (b)). The use of a non-orthogonal procrustes transformation [3] assures the best superimposition of the two partial networks MDS (fig. 3).

Here we can visually appreciate the changes in relational node distances. For example, let suppose that we are interested in the instructors and supervisors relational changes in the two configurations. We can detect the relative positions, with respect to all the other nodes, in a given configuration. For example, instructor 3 is more central in the configuration G_2 than in G_1. This means that I3 relative distance (instructor 3 in G_1) is comparatively larger than the one observed for the same actor (i3) in G_2. Similarly, for actor S2 (supervisor 2 in G_1), its ECTD sharply decreases.

[2] We adopt the population size of 14 units and some particular actor labels to recall the Elton Mayo's experiments to which this example is inspired.

[3] Here we choose this particular kind of transformation for simplicity but it is possible also to choose different procrustes transforms. For example the orthogonal procrustes transform allows to select whether or not a translation and a scaling are allowed in the transform.

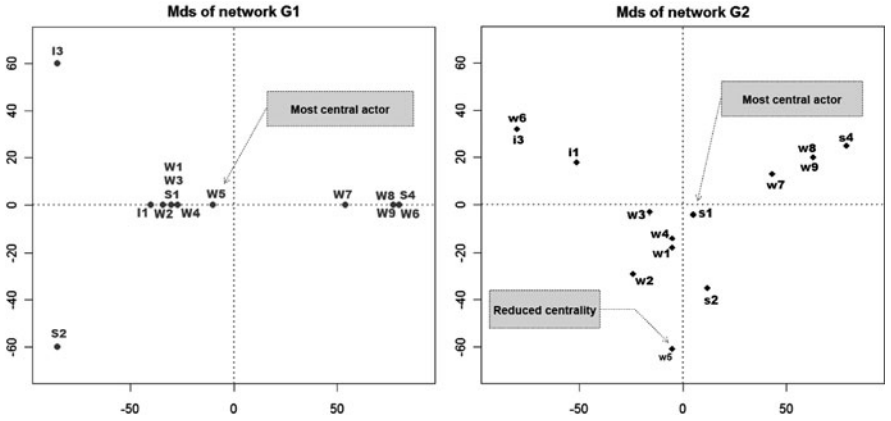


Fig. 2 **(a)** MDS on the Δ_1 corresponding to the initial network configuration G_1; **(b)** MDS on the Δ_2 corresponding to the configuration G_2. On the axis there are the ECTD distances. In the text-boxes some examples of data interpretation

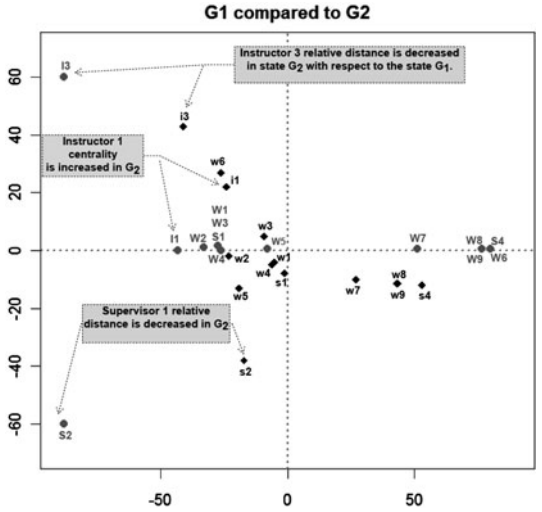


Fig. 3 Representation of G_1 and G_2 in the common 2-dimensional Euclidean space obtained by Procrustean transformation. With circles we indicate actors in G_1 (capital letters), with rhombus actors in G_2 (small letters). On the axis there are the ECTD distances. In the text-boxes some examples of data interpretation

Indeed, its position in second configuration G_2 (s2) is closer to the other points. It is possible to be more precise by measuring ECTD between nodes by means of their coordinates. Since ECTD expresses the centrality and the connectivity of the actors we can interpret it in a relational fashion and use it for comparison purposes.

5 Concluding Remarks

In the present paper we showed how the use of spectral graph theory quantities can be used to represent distance in a social network, and how this distance may have a nice relational interpretation. The procedure is mainly based on the consideration that distance between networks should always be interpreted in a relational sense and in an elementwise fashion, taking into account both direct and indirect nodes interrelations. Indeed, networks are complex relational structures whose differences lie in the relative node connections changes. Therefore, the principal advantage of this approach, based on pattern recognition methods, with respect the usual social network analysis matching procedures (MA and MB) consists in the fact that we can operate elementwise comparisons rather than obtain global measures. For example, MA simply returns an index, which is very useful for quick comparisons but is clearly not suitable for more complex issues. Also in MB approaches we consider whole networks just as single points in a factorial space. The presented procedure allows to compare, principally in a visual fashion (but also analytically, by means of nodes coordinates), the actors relational distance (in ECTD) changes in order to have information on the whole network modifications occurred in the passage from G_1 to G_2. However, also in the presented approach a global matrix index can be used in order to have a synthetic comparison. It could be use *canonical matrix norm* between Δ_1, Δ_2, though just one quantity could be not enough for a complex task.

References

1. Bai, X., Yu, H., Hancock, E.R.: Graph matching using spectral embedding and alignment. In: Proceeding of the 17th International Conference on Pattern Recognition, Cambridge, pp. 398–401 (2004)
2. Borg, I., Groenen, P.J.F.: Modern Multidimensional Scaling. Springer, New York, NY (2005)
3. Borgatti, S.P., Everett, M.G., Freeman, L.C.: UCINET 6 for Windows: Software for Statistical Analysis of Social Networks. Analytic Technologies, Harvard, CA (2002)
4. Bunke, H., Irniger, C., Neuhaus, M.: Graph matching: challenges and potential solutions. In: Roli, F., Vitulano, S. (eds.) ICIAP 2005, Lecture Notes in Computer Science, vol. 3617, pp. 1–10. Springer, Heidelberg (2005)
5. Chung, F.: Spectral Graph Theory. AMS, New York, NY (1997)
6. Conte, D., Foggia, P., Sansone, C., Vento, M.: Thirty years of graph matching in pattern recognition. Int. J. Pattern Recognit. Artif. Intell. **18**, 265–298 (2004)
7. Faust, K.: Comparing social networks: size, density and local structures. Methodoloski Zvezki **3**, 185–216 (2006)
8. Gobel, F., Jagers, A.: Random walks on graph. Stoch. Process. Appl. **2**, 311–336 (1974)
9. Kashima, H., Tsuda, K., Inokuchi, A.: Kernels for graphs. In: Scholkopf, B., Tsuda, K., Vert, J.P. (eds.) Kernel Methods in Computational Biology, Bradford Books, pp. 155–170. The MIT Press, Cambridge, MA (2004)
10. Kemeny, J.G., Snell, J.L.: Finite Markov Chains. Springer, New York, NY (1976)
11. Klein, D.J., Randic, M.: Resistance distance. J. Math. Chem. **12**, 81–95 (1974)
12. Mayo, E.: The Social Problems of an Industrial Civilization. Routledge & Kegan Paul, London (1949)

13. Saerens, M., Fouss, F.: A novel way of computing similarities between nodes of a graph, with application. Technical report IAG WP 06/08, Université Catholique de Louvain (2006)
14. Valveny, E., Ferrer, M.: Application of Graph Embedding to Solve Graph Matching Problems. In: CIFED 2008, France (2008)
15. Wasserman, S., Faust, K.: Social Network Analysis. Cambridge University Press, Cambridge (1994)

Improving the MHIST-p Algorithm for Multivariate Histograms of Continuous Data

Mauro Iacono and Antonio Irpino

Abstract In many different applications (ranging from OLAP databases to query optimization) having an approximate distribution of values in a data set is an important improvement that allows a relevant saving of time or resources during computations. Histograms are a good solution, offering a good balance between computation cost and accuracy. Multidimensional data require more complicated handling in order to keep these two requirements within significant usefulness. In this paper we propose an improvement of the MHIST-p algorithm for the generation of multidimensional histograms and compare it with other approaches from literature.

1 Introduction

Important areas of scientific, civil and commercial applications are characterized by considerable amounts of relevant data, scattered in a known defined range. These data are usually specially relevant if considered inside (arbitrary) intervals, and the most of the information represented by these data can be obtained by manipulating an aggregate description of them. This is the case of results of measurements, such as in scientific applications, that are usually raw values affected by error, or geographical coordinates, such as in georeferenced information, which locate a single spot but usually tag other information that is relevant for the surrounding area. These data tend to assume almost as many different values as the number of the available observations, but cannot be discretized a priori into intervals because of the different possible applications and the intrinsic value of the information. Nevertheless, given the application, the availability of synthesis tools that enable a fast approximate knowledge of data with low computational demands is highly advisable. This is the case of many applications that span from advanced database manipulations to low-level database query optimization [7, 3].

M. Iacono (✉)
DEM, Seconda Universitá degli Studi di Napoli Belvedere Reale di San Leucio, Caserta, Italy
e-mail: mauro.iacono@unina2.it

B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-13312-1_15,

2 Motivation

The need for efficient and approximate synthetic representation of data can be easily motivated. A first "classic" example is given by OLAP (On-Line Analytical Processing) database applications. In this field data are not important in their single atomic informational content but as a whole, and the goal is to be able to manipulate synthetic (approximated) information by considering aggregations over the different aspects of the facts represented by data. OLAP applications do not focus on punctual data but on trends over (multidimensional) aggregations obtained by fast tools for slicing and dicing data according to the needs, so an approximate but manageable description is perfect for the task.

A fundamental field of application is completely different but crucial in relational databases. User queries are analyzed by the query optimizer of the DBMS and an execution plan is obtained by the analysis that minimizes the use of resources in intermediate tables during JOIN operations. In order to compute the approximate dimensions of intermediate tables and to choose between alternatives, a compact and computationally inexpensive approximate description of data distributions is needed.

In all these cases, the complete description of data is complex, while the punctual description is sparse. An intermediate description suitable for our needs must satisfy three conditions:

- compactness;
- low space cost and complexity for the representation;
- incremental update.

A fourth, desirable condition is the possibility of handling multi-dimensional data. An appropriate tool is given by histograms.

3 Histograms and Related Works

In [6] we can find the state of the art of the histogram generation procedures from large databases. One of the main criticism, that in our opinion biased the research, is about the two different definitions of histograms reported by the author. Trying to identify a historical birthday for the term and use of histograms in statistics, the author correctly reports the definition of Karl Pearson which created the term "histogram" to refer to a "common form of graphical representation". He then goes on:

> In the Oxford English Dictionary quotes from "Philosophical Transactions of the Royal Society of London" Series A, Vol. CLXXXVI, (1895), p. 399, it is mentioned that "[The word 'histogram' was] introduced by the writer in his lectures on statistics as a term for a common form of graphical representation, i.e., by columns marking as area the frequency corresponding to the range of their base."

In this definition the proportionality relationship between the frequency and the area of the rectangles (columns) is clear. Further in the text, the common error of

identifying a histogram as a particular case of a (vertical) "bar chart" clearly appears (the same error is also present in Microsoft Excel®). The error is related to the definition also proposed in [6], where the author assumes as common definition the following:

> Even today this point [the histogram are conceived as a visual aid to statistical approximation] is still emphasized in the common conception of histogram: Webster's defines a histogram as "a bar graph of a frequency distribution in which the widths of the bars are proportional to the classes into which the variable has been divided and the heights of bars are proportional to the class frequencies".

In statistics, this is true only if the widths (or the classes) of the bars are all equal!

We assume the definition of Karl Pearson as correct, considering that a histogram is a way of representing a density distribution.

In many interesting relevant cases (e. g. databases recording raw instrumental measures from sensors) two hypotheses are valid:

- a) data assume continuous values into a specified range, generally distributed over all possible samples: we will refer to this hypothesis as the *continuous descriptors* hypothesis;
- b) the scale of different measures is quite different: we will refer to this hypothesis as the *non-homogeneity* hypothesis.

Among several algorithms proposed in the literature [7, 10, 2, 8], in order to obtain improved performance in the relevant cases we propose to modify the MHIST-p [2] algorithm for building a multivariate histogram. Good histograms partition data sets into buckets with close-to-uniform internal tuple density. Unfortunately, current multidimensional histogram techniques do not always produce close-to-uniform partitions of the data sets, as we discuss in the following. A partition of a multidimensional data domain results in a set of disjoint rectangular buckets that covers all the points in the domain and assigns to each bucket some aggregated information, usually the number of tuples enclosed. The choice of rectangular buckets is justified by two main reasons: first, rectangular buckets make it easy and efficient to intersect each bucket and a given range query to estimate selectivities. Second, rectangular buckets can be represented concisely, which allows a large number of buckets to be stored using the given budget constraints.

In [7] a taxonomy of partitioning schemes for building multidimensional histograms is presented, which is illustrated in Fig. 1. In the grid partitioning scheme (Figure 1(a)), each dimension d_i is divided into p_i disjoint intervals, obtaining a grid

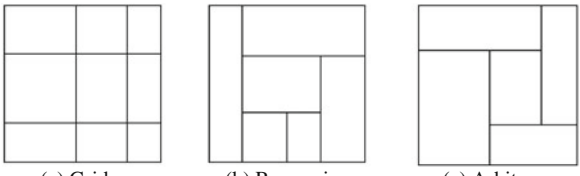


Fig. 1 Three strategies for bucketing data

of Q in p_i buckets. A recursive partition (Fig. 1b) starts with one bucket covering the whole domain, and repeatedly divides some existing bucket in two parts along some dimension. Finally, the arbitrary partition scheme (Fig. 1c) imposes no restrictions on the arrangement of buckets. In principle, all schemes are equivalent in the sense that we can simulate any partition that follows one scheme with the others (possibly using more buckets).

4 Improving MHIST-p

We improve the MHIST algorithm in order to overcome two main drawbacks deriving from the nature of the data to be summarized. In the general MHIST-p algorithm, data are described by a data table $\mathbf{X}_{n\times p} = [x_{i,j}]$, in which each row is a record and each column is a numerical field. According to our hypotheses, the $j-th$ $(j = 1, \ldots, p)$ variable describing each record is a continuous variable, i.e. we observe n different values for each row.

The MHIST-p algorithm belongs to a family of algorithms that search the best rectangular partition of a p-dimensional description space by using non-overlapping p-dimensional rectangular regions. The general MHIST-p algorithm operates as follows:

Step 0 A maximum number (denoted with k) of p-dimensional rectangles for the partition of the data domain is chosen. In the beginning, there is only one $(s = 1)$ p-dimensional rectangle (including the whole data domain) associated with p univariate distributions (one for each dimension).

Step 1 From the set of $T_{p\times s} = [t_j]$ the $j-th$ marginal distribution that is *the most in need of partitioning* is chosen.

Step 2 Next, the t_j is split at the point of maximum difference (in absolute value) with the uniform distribution, that is at the point where the Kolmogoroff distance between the data distribution and the uniform cdf has a maximum. s is augmented of one, i.e. the initial rectangle is split in two non overlapping rectangles. The algorithm iterates from *Step 1* until $s \leq k$.

Performances of the algorithm strongly depend on the criterion used to define which distribution is *the most in need of partitioning*. For V-Optimal histograms [6] considered by the standard MHIST-p algorithm, this means a marginal distribution that has the maximum variance of the source parameter value (the observed values or the frequency associated with an observed value): but considering the *continuous descriptors* and the *non-homogeneity* hypotheses, the classic [2, 6] criteria are inadequate. Moreover, the main assumption in designing histograms is the *uniformity* hypothesis (i.e. data are uniformly distributed within each p-dimensional rectangle). To cope with this drawback we propose to use a different criterion to select and to split the variable needing to be partitioned. For each distribution of the set T, we compute the *mean square error* between the observed distribution function and the uniform distribution defined by the bounds of the domain of t_j. This criterion allows

the algorithm to be more consistent with the *uniformity*, the *continuous descriptors* and the *non-homogeneity* hypotheses. In order to demonstrate improvements, two kinds of validation procedures can be used:

Goodness of fit: comparison of the real data multivariate distribution with the distribution associated to the new multivariate histogram representation. The main drawback is related to the *continuous descriptors* hypothesis that can considerably slow down the validation process;

Selectivity estimation: one of the strategies proposed in [2, 8] for preparing a collection of a large set of validating queries (to simulate user behavior), followed by a verification of the results returned without using the multivariate histogram in terms of absolute frequencies and those obtained using the multivariate histogram.

5 An Application on Real and Artificial Datasets

To show different situations, the process will be applied on two data sets: a real data set and a synthetic data set. For the sake of simplicity, we refer to two 2D datasets, but the method is extensible to more than two dimensions. The first dataset is extracted by the Forest Covertype data [1] and considers the "Elevation" and "Horizontal distance from roadways" variables on the first 11,340 tuples, that have been used as training set for classification purposes. The second is a synthetic dataset of 40,000 tuples generated from a mixture of 40 bivariate and independent gaussian variables. We generated three histograms for each dataset respectively consisting in 20, 40 and 100 buckets. We compared our proposal with the MHIST algorithm variant that uses the Maxdiff criterion for splitting buckets, using the observed values as source parameter (see Fig. 2. In a Maxdiff histogram, there is a bucket boundary between two adjacent source parameter values if the difference between these values is the largest difference possible (for further details please refer to [4]). Both the algorithms have the same computing cost in terms of memory usage ($2 \times dimensions \times No.Buckets$ plus the frequencies of each bucket) and executed operations ($No.Buckets \times tuples$ in the worst case for step 2). Given a data set D, a histogram H, and a validation workload W, the average absolute error $E(D, H, W)$ is calculated according to [2] as the mean absolute error between predicted frequency by the model and the effective frequency of data as:

$$E(D, H, W) = \frac{1}{|W|} \sum_{q \in W} |Pre(H, q) - Obs(D, q)| \tag{1}$$

where $Pre(H, q)$ is the estimated number of tuples returned by query q, using histogram H for the estimation, and $Obs(D, q)$ is the actual number of D tuples returned by q. We choose average absolute errors as the accuracy metric, since relative errors tend to be less robust when the actual number of tuples for some queries is zero or near zero. In general, however, absolute errors greatly vary across data

sets, making it difficult to report results for different data sets. Therefore, for each experiment, we normalize the average absolute error by dividing it by the estimated result size obtained by assuming a single bucket over all the data domain, i.e. in the case where no histogram is available. We refer to the resulting metric as *Normalized Absolute Error*. We compute the *Normalized Absolute Error* in order to validate the technique by using a goodness of fit approach, comparing the histograms with respect to the data distribution, and by using a selectivity estimation strategy. Effectiveness of the description can be evaluated by properly simulating the behavior of users. We focus on the performance evaluation of this technique in the case of range queries. According to [8], users can be modeled with two general behaviors while accessing data: access by windows with constant dimensions (while exploring data, with no idea of their nature) or by windows with a constant number of elements in it (in case of expertise about the nature of data, examining a quantity of the information that is considered manageable). Accesses happen according to a strategy for choosing the center of the window: distributed as a uniform (or gaussian) in a first approach, depending on the hypotheses about data, or distributed according to the data distribution, in the case in which knowledge about data is available and every element of the distribution is equally important for the user. In order to evaluate the selectivity estimation, in analogy with [2] we must select:

- proper data sets on which the validation has to be performed;
- a type of query on which the validation can be based;
- a distribution for the generation of the queries.

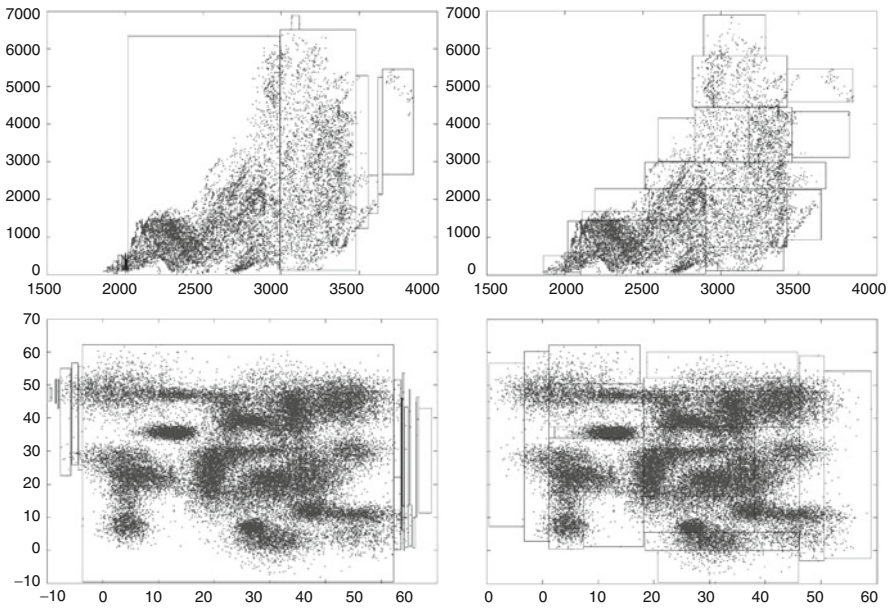


Fig. 2 The Cover type 2D (upper) and the Multigauss (lower) dataset bucketed into 20 2-dimensional classes using Mhist with Maxdiff (left) and Kolmogoroff (right)

Table 1 Normalized absolute errors in percentage

		No.Buckets = 20		No.Buckets = 40		No.Buckets = 100	
Dataset	Validation	Mhist-MD	Mhist-K	Mhist-MD	Mhist-K	Mhist-MD	Mhist-K
Forest 2d	GOF	80.3%	2.2%	66.7%	1.8%	45.8%	0.6%
	Uniform V[1%]	5.5%	0.8%	4.4%	0.5%	3.3%	0.3%
	Uniform T[1%]	13.9%	0.8%	9.6%	0.5%	5.3%	0.3%
	Gauss V[1%]	6.4%	1.2%	5.9%	0.8%	5.2%	0.4%
	Gauss T[1%]	6.7%	1.0%	7.3%	0.7%	7.3%	0.3%
	Data V[1%]	10.9%	1.8%	9.7%	1.4%	7.9%	0.7%
	Data T[1%]	2.9%	0.9%	2.3%	0.7%	2.0%	0.4%
40 Multigauss	GOF	84.4%	10.1%	66.2%	8.2%	21.0%	2.1%
	Uniform V[1%]	11.2%	5.7%	10.3%	3.3%	7.2%	1.6%
	Uniform T[1%]	12.0%	5.3%	11.6%	3.3%	7.7%	1.6%
	Gauss V[1%]	13.6%	6.4%	11.7%	3.4%	8.3%	1.7%
	Gauss T[1%]	9.4%	4.8%	8.8%	2.9%	6.8%	1.5%
	Data V[1%]	17.9%	8.6%	15.1%	4.3%	10.9%	2.0%
	Data T[1%]	6.9%	4.2%	6.1%	2.6%	4.9%	1.3%

We start with a training workload that consists of 10,000 square queries, the centers of which follow the three cited distributions:

Uniform: The query centers are uniformly distributed in the data domain.
Gauss: The query centers follow a Gauss distribution independent of the data distribution.
Data: The query centers follow the data distribution.

The range constraints we used for our experiments are:

V[1%]: The range queries are squares the area of which is 1% of the area defined by the data domain, to model the case in which the user specifies the query values in terms of a window area.
T[1%]: The range queries are squares covering a region containing 1% of total tuples, to model the situation in which the user has knowledge about the data distribution and issues queries with the intention of retrieving a given number of tuples.

The Normalized Absolute Error results are reported in Table 1 with histograms generated on data consisting of 20, 40 and 100 buckets.

6 Conclusions and Perspectives

In this paper we presented an improvement for the MHIST algorithm that considerably improves the quality of the histograms in case of very scattered data. We demonstrated by simulation that our splitting criterion gives a better description of data with the same amount of buckets. We are convinced that techniques for the extraction of multivariate histograms will get more and more important because of the exponential growth of databases and the increasing use of data stream analysis techniques and tools, since of their inherent capability of preserving a good approximation of statistical dependence of multivariate data. Future works include further investigation on possible improvements and the application of the technique.

References

1. Blake, C. Merz, C.: UCI repository of machine learning databases (1998)
2. Bruno, N., Chaudhuri, S., Gravano, L.: STHoles: a multidimensional workload-aware histogram. Technical Report MSR-TR-2001-36, Microsoft Research (2001)
3. Deshpande, A., Garofalakis, M., Rastogi, R.: Independence is good: dependency-based histogram synopses for high-dimensional data. In: Proceeding of the 2001 ACM SIGMOD International Conference on Management of Data, pp. 199–210, 21–24 May 2001, Santa Barbara, CA (2001)
4. Dumouchel, W., Faloutsos, C., Haas, P.J., Hellerstein, J.M., Ioannidis, Y., Jagadish, H.V., Johnson, T., Ng, R., Poosala, V., Ross, K.A., Sevcik, K.C.: The New Jersey data reduction report. IEEE Data Eng. Bull. **20**, 3–45 (1997)

5. Frank, A., Asuncion, A.: UCI Machine Learning Repository. Irvine, CA, University of California, School of Information and Computer Science. http://archive.ics.uci.edu/ml (2010)
6. Ioannidis, Y.: The history of histograms (abridged). In: Proceeding of the 29th International Conference on Very Large Data Bases, 09–12 Sept 2003, pp. 19–30, Berlin, Germany (2003)
7. Muthukrishnan, S., Poosala, V., Suel, T.: On rectangular partitionings in two dimensions: Algorithms, complexity, and applications. In: Beeri, C., Buneman, P. (eds.) Database Theory. In: ICDT '99, 7th International Conference, Jerusalem, Israel, 10–12 Jan 1999, Proceedings. LNCS, vol. 1540, pp. 236–256. Springer, Heidelberg (1999)
8. Pagel, B.-U., Six, H.-W., Toben, H., Widmayer, P.: Towards an analysis of range query performance in spatial data structures. In: Proceedings of 12th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database System Washington, DC (1993)
9. Poosala, V. Ioannidis Y.: Selectivity estimation without the attribute value independence assumption. In: Proceedings of the 23rd International Conference on Very Large Databases, Athens, Greece (1997)
10. Wang, H., Sevcik, K.C.: A multi-dimensional histogram for selectivity estimation and fast approximate query answering. In: Proceedings of the 2003 Conference of the Centre for Advanced Studies on Collaborative Research, Toronto, ON (2003)

On Building and Visualizing Proximity Graphs for Large Data Sets with Artificial Ants

Julien Lavergne, Hanane Azzag, Christiane Guinot, and Gilles Venturini

Abstract We present in this paper a new incremental and bio-inspired algorithm that builds proximity graphs for large amounts of data (i.e. 1 million). It is inspired from the self-assembly behavior of real ants where each ant progressively becomes attached to an existing support and then successively to other attached ants. The ants that we have defined will similarly build a complex hierarchical graph structure. Each artificial ant represents one data. The way ants move and connect depends on the similarity between data. Our hierarchical extension, for huge amounts of data, gives encouraging running times compared to other incremental building methods and is particularly well adapted to the visualization of groups of data (i.e. clusters) thanks to the super-node structure. In addition the visualization using a force-directed algorithm respects the real distances between data.

1 Introduction

We deal in this paper with the following problem: given a large data set of n data $d_1, \ldots, d_n$ (i.e. 1 million) and a similarity measure between these data, we would like to allow the domain expert to explore this large amount of data in a visual and content-based way. We consider that the expert would like to get an overview of the data as well as details [11] obtained by the local exploration of a neighborhood relation. This relation is based on the similarity between data. The problem of exploring a large data set can be solved by incrementally etablishing a hierarchical proximity graph between the data according to their similarities and by visualizing the graph and facilitating its exploration.

We thus concentrate on methods which can be used for building neighborhood graphs (see a survey in [6]). These graphs, also called proximity graphs, are used in data mining, machine learning or clustering. However, these standard algorithms have a high complexity, and therefore they may not be efficient for large data sets. To overcome this limitation, extensions to these methods have been proposed in order

J. Lavergne (✉)
Computer Science Laboratory, François-Rabelais University, Tours, France,
e-mail: julien.lavergne@univ-tours.fr

B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-13312-1_16,

to incrementally build a proximity graph, but these are not suitable for exploration tasks [6]. For this purpose, we have studied a biomimetic method for the incremental construction of such graphs. We proposed an approach which builds a hierarchical proximity graph suitable for interactive visual exploration tasks.

The remaining of this paper is organized as follows: in Sect. 2, we present the principles of self-assembly behavior and our initial algorithm, AntGraph, which incrementally builds a proximity graph from a data set. We also detail the local decision rules which govern the ants behavior. In Sect. 3, we present the main principles of our approach, which extends our initial algorithm, to build a hierarchical proximity graph from a large data set (i.e. 1 million). We detail a new model which introduces a hierarchical structure in the proximity graph construction. Section 4 is devoted to the experimental results obtained on several databases and a comparison with another incremental method [6]. Finally Sect. 5 concludes on this work and presents perspectives.

2 Initial Bio-Inspired Algorithm

Several clustering problems have been solved with bio-inspired algorithms like genetic algorithms [7] or swarm intelligence [3]. These population-based methods have several advantages: they can be run in a distributed way and do not require data pre-classification (or initial information such as the number of classes). They can deal with numerical data as well as symbolic and textual data (provided that a similarity or distance can be defined). Several of them can be easily extended to deal with incremental clustering. In this paper, we deal with such approaches in order to cluster data with artificial ants by building a proximity graph [6].

The initial algorithm that we have extended is called AntGraph [9], and is itself inspired from AntTree [1], a hierarchical clustering algorithm. AntGraph learns a graph-based clustering of data. It is inspired by the self-assembly behavior of real ants that progressively build a living structure by connecting their bodies together [10]. We have generalized this principle for building proximity graphs. The algorithm takes as input a data set of individuals described either by features or by a similarity matrix. The set is large and considered as a stream (which may be randomly sorted). The first ant of the stream is selected and denoted by a_1. This ant will be the support of the structure and the input node of the graph. Then, the remaining ants are introduced one by one in the graph. Let a_i denotes such an ant. a_i moves in the graph until it finds a convenient location where to connect. For this purpose, a_i follows the path of maximum similarity (see Fig. 1). Let a_{pos} denotes the ant (node) where a_i is located. The following cases have to be considered: (1) a_{pos} does not have any neighbor to explore: a_i connects to a_{pos}; (2) a neighbor of a_{pos} is more similar to a_i than a_{pos}: a_i moves onto this neighbor; (3) a_{pos} is the most similar ant to a_i (in a_{pos} local neighborhood): a_i connects to a_{pos} and to all neighbors of a_{pos} which are similar enough (i.e. their similarity to a_i is above a given similarity threshold S_t). This threshold is locally computed as follows: $S_t = \alpha * sim(a_i, a_{pos})$ with $\{\alpha \in \mathbb{R}, 0 \leq \alpha \leq 1\}$.

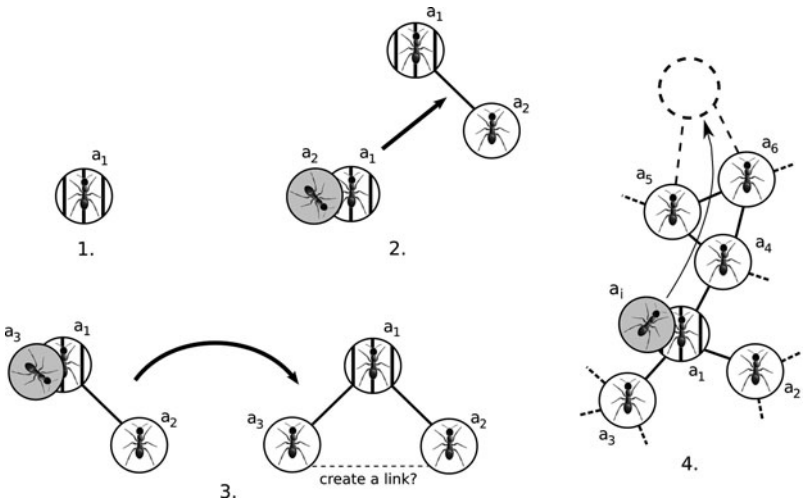


Fig. 1 Building principle with artificial ants. a_i moves from ant to ant (i.e. it follows the path of greatest local similarity). Steps 1, 2 and 3 illustrate the beginning of the construction of a proximity graph whereas step 4 generalizes this construction

Ants become rapidly connected because they locally follow the path of greatest similarity, a way to cut through large data sets. One may notice that we can easily add some data without restarting the algorithm. AntGraph is thus incremental. We have compared its results with standard algorithms for building proximity graphs (Relative Neighborhood Graph [13]) with data sets containing up to 5620 data (Optdigits data set, [2]). We propose in the next section an extension of this algorithm in order to (1) cluster a much larger amount of data (up to 1 million) in a very short time; (2) visualize and navigate through the obtained hierarchical proximity graph.

3 Hierarchical Approach and Visualization

Our approach deals with the construction of a hierarchical graph for a large amount of data using the AntGraph building principle [9]. Consider a graph built with AntGraph for a large amount of data (i.e. 1 million). The visualization and exploration of this graph (i.e. 1 million of nodes and many more links) with a force-directed algorithm is impossible [5]. We cannot visualize distinct clusters with all data and their neighborhood relations.

We propose the use of a *super-node* model which introduces a hierarchical structure in the graph. Our idea is the following one: we consider that each node (data/ant) of a proximity graph can become a super-node which can be the support of a sub-graph (with its own localized neighborhood relations). Each super node thus contains a proximity graph, and each node of this graph can recursively become a super-node. More precisely, each ant in a graph is basically a standard node (i.e. it

contains one data only as in the previous approach) but when it becomes a super-node (sub-graph) it may contain a graph with a maximum of r data (for instance, $r = 500$). The building principle described in the previous section has been modified in the following way.

First, the initial support ant a_1 is considered as a super-node. The other ants enter the graph by this super-node. When such a moving ant a_i has found a convenient location a_{pos} where to connect, it becomes connected to a_{pos} if and only if the super-node is not saturated. If the maximum number of nodes in the current super-node has been reached, then a_{pos} is turned into a super-node. a_i enters this super-node and repeats the same algorithm until it becomes connected. When all ants are connected, we obtain a hierarchical graph, where each sub-graph is "small" enough in order to be displayed with force-directed algorithm [5]). This algorithm benefits from the previous properties of AntGraph: the construction is incremental and fast (see next section), the similarities between data are represented in a proximity graph. Using the hierarchical structure, we can deal with much larger data sets.

As far as visualization is concerned, the structured graph can be displayed in the following way: let us consider a given super-node. The size of this node is limited and can be displayed in 2D or 3D. We illustrate this with Figs. 2 and 3, which respectively represents the 2D visualizations of the first super-node of two hierarchical proximity graphs built with our extension from Optdigits and Waveform data sets [2] whereas the Fig. 4 displays in 3D the first subgraph of another hierarchical proximity graph built from the Gen-1,000,000 data set.

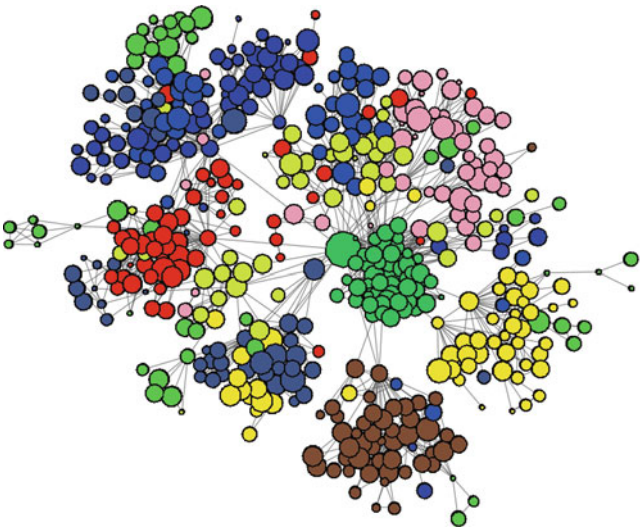


Fig. 2 Visualization of the super-node ant a_1 for the Optdigits data set with a $r = 500$. Different colored clusters can be seen which correspond to 10 real classes. We can notice that the visual size of nodes is relative to the number of data/nodes they contain

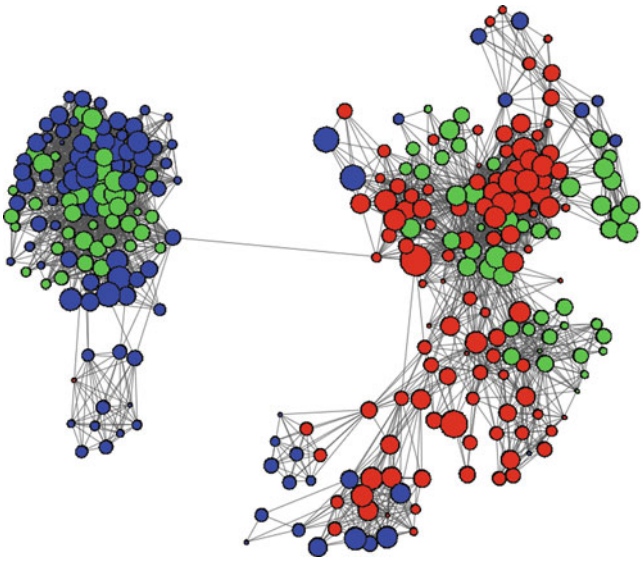


Fig. 3 Visualization of the super-node ant a_1 for the Waveform data set with a r value fixed to 300 (maximum size of a super-node (subgraph)). Different colored clusters can be seen which correspond to 3 real classes

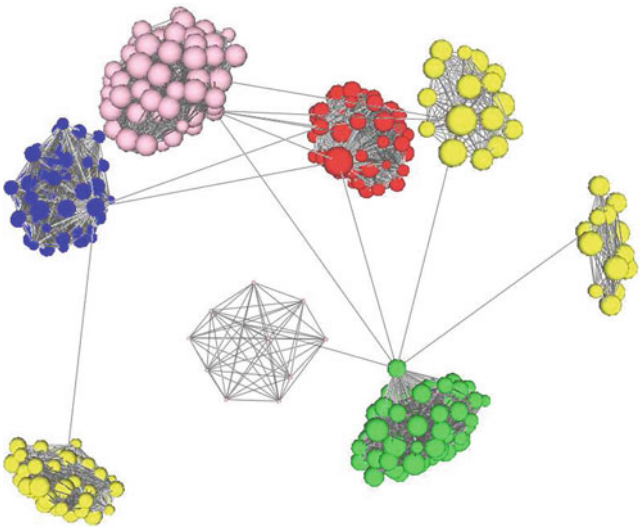


Fig. 4 Visualization of the super-node ant a_1 for the Gen-1000000 data set with a r value fixed to 300. Different colored clusters can be seen which correspond to 5 real classes

As we can see, the user (domain expert or not) may easily view the groups of data. He can also distinguish centered or isolated data. Several interactions are possible like 3D displacements around the graph, zoom with distortion, visualizing the size of a super-node, and a stereoscopic mode (i.e. immersion in a virtual reality environment). Moreover when clicking on a super-node, the user may enter this super-node and our method displays the content of this super node (i.e. force-directed visualization of the graph takes a few seconds only for limited number of nodes, i.e. 500). Thus we have developed an application to evaluate the clustering quality of our algorithm with real users (user validation process). We compared our method to three other clustering methods (one interactive, i.e. POI [4] and two automatic, i.e. Ascendant Hierarchical Clustering (AHC) [8, 12] and AntTree) on several artificial and real-world databases [1, 2]. To respect the limited number of pages, we can summarize that our method is able to cluster data with a quality close to POI and AHC, even slightly better than AntTree (all these methods encounter difficulties on databases with overlaps). Currently we continue experiments with real users on larger databases.

4 Results

We have performed an experimental study on numerical databases (artificial and real-world) having from 5,000 to 1 million data. We have generated some artificial data sets (i.e. Gen-{10000, 25000, 60000, 100000 and 1000000}) with uniform laws (i.e. each set owns five non-overlapping classes). The real-world data sets (Waveform with 5,000 data, Letter Recognition with 20,000 data and Forest Covertype with 581,012 data) come from the UCI Repository of Machine Learning [2]. All sets have been randomly sorted.

We have measured execution times for building hierarchical proximity graphs for different sizes of databases (see above). In the Table 1, our method gives very interesting execution times for all tested databases. For a set of 1 million of data, we respectively build, according to the maximal size of a super-node (i.e. tested values {300, 500 and 700}), a hierarchical proximity graph in less than 27 min, 51 min and 71 min. The Forest Covertype data set requires additional running time because the number of attributes is more important (i.e. 54): in our incremental approach, the similarities are computed in real-time (and cannot be pre-computed due to the size of the data set).

We have also performed a comparative study between our approach and another incremental proximity graphs construction method [6]. Authors consider in their experiments large data sets from 5,000 to 75,000 with a dimension of 50. We have reported execution times obtained with that method in Table 2. We can notice that to build a proximity graph from the 75,000 data set, that algorithm requires 156 min whereas our approach etablishes a proximity graph from the Forest Covertype data set (i.e. 581,012 data with 54 attributes) in less than 43 min (for a size of super-node $r = 300$). Even for a maximum of super-node cardinality of 700, we obtain

Table 1 T_{exec} is execution times in seconds for tested data sets with $r = \{300, 500 \text{ and } 700\}$ (maximum number of nodes in a super-node). C, M and N are respectively the number of real classes, data, and attributes

Bases	C	M	N	$r = 300$ T_{exec}		$r = 500$ T_{exec}		$r = 700$ T_{exec}	
Waveform	3	5,000	21	3.324	0.389	3.872	0.373	5.610	0.722
Gen-10,000	5	10,000	10	6.196	1.026	5.908	0.534	6.200	0.151
Letter Recognition	26	20,000	16	16.357	2.456	15.065	1.542	14.989	1.924
Gen-25,000	5	25,000	10	22.720	3.252	20.605	1.895	28.124	4.496
Gen-60,000	5	60,000	10	94.273	5.452	88.342	6.379	82.519	9.328
Gen-100,000	5	100,000	10	184.946	8.392	225.932	13.387	162.836	16.652
Forest Covertype	7	581,012	54	2,536.986	39.028	4,545.774	89.843	6,365.348	103.234
Gen-1,000,000	5	1,000,000	10	1,590.067	32.958	3,075.201	60.34	4,305.867	78.982

Table 2 T_{exec} is execution times in seconds for tested data sets with the other incremental method [6]. M and N are respectively the number of data and attributes

M	N	T_{exec} in seconds
5,000	50	578.4
10,000	50	1,203.0
20,000	50	2,453.4
40,000	50	4,954.8
50,000	50	6,385.8
75,000	50	9,333.6

a proximity graph in 71 min. We point out that the domain expert can visualize neighborhood relations between all data in a hierarchical proximity graph built with our approach. Finally for about the same number of attributes between these data sets and Forest Covertype, we can notice that our method is much faster than [6].

5 Conclusions

As a conclusion, our approach offers the possibility to build, visualize and explore very large proximity graphs. Our approach has several advantages thanks to the properties of our algorithm (both for graphs construction and visualization). It benefits from the initial method AntGraph and also has a simple setup.

It is suitable for processing large data sets and achieves a partitioning of data in the form of a large hierarchical proximity graph in competitive execution times compared to other incremental building approaches [6]. Therefore the domain expert can adapt, according to his needs, the desired number of data to visualize and explore at a time (maximum cardinality of a super-node).

As perspectives, we are currently studying several improvements like providing the user with a global visualization (in the current implementation, the user loses the context, i.e. global shape of the graph, when entering a super-node). We will apply this algorithm to the indexing of large data sets (i.e. documents from the web) in the context of strategic watch and content-based search.

References

1. Azzag, H., Guinot, C., Venturini, G.: Data and text mining with hierarchical clustering ants. In: Swarm Intelligence and Data Mining, Studies in Computational Intelligence, vol. 34, pp. 153–190. Springer, Heidelberg (2006)
2. Blake, C., Merz, C.: UCI repository of machine learning databases. http://archive.ics.uci.edu/ml/ (1998)
3. Bonabeau, E., Dorigo, M., Theraulaz, G.: Swarm Intelligence: From Natural to Artificial Systems. Oxford University Press, New York, NY (1999)
4. Da Costa, D., Venturini, G.: A visual and interactive data exploration method for large data sets and clustering. In: ADMA '07: Proceedings of the 3rd International Conference on Advanced Data Mining and Applications, pp. 553–561. Springer, Berlin Heidelberg (2007)
5. Fruchterman, T., Reingold, E.: Graph drawing by force-directed placement. Softw. Pract. Exp. **21**, 1129–1164 (1991)
6. Hacid, H., Yoshida, T.: Incremental neighborhood graphs construction for multidimensional databases indexing. In: The Canadian Artificial Intelligence Conference, Canadian AI 2007, Montreal, QC, Canada, Lecture Notes in Artificial Intelligence 4509, pp. 405–416 (2007)
7. Jones, D., Beltramo, M.: Solving partitioning problems with genetic algorithms. In: Belew, R., Booker, L. (eds.) ICGA 1991, pp. 442–449. Morgan Kaufmann, San Diego, CA (1991)
8. Lance, G., Williams, W.: A general theory of classificatory sorting strategies: I. Hierarchical systems. Comput. J. **9**(4), 373–380 (1967)
9. Lavergne, J., Azzag, H., Guinot, C., Venturini, G.: Incremental construction of neighborhood graphs using the ants self-assembly behavior. In: 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007), Patras, Greece, vol. 1, pp. 399–406. http://doi.ieeecomputersociety.org/10.1109/ICTAI.2007.115 (2007)
10. Lioni, A., Sauwens, C., Theraulaz, G., Deneubourg, J.L.: The dynamics of chain formation in oecophylla longinoda. J. Insect Behav. **14**, 679–696 (2001)
11. Shneiderman, B.: The eyes have it: a task by data type taxonomy for information visualizations. In: VL '96: Proceedings of the 1996 IEEE Symposium on Visual Languages, Boulder, CO, p. 336. IEEE Computer Society, Washington, DC (1996)
12. Sneath, P.H., Sokal, R.R.: Numerical Taxonomy. W.H. Freeman, San Francisco, CA (1973)
13. Toussaint, G.T.: The relative neighborhood graphs in a finite planar set. Pattern Recognit. **12**, 261–268. http://citeseer.ist.psu.edu/toussaint80relative.html (1980)

Including Empirical Prior Information in Test Administration

Mariagiulia Matteucci and Bernard P. Veldkamp

Abstract In this work, the issue of using prior information in test administration is taken into account. The focus is on the development of procedures to include background variables which are strongly related to the latent ability, adopting a Bayesian approach. Because the desirability of prior information for the ability estimation in item response modelling depends on the goals of the test, only some kinds of educational tests might profit of this approach. The procedures will be evaluated in an empirical context and some recommendations about the use of prior information will be given.

1 Introduction

The typical situation in educational testing is to have a collection of k items designed to measure single or multiple latent traits, commonly denominated *ability*. Furthermore, some kind of collateral information on the individuals may be available. In this paper, the issue of including prior information in an empirical context is considered. In particular, the inclusion of background variables which are strongly related to the latent ability in the item response model is investigated. A linear relation between the background and the latent variables is considered and the empirical prior is modelled through a Bayesian approach. It depends on the goals of the test whether it is desirable to include prior information for ability estimation.

2 Joint Modelling of Measurement Model and Prior Information

Given the responses to k binary items, a single ability to be measured, and collateral information about the examinees, three points should be specified in order to include prior information in the item response model: the relation between the response

M. Matteucci (✉)
Statistics Department "Paolo Fortunati", University of Bologna, 40126 Bologna, Italy
e-mail: m.matteucci@unibo.it

B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-13312-1_17,

variables and the ability, the relation between the ability and the background variables, and the estimation method. From an item response theory (IRT) perspective, the mathematical relationship between the probability of a correct response and the ability may be described by a wide range of models. When both item difficulty and discrimination are considered, the two-parameter normal ogive model (see [5, 6]) may be appropriate in case of binary items

$$P(Y_{ij} = 1|\theta_i, \alpha_j, \delta_j) = \Phi(\alpha_j\theta_i - \delta_j) = \int_{-\infty}^{\alpha_j\theta_i - \delta_j} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz, \qquad (1)$$

where Y_{ij} is the binary response variable of individual i to item j, with $i = 1, \ldots, n$ and $j = 1, .., k$, α_j and δ_j are the item discrimination and difficulty, respectively, θ_i is the latent ability of person i, and Φ is the standard normal cumulative distribution function. Abilities $\theta_1, \ldots, \theta_n$ are assumed to be a random sample from a normal distribution with mean equal to 0 and standard deviation equal to 1 for identification purposes. Model (1) is usually preferred for Bayesian estimation, particularly using the Gibbs sampler [4] within the MCMC methods.

The introduction of prior information in the model can be performed by considering a set of P covariates or background variables X_p, with $p = 1, \ldots, P$, which are strongly related to the latent ability θ, as described in the following equation

$$\theta_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_P X_{iP} + \varepsilon_i, \qquad (2)$$

where the error terms are assumed to be independent and normally distributed as $\varepsilon_i \sim N(0, \sigma^2)$. Following Eq. (2), the ability is assumed to be measured by the background variables with error. Therefore, the conditional distribution of θ_i, given the X_p's, is normal

$$\theta_i|X_{i1}, \ldots, X_{iP} \sim N(\beta_0 + \beta_1 X_{i1} + \ldots + \beta_P X_{iP}; \sigma^2). \qquad (3)$$

The measurement model together with the structural part represents a particular case of the MIMIC model, where indicators are allowed to be also continuous (see e.g. [10]). When cluster effects are present, a multilevel factorial model can be used. The direct estimation of the β's coefficients and the residual variance σ^2 may be conducted substituting relation (2) into the IRT model. This kind of estimation has been conducted through MML via EM algorithm for the Rasch model by [9, 10, 11] and for the two-parameter logistic model by [12]. On the other hand, the inclusion of prior information is possible within the MCMC methods. The Gibbs sampler has been implemented to estimate a general multilevel IRT model, where covariates of first and second order are included [2]. Furthermore, the same algorithm has been used to model hierarchically the measurement model (1) and prior information in the form of response times [3].

Starting from the application of the Gibbs sampler to model (1) by [1], the idea is to extend the algorithm with the inclusion of prior information on θ. From another point of view, the work can be seen as the equivalent of simplifying the

multilevel IRT model described in [3] to a one-level model with covariates on the latent ability. The presence of a dichotomous variable Y_{ij}, indicating correct or incorrect response of person i to item j, is modeled through the introduction of the underlying variables Z_{ij}, conditionally independent and identically distributed as $Z_{ij} \sim N(\alpha_j\theta_i - \delta_j; 1)$, with $i = 1, \ldots, n$ individuals and $j = 1, \ldots, k$ items. From a fully Bayesian perspective, the joint posterior distribution of interest is

$$P(\mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{\beta}, \sigma^2|\mathbf{Y}, \mathbf{X}) = P(\mathbf{Z}|\boldsymbol{\theta}, \boldsymbol{\xi}, \mathbf{Y})P(\boldsymbol{\theta}|\boldsymbol{\beta}, \sigma^2, \mathbf{X})P(\boldsymbol{\xi})P(\boldsymbol{\beta})P(\sigma^2), \tag{4}$$

where $\boldsymbol{\xi}$ is the vector of the item parameters and the other variables are in matrix notation. Because distribution (4) has an intractable form, the Gibbs sampler algorithm can be applied in order to iteratively sample from the following conditional distributions:

1. $\mathbf{Z}|\boldsymbol{\theta}, \boldsymbol{\xi}, \mathbf{Y}$;
2. $\boldsymbol{\theta}|\mathbf{Z}, \boldsymbol{\xi}, \boldsymbol{\beta}, \sigma^2$;
3. $\boldsymbol{\xi}|\boldsymbol{\theta}, \mathbf{Z}$;
4. $\boldsymbol{\beta}|\boldsymbol{\theta}, \sigma^2$;
5. $\sigma^2|\boldsymbol{\theta}, \boldsymbol{\beta}$.

The conditional distribution of the independent Z_{ij} is normal, with expected value $\eta_{ij} = \alpha_j\theta_i - \delta_j$ and variance equal to 1, truncated by 0 to the left if $Y_{ij} = 1$ and to the right if $Y_{ij} = 0$

$$Z_{ij}|\boldsymbol{\theta}, \boldsymbol{\xi} \sim \begin{cases} N(\eta_{ij}, 1) \text{ with } Z_{ij} > 0 \text{ if } Y_{ij} = 1, \\ N(\eta_{ij}, 1) \text{ with } Z_{ij} \leq 0 \text{ if } Y_{ij} = 0. \end{cases} \tag{5}$$

To obtain the conditional distribution of the $\boldsymbol{\theta}$, we should start from the normal regression model $Z_{ij} = \alpha_j\theta_i - \delta_j + \upsilon_{ij}$ to obtain

$$Z_{ij} + \delta_j = \alpha_j\theta_i + \upsilon_{ij}, \tag{6}$$

where υ_{ij} i.i.d.$\sim N(0, 1)$. Eq. (6) describes the multiple regression of $(Z_{ij} + \delta_j)$ on the regressors α_j, with $j = 1, \ldots, k$, considering θ_i as regression coefficient. Therefore, the likelihood function of θ_i follows the normal distribution with mean equal to the least square estimate of θ_i, specifically $\hat{\theta}_i = \sum_{j=1}^{k} \alpha_j(Z_{ij} + \delta_j)/\sum_{j=1}^{k} \alpha_j^2$, and variance $v = 1 \Big/ \sum_{j=1}^{k} \alpha_j^2$. The prior distribution for θ_i is expressed by (3); it follows that the θ_i's are independent and their posterior distribution is normal and parameterized as

$$\theta_i|\mathbf{Z}, \boldsymbol{\xi}, \boldsymbol{\beta}, \sigma^2 \sim N\left(\frac{\hat{\theta}_i/v + \mathbf{X}_i\boldsymbol{\beta}/\sigma^2}{1/v + 1/\sigma^2}; \frac{1}{1/v + 1/\sigma^2}\right). \tag{7}$$

By following the same approach, consider the normal regression model

$$\mathbf{Z}_j = [\boldsymbol{\theta} \ -\mathbf{1}]\boldsymbol{\xi}_j + \boldsymbol{\upsilon}_j, \tag{8}$$

where $\boldsymbol{\theta}$ is the n-dimensional vector of individual abilities, -$\mathbf{1}$ is a n-dimensional vector with entries equal to -1, $\boldsymbol{\xi}_j = [\alpha_j; \delta_j]'$ is the vector of item parameters for item j, and $\boldsymbol{\upsilon}_j = (\upsilon_{1j}, \ldots, \upsilon_{nj})'$ is a random sample from a standard normal distribution. The model can be interpreted as the regression of $\mathbf{Z}_j$ on the explanatory variables $\mathbf{W} = [\boldsymbol{\theta} \ -\mathbf{1}]$, considering the $\boldsymbol{\xi}_j$ as regression coefficients. Therefore, the likelihood function for the $\boldsymbol{\xi}_j$ follows a multivariate normal distribution with mean equal to the usual least squares estimate $\hat{\boldsymbol{\xi}}_j = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{Z}_j$ and variance equal to $(\mathbf{W}'\mathbf{W})^{-1}$. A possibile choice for the prior distribution is a prior covariance matrix for the item parameters denoted by

$$\boldsymbol{\Sigma}_{\mathbf{0}} = \begin{pmatrix} s_\alpha^2 & 0 \\ 0 & s_\delta^2 \end{pmatrix},$$

where s_α and s_δ are the prior standard deviations for α_j and δ_j. In this case, the conditional posterior distribution of $\boldsymbol{\xi}_j$ is a multivariate normal as follows

$$\boldsymbol{\xi}_j|\boldsymbol{\theta}, \mathbf{Z} \sim N((\mathbf{W}'\mathbf{W} + \boldsymbol{\Sigma}_{\mathbf{0}}^{-1})^{-1}\mathbf{W}'\mathbf{Z}_j; (\mathbf{W}'\mathbf{W} + \boldsymbol{\Sigma}_{\mathbf{0}}^{-1})^{-1}). \tag{9}$$

The algorithm can be applied with different prior distributions for item parameters: the most common is to use $P(\boldsymbol{\xi}) \propto \prod_{j=1}^{k} I(\alpha_j > 0)$ to ensure positive discriminations (see [1, 2]). The posterior distribution of $\boldsymbol{\beta}$, depending on $\boldsymbol{\theta}$ and σ^2, can be found starting from the regression (2), which in vector notation is

$$\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{10}$$

where $\boldsymbol{\theta}$ is the n-dimensional vector of abilities, $\mathbf{X}$ is the $n \times (P + 1)$ matrix of covariates, with the first column of ones to model the intercept term, $\boldsymbol{\beta}$ is the vector of regression coefficients of length $P+1$ and $\boldsymbol{\varepsilon}$ is the n-dimensional vector of the error terms. With a non informative prior on $\boldsymbol{\beta}$, the posterior distribution follows

$$\boldsymbol{\beta}|\boldsymbol{\theta}, \sigma^2 \sim N(\hat{\boldsymbol{\beta}}; \sigma^2(\mathbf{X}'\mathbf{X})^{-1}), \tag{11}$$

where $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\theta}$.

Finally, a conjugate prior for the variance σ^2 is a scaled inverse Chi-square distribution with parameters ν_0 and σ_0^2. If $\nu_0 = 0$ we obtain a non informative prior for σ^2, i.e. $P(\sigma^2) \propto \sigma^{-2}$. The likelihood function for σ^2 is proportional to $\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\theta_i - \mathbf{X}_i\boldsymbol{\beta})^2\right]$. Therefore, the posterior distribution of σ^2 becomes

$$\sigma^2|\boldsymbol{\theta}, \boldsymbol{\beta} \sim Inv - \chi^2(n; S^2), \tag{12}$$

where $S^2 = \frac{1}{n}(\boldsymbol{\theta} - \mathbf{X}\boldsymbol{\beta})'(\boldsymbol{\theta} - \mathbf{X}\boldsymbol{\beta})$.

Starting from a set of starting values, the Gibbs sampler proceeds with the iterative sampling from the conditional distributions until convergence. The choice of starting values is not crucial in MCMC but reasonable initial points may speed convergence.

3 Simulation Study

In this section, simulation studies are presented to understand the parameter recovery of the Gibbs sampler in the estimation of the 2PNO model extended to the presence of background variables on latent variables described in Sect. 2. The algorithm has been implemented in MATLAB [7].

In the first study, binary responses to 14 items have been generated for 1,500 individuals according to model (1). Furthermore, a simple regression has been considered to model the relationship between the ability and the covariate, i.e. $\theta = -0.3 + 0.6X_1 + \varepsilon$, where $\varepsilon \sim N(0, 0.46)$ and X_1 has been drawn from $N(0.5, 1.5)$. True values for the item parameters are shown in the second and third column of Table 1, respectively. For the simulation, 100 dataset have been generated according to the above specifications. For a single dataset, the Gibbs sampler has a run length of 5,000 iterations, with a burn-in of 500, and took around 3 minutes on a Intel(R) Pentium(R) M processor 1.73 GHz. Totally, the algorithm took around 4 h to estimate the joint model for 100 samples. The starting values have been fixed to zeros for both difficulty parameters and ability scores, while all initial discrimination parameters are set to one. Prior standard deviations $s_\alpha = s_\delta = 1$ have been chosen for the item parameters. Results for the item parameter estimates are shown in Table 1.

Table 1 Item parameter recovery[a]

	True values		Gibbs sampler	
Item	α_j	δ_j	$\hat{\alpha_j}$	$\hat{\delta_j}$
01	0.675	−1.041	0.688 (0.058)	−1.042 (0.051)
02	0.585	0.480	0.591 (0.048)	0.474 (0.040)
03	0.240	0.868	0.242 (0.042)	0.869 (0.038)
04	0.662	0.688	0.678 (0.048)	0.692 (0.045)
05	0.143	−0.086	0.146 (0.043)	−0.086 (0.036)
06	1.272	0.093	1.269 (0.086)	0.095 (0.046)
07	0.369	−0.031	0.375 (0.038)	−0.030 (0.034)
08	0.644	−0.277	0.649 (0.047)	−0.277 (0.035)
09	0.681	−1.079	0.680 (0.052)	−1.081 (0.055)
10	0.621	−0.161	0.625 (0.050)	−0.168 (0.042)
11	0.532	0.543	0.537 (0.043)	0.538 (0.041)
12	0.984	0.673	0.995 (0.075)	0.669 (0.055)
13	0.667	0.457	0.670 (0.050)	0.462 (0.038)
14	0.345	−0.654	0.349 (0.040)	−0.656 (0.037)

[a]standard deviations in brackets

Table 2 Parameter recovery for the regression coefficients and the variance of the error term[b]

	True values	Gibbs sampler
β_0	−0.3	−0.299 (0.018)
β_1	0.6	0.602 (0.012)
σ^2	0.46	0.459 (0.021)

[b]standard deviations in brackets

The estimated parameters seem to reflect quite precisely the true values for the item parameters. In Table 2 the estimates for the regression coefficients and the variance of the error term are reported.

The algorithm is able to recover with high precision the parameters of the regression between the ability and the covariate. The Gibbs sampler is not sensitive to the choice of starting values; furthermore, the algorithm has shown a fast convergence to the true values.

The second simulation study has been conducted to show the parameter recovery of the regression parameters according to different sample sizes and test lengths (7 and 14 items). Data have been simulated for 14 items (see Table 1 for the item parameters) and a sample size of 1, 500, 700 and 300. The regression model underlying the data is the same of the first simulation study. Then, data have been simulated for 1, 500, 700 and 300 individuals only for the first 7 items. The results of the MCMC procedure with 5, 000 iterations and a burn-in phase of 500 are shown in Table 3.

The regression parameters are reproduced quite precisely for all the combinations of test length and sample size and this encourages the use of the algorithm in different situations. Anyway, we can notice that as both the sample size and the number of items decrease, the precision of estimates decreases, especially for the variance σ^2.

4 Real Data Application

Data were available from 666 test takers that completed an intelligence test in the Netherlands [8]. The items have been pre-calibrated by using the two-parameter logistic model, therefore maximum likelihood estimates of the ability are known for all the examinees.

In the analysis, the relation between Number series responses as collateral information for the estimation of the Raven's matrices test is investigated. A multiple regression model for the ability in Raven's matrices on Number series performance, educational background and minority orientation has been specified. The model is $\theta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \varepsilon_i$, where X_1 represents the estimated abilities on the Number series items, X_2 is the education variable equal to 1 for university and 0 for higher vocational, and X_3, X_4 are indicator variables for the minority orientation ($X_3 = 1$ for Dutch native, $X_4 = 1$ for western immigrant, and non-western immigrant reference category). The Gibbs sampler has been run for 5,000 iterations and a burn-in of 500 iterations has been considered to estimate the

Table 3 Parameter recovery for the regression coefficients and the variance of the error term with different sample size and number of items[c]

		Gibbs sampler					
		$k = 14$			$k = 7$		
True values		$n = 1,500$	$n = 700$	$n = 300$	$n = 1,500$	$n = 700$	$n = 300$
β_0	−0.3	−0.299 (0.018)	−0.297 (0.029)	−0.297 (0.043)	−0.298 (0.022)	−0.297 (0.029)	−0.289 (0.045)
β_1	0.6	0.602 (0.012)	0.598 (0.019)	0.599 (0.033)	0.597 (0.022)	0.595 (0.025)	0.591 (0.046)
σ^2	0.46	0.459 (0.021)	0.469 (0.034)	0.476 (0.052)	0.468 (0.036)	0.471 (0.043)	0.484 (0.079)

[c]standard deviations in brackets

Table 4 Item parameter estimates for the Raven's matrices test

$\hat{\alpha_j}$	sd($\hat{\alpha_j}$)	$\hat{\delta_j}$	sd($\hat{\delta_j}$)
0.476	0.075	−0.503	0.055
1.033	0.141	−0.720	0.076
0.599	0.085	−0.341	0.053
0.181	0.065	0.363	0.050
0.526	0.077	−0.346	0.053
0.952	0.124	−0.418	0.063
0.903	0.138	−1.354	0.116

Table 5 Estimated parameters for the linear regression of ability on Raven's matrices on all covariates [d]

		Estimates
β_0	intercept	0.037 (0.078)
β_1	number series	0.655 (0.068)
β_2	university	0.118 (0.095)
β_3	Dutch	0.176 (0.104)
β_4	western immigrant	−0.012 (0.134)
σ^2	variance	0.763 (0.052)

[d]standard deviations in brackets

IRT model and the regression equation for the ability jointly. Table 4 shows the item parameter estimates for the Raven's matrices subscale.

The results on the coefficients and on σ^2 for the complete regression model are described in Table 5.

The estimated $\beta_1 = 0.655$ highlights a positive and large effect of the Number series predictor on the Raven's matrices ability, given the standard normal scale of the ability. The results show a small positive effect of the university respect to the higher vocational education because $\beta_2 = 0.118$. Furthermore, a moderate positive effect is noticed for Dutch natives with respect to immigrants ($\beta_3 = 0.176$) while the effect of being a western immigrant is not significant and almost null ($\beta_4 = -0.012$).

5 Concluding Remarks

In this work, a Gibbs sampler algorithm has been developed to include in the estimation of the 2PNO model a series of background variables which can be predictive of the ability level. The simulation study has shown the capability of the algorithm of recovering both item and regression parameters. Finally, the application to data from intelligence tests has been given as an example of the context in which the algorithm can be used. In operational testing, the algorithm can be used to initialize the ability estimation, given the estimated coefficients for a set of related covariates. This can be especially useful in adaptive testing.

References

1. Albert, J.H.: Bayesian estimation of normal ogive item response curves using Gibbs sampling. J. Educ. Stat. **17**(3), 251–269 (1992)
2. Fox, J.P., Glas, C.A.W.: Bayesian estimation of a multilevel IRT model using Gibbs sampling. Psychometrika **66**(2), 271–288 (2001)
3. Fox, J.P., Klein Entink, R., Van der Linden, W.J.: Modeling of response times with the package cirt. J. Stat. Softw. **20**(7), (2007)
4. Geman, S., Geman, D.: Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. **6**, 721–741 (1984)
5. Lord, F.: A theory of test scores. Psychom. Monogr. **7** (1952)
6. Lord, F., Novick, M.: Statistical Theories of Mental Test Scores. Addison-Wesley, Reading, MA (1968)
7. MATLAB 7.1. The MathWorks Inc., Natick, MA (2005)
8. Schakel, L., Smid, N., Maij-de Meij, A.M.: Connector Ability Manual. PiCompany B.V., Utrecht, The Netherlands (2009)
9. Skrondal, A., Rabe-Hesketh, S.: Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models. Chapman & Hall/CRC, Boca Raton, FL (2004)
10. Zwinderman, A.H.: A generalized Rasch model for manifest predictors. Psychometrika **56**(4), 589–600 (1991)
11. Zwinderman, A.H.: Response models with manifest predictors. In: Van der Linder, W.J., Hambleton, R.K. (eds.) Handbook of Modern Item Response Theory, pp. 245–256. Springer, New York (1997)
12. Van der Linden, W.J.: Empirical initialization of the trait estimation in adaptive testing. Appl. Psychol. Meas. **23**(1), 21–29 (1999)

Part IV
Robustness and Classification

Italian Firms' Geographical Location in High-tech Industries: A Robust Analysis

Matilde Bini and Margherita Velucchi

Abstract Recent debates in economic-statistical research concern the relationship between firms' performance and their capabilities to develop new technologies and products. Several studies argue that economic performance and geographical proximity strongly affect firms' level of technology. The aim of the paper is twofold. Firstly, we propose to generalize this approach and to develop a model to identify the relationship between the firm's technology level and some firm's characteristics. Secondly, we use an outlier detection method to identify units that affect the analysis results and the estimates stability. This analysis is implemented using a generalized regression model with a diagnostic robust approach based on forward search. The method we use reveals how the fitted regression model depends on individual observations and the results show how the firms' technology level is influenced by their geographical proximity.

1 Introduction

Recent economic-statistical literature investigates the relationship between firms' performance and their capabilities to develop new technologies and products. Several studies argue that economic performance and geographical proximity strongly affect firms' level of technology (high-tech and low-tech), proxied by sectors in which they operate. Following this literature, a technological clustering approach has been developed [2]. A technological cluster is a geographical concentration of high technology firms; they often form around scientific research centers, such as universities or national labs ([5]; Fallah and Ibrahim, 2004) or close to larger innovative and internationalized firms [4]. Innovative activities are strongly geographically agglomerated and this has led many researchers to investigate the causes of this phenomenon [7]. Also, innovative behavior varies considerably across regions and sectors [3, 9] and the existence of clusters reveals important insights about the microeconomics of competition and the role of location in competitive advantage of

M. Bini (✉)
Department of Economics, European University of Rome Via degli Aldobrandeschi, 190 00163 Roma
e-mail: mbini@unier.it

B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-13312-1_18,

firms. Clusters' influences on competition have taken on growing importance in an increasingly complex, knowledge-based, and dynamic economies.

The aim of the paper is twofold. Firstly, we suggest a model to identify the relationship between the sector of the firm and some firm's characteristics such as size (proxied by total sales and added value in log terms),[1] age and geographical location. We use large agglomeration as a proxy of geographical location because it is the classification adopted by ISTAT (Istituto Nazionale di Statistica) although this is just a very rough definition of geographical proximity. Secondly, to test this model, we use an outlier detection method so we identify units (firms) that affect the analysis results and the stability of estimates obtained by fitting the model. The method we use reveals how the fitted regression model depends on individual observations and the results show how the firms' technology level is influenced by their geographical proximity. The analysis is implemented using a generalized regression model with a diagnostic robust approach based on forward search [1]. Since the response variable in our model is dichotomous (high-tech and low-tech level), a logistic regression model is fitted. We use a cross section data set of Italian entreprises that presented their balance sheet in 2006 for manufacturing sectors provided by AIDA (Analisi Informatizzata Delle Aziende, Bureau Van Dijk).

2 The Model: A Robust Approach to Detect Outliers

The approach we follow is proposed by Atkinson and Riani [1]. The fit of a logit model on this data set allows to select the significant covariates. We aim at identifying the relationships between the firm's technology level and their characteristics such as size, age and geographical location. The *forward search* algorithm is applied to a regression analysis and it reveals the hidden structure of the data. This approach is useful when extreme, anomalous units are present, helping downweigthing or discarding them. The algorithm is based on a robust fit on very few observations. Then, a successive fit is done with larger subsets. The initial subset is identified using the *least median of squares method* [10] that guarantees that no outliers are included in the initial subset.

Formally, following Atkinson and Riani [1], let $Z = (X, y)$ be a data matrix of dimension $n \times (p + 1)$. If n is moderate and $p << n$, the choice of the initial subset can be performed by exhaustive enumeration of all $\binom{n}{p}$ distinct p-tuple:

$S^{(p)}_{i_1,\ldots,i_p} \equiv \{z_{i_1}, \ldots, z_{i_p}\}$,

where $z^T_{i_j}$ is the ij-th row of Z, for $j = 1, \ldots, p$ and $1 \leq i_j \neq i_{j*} \leq n$.

Specifically, let $\iota^T = [i_1, \ldots, i_p]$ and let $e_{i,S^{(p)}_\iota}$ be the least squares residual for the unit i given the model has been fitted with the observations in $S^{(p)}_\iota$. The initial

[1] Sales and added value have been transformed in log terms for scale reasons and for an easier interpretation of the results.

subset is $S_*^{(p)}$ which satisfies

$$e^2_{[med],S_*^{(p)}} = \min_{\iota} \left[e^2_{[med],S_\iota^{(p)}} \right]$$

where $e^2_{[k],S_\iota^{(p)}}$ is the k-th ordered squared residual among $e^2_{i,S_\iota^{(p)}}$, with $i = 1 \ldots . n$ and med is the integer part of $(n + p + 1)/2$. If $\binom{n}{p}$ is too large, the choice is made using $3,000$ p-tuples sampled from Z matrix. The subset size is increased by one and the model refitted to the observations with the smallest residuals for the increased subset size. The result is an ordering of the observations by closeness to the assumed model. Usually one observation enters the subset used for fitting, but sometimes two or more observations enter the subset as one or more leave. GLMs have been accomplished according to this forward search.

3 Data Set Description and Some Results

The data set we use is a cross section of Italian manufacturing firms. We consider 785 medium sized limited companies (with one associate) having different levels of technological intensity [8], firm's age and geographical location (North-West, North-East, Center, South, Sicily and Sardinia). We limit our analysis to medium sized limited companies because they are the most dynamic and growing among Italian firms, representing the so called *fourth capitalism* [6]. Descriptive statistics in Table 1 show that, as expected, only 28% of the firms operate in a high tech sector and their age is quite high (around 24 year old). From further results that are not included for lack of space, we also see that they live longer than their larger or smaller counterparts. This is in line with the definition of the *fourth capitalism* firms: successful, dynamic and flexible medium sized entreprises. Following the ISTAT classification, geographical location statistics show that the vast majority of these firms are in the North of Italy: 46.8% in the North-East and 35.6% in North-West. 11.8% are from Center and 5% from South and Islands.

We estimate an extended model including all covariates (geographical location, firm's age, size proxied both by added value and sales) and, by means of the forward search algorithm, we select the link logit as the best among the alternatives. Fig. 1 shows that this link fits well but for some units. The plot of the Goodness of Link

Table 1 Descriptive Statistics for medium–sized Italian firms (2006). 785 observations from AIDA data set

	Tech	Sales (log)	Added value (log)	Age
Mean:	0.28	3.77	4.07	24.14
Median:	0.00	3.71	4.02	22.00
St. Dev.:	0.45	0.56	0.14	10.89
Skewness:	0.98	0.47	3.60	1.92

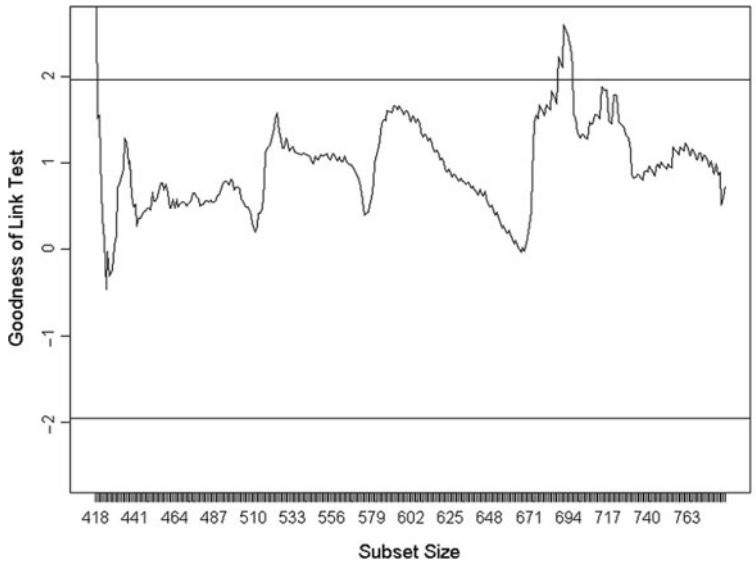


Fig. 1 Forward Search: Goodness of Link Test

Test shows anomalous values at the end and at steps 688–698 of the forward search that lie out of the significant bounds (5% level).[2]

From the preliminary analysis on links emerges that a deeper study is necessary not only for units entering the last steps of the search but also for units that affect the behavior of the link logit (steps 688–698). The last observations entering the model cause an increase in the residual deviance and a decrease in the Pseudo-R^2 statistic.[3]

Figure 2 reports the stability of the estimates showing that the same group of units influences the significance. This figure shows that the estimates are very unstable especially at the beginning of the procedure. Moreover, the significance of the estimates change as sample size increases. Concerning the relevant variables of our analysis, geographical location and firm's age are significant throughout the procedure: mature firms are less likely to operate in high-tech, risky industries and geographical location strongly affects the technological intensity of firms. In particular, it emerges a North-South gradient in technological intensity: working in the northern regions, particularly in the north-east ones, increases the probability of a higher technological intensity while working in the central and southern regions does not statistically affect it. Table 2 reports the estimates results for the whole sample at the final step of the forward search; they coincide with results obtained without the procedure.

[2] This result is due to the presence of units differing from the bulk of the data that may derive from different populations. Further results on links comparison are available upon request.

[3] Detailed results are available upon request.

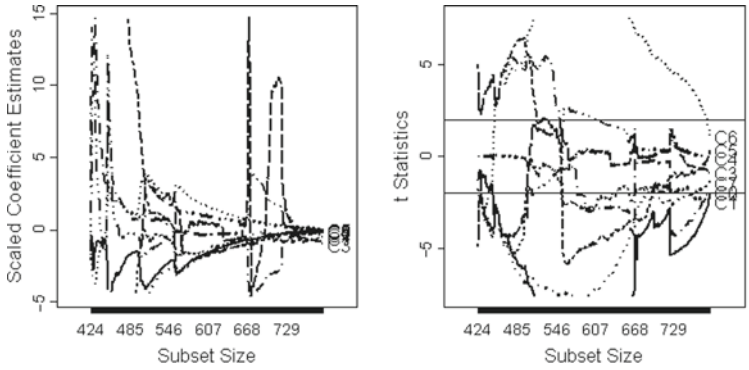


Fig. 2 Forward plots of scaled coefficient estimates and of t values for the coefficients. Whole sample

Variables in Figure 2 are labelled as C1: North-West area, C2: North-East area, C3: Center area, C4: South area, C5: sales (log), C6: added value (log), C7: age.

Table 2 Results of the estimated model. Whole sample

	Coefficient	St.err	t-stat	p-value
Intercept	−4.136	2.132	−1.940	0.053
Area: North-West	−0.213	0.089	−2.381	0.018
Area: North-East	−0.177	0.092	−1.917	0.055
Area: Center	−0.092	0.103	−0.890	0.374
Area: South	−0.013	0.162	−0.078	0.936
Sales	0.073	0.213	0.342	0.734
Added value	0.723	0.662	1.093	0.276
Age	−0.011	0.008	−1.444	0.150

The comparison between Table 1 and Fig. 2 shows the relevance of the procedure in detecting anomalous units affecting both the estimates and their significance as sample size increases. For example, estimates show a sudden peak at steps 688–698 suggesting a further and deeper analysis on those units. Fig. 3 reports the Goodness of Link Test after deletion of units affecting the stability of estimates and the trajectory of the logit Link Test remains within the boundaries (5%). In Fig. 4, coefficients and t-stats from the estimation on the reduced sample point out that geographical location is still significant while the firms' age is only weakly significant. Added value becomes extremely significant throughout the procedure, as confirmed by the estimates from the last step, reported in Table 3. On the one hand, high added value firms are more likely to have high technological intensity while, on the other, mature firms risk less and tend to operate in low tech sectors.

Our analysis reveals that a small number of units (steps 688–698) and the last units entering the forward search (last 26 steps) alter the results. A detailed focus on these units show that they have very peculiar characteristics. The firms identified as outliers entering at steps 688–698 (11 units), turn out to be "made in Italy" (fashion and leather goods), quite large firms with high sales and added value (281,013,000 euro sales and 85,559 euro added value, on average) mainly located in

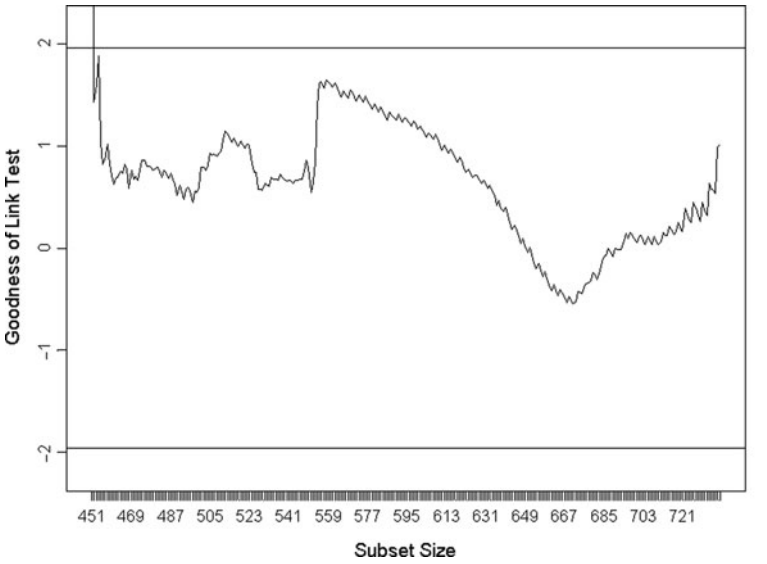


Fig. 3 Forward Search: Goodness of link test after units deletion

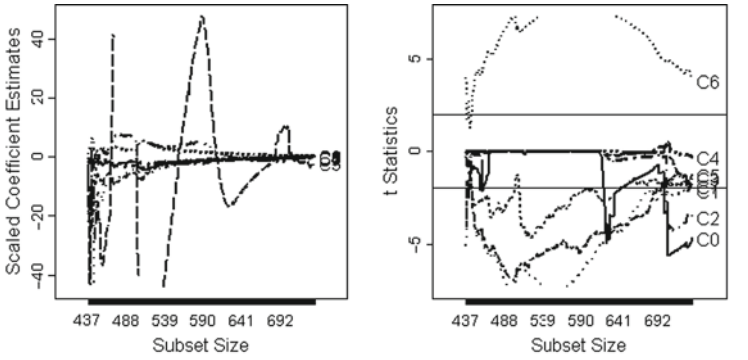


Fig. 4 Forward plots of scaled coefficient estimates and of t values for the coefficients. Whole sample

Variables in Figure 4 are labelled as C1: North-West area, C2: North-East area, C3: Center area, C4: South area, C5: sales (log), C6: added value (log), C7: age.

the North-West industrial districts as shown in Table 4. These firms are internationalized and famous brands; among these, Max Mara, Marina Rinaldi and Marella (textiles), Luxottica (eye glasses), Bonomelli (food). Firms entering last steps of the procedure (last 26 steps), instead, are smaller with both sales and added value lower than the whole sample (6,803,100 euro sales and 1,812 euro added value, on average) located mainly in the North and in the Center of Italy. On average, the outliers are also younger (18.6 years old) and more dynamic firms than these units but they both strongly deviate from the assumed model.

Table 3 Results of the estimated model after outliers deletion

	Coefficient	St.err	t-stat	p-value
Intercept	−16.280	3.506	−4.644	0.000
Area: North-West	−0.194	0.092	−2.096	0.036
Area: North-East	−0.416	0.119	−3.506	0.000
Area: Center	−0.189	0.130	−1.453	0.147
Area: South	−0.061	0.219	−0.278	0.781
Sales	−0.306	0.247	−1.240	0.215
Added value	4.009	1.027	3.904	0.000
Age	−0.016	0.008	−1.849	0.065

Table 4 Results of the estimated model after outliers deletion

Means	obs	High tech	Sales	Added value	North-west	North-east	Center	South	Islands	Age
Sample	785	28%	24,093	5,461	47%	35%	12%	5%	1%	24.2
Outliers	11	33%	281,013	85,559	50%	25%	0%	25%	0%	18.6
Last units	26	28%	6,831	1,812	38%	31%	27%	4%	0%	27.2

The robust analysis shows that the technological level is strongly influenced by added value, geographical location and only weakly by age of firms. This relationship is strong for medium sized firms while it is weak for large sized and internationalized enterprises.

4 Conclusive Remarks

This paper deals with the technological clustering approach and suggests a model to identify the relationships between the firm's technological level and its characteristics (size, age and geographical location). An outlier detection method is used to identify units that affect the results and the stability of estimates. We use a generalized regression model with a diagnostic robust approach based on the forward search algorithm that allows to identify clusters of units (outliers) having peculiar characteristics. This approach better explains firms' heterogeneity and the role of added value and geographical location in creating the fertile context for technological clusters. Our model shows that larger firms in North-East and North-West industrial districts are more likely to have high technological intensity. We identify also a group of firms, outliers, in Made in Italy sectors like fashion and food that do not fit the model and over-perform with respect to the whole sample.

Acknowledgments We are grateful to Luigi Biggeri and Marco Riani for their comments and suggestions that strongly improved this work.

References

1. Atkinson, A., Riani, M.: Robust Diagnostic Regression Analysis. Springer, New York, NY (2000)
2. Breschi, S., Malerba, F.: Clusters, Networks, and Innovation. Oxford University Press, Oxford (2005)
3. Cincera, M.: Patents, R&D and technological spillovers at the firm level: some evidence from econometric count models for panel data. J. Appl. Econ. **12**, 265–280 (1997)
4. De Clercq, D., Hessels, J., Van Stel, A.: Knowledge spillovers and new ventures' export orientation. Small. Bus. Econ. **31**, 283–303 (2008)
5. Galbraith, C.S., Rodriguez, C.L., De Noble, A.F.: SME competitive strategy and location behavior: an exploratory study of high-technology manufacturing. J. Small. Bus. Manage. **46**(2), 183–202 (2008)
6. ISTAT: Rapporto Annuale. La situazione del paese nel 2007, ISTAT, Roma (2008)
7. Nieto, M., Quivedo, P.: Absorptive capacity, technological opportunity, knowledge spillovers and innovative effort. Technovation **25**, 1141–1157 (2005)
8. OECD: Science, Technology and Industry Scoreboard, OECD, Paris (2006)
9. Porter, M.E.: Location, competition, and economic development: local clusters in a global economy, Econ. Devel. Quart. **14**(1), 15–34 (2000)
10. Rousseeuw, P.J.: Least median of square regression. J. Am. Stat. Assoc. **85**, 633–639 (1984)

Robust Tests for Pareto Density Estimation

Aldo Corbellini and Lisa Crosato

Abstract A common practice to determine the extension and heaviness of heavy tails of income, return and size distributions is the sequential estimation and fitting of one or several models, starting from a group of the largest observations and adding one observation at a time [14]. In the early stages this kind of procedure shows high sensitivity of the shape parameter estimates to single observations, the end of the search being fixed when the shape parameter value estimates reach a plateau. In this paper we propose a stepwise fitting of a heavy-tailed model, the Pareto II distribution [1], previously applied to the size distribution of business firms. The procedure, based on the forward search technique [2], is data-driven since observations to be added at each iteration are determined according to the results of the estimation carried out at the preceding step and not, as in sequential fitting, according to their rank.

1 Introduction

Recent literature, in both economics and finance, has dealt with the parametric modelling of heavy tails recurrent in several variables, from personal income to firm size [9], from returns to innovation [18] to financial returns [10, 14].

In cross section datasets observations belonging to the heavy tails are often the most important, because of their relative weight in terms of income, total assets, total returns etc. Parameters that measure the tail thickness are then used in income analysis and industrial economics as inequality and concentration indicators respectively. In financial return time series data, the presence of heavy tails indicates the occurrence of profits under or above the average. Investors and, more generally, stakeholders are obviously concerned about the frequency and magnitude of

A. Corbellini (✉)
Economics Department, Università di Parma, Parma, Italy
e-mail: aldo.corbellini@unipr.it

B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-13312-1_19,

extremal returns, so determining a cut-off value for the tails and estimating their weight is not a secondary matter.

The benchmark distribution to assess the extension and thickness of heavy tails in industrial economics and income studies is the Pareto I distribution [13], whose fitting capacity is unfortunately often restricted to a small percentage of observations (usually the largest). In the empirical financial literature, instead, data are mainly modelled through the stable distribution family [10, 14], which allows for heavy tails both on the positive and on the negative side. In particular, in extreme value theory [3, 7, 11, 15], the limiting threshold of heavy tails and the estimation of the tail index are determined via a sequential procedure. Estimation and fitting start from a group of the largest observations, then proceed adding one observation at each step according to its rank (where the observation number 1 is the largest), and stopping when the estimated value for the shape parameter reaches a certain degree of stability. A well known problem of this kind of procedure is the trade-off between the efficiency and the bias of the estimate, since adding observations leads to a smaller variance but also to higher residuals between true data and the model [18]. As a consequence, the literature has mainly focused on comparing different estimator performances [16] in terms of asymptotic efficiency.

In this paper we propose a robust stepwise testing procedure for the Pareto II [1] distribution [1], which has previously been shown to describe adequately the distribution of business firms in Italy [6]. Our procedure relies on the forward search technique [2], a powerful overall approach suited for detection of masked outliers, for evaluation of their effects on models fitted to data and for investigation of model inadequacy.

Our search procedure has two steps. At the preliminary stage, a group of observations is selected to form a Basic Subset (BSB). The BSB is set up by isolating a group of observations that, by itself, yields the best-fitting Pareto II model to all the observations. Secondly, the observations added at each step are selected only according to their contribution to the goodness-of-fit of the model, and not according to their ranked size. Finally, using Monte Carlo confidence intervals, we are able to determine the percentage of observations consistent with the null hypothesis of Pareto II and to highlight the effect exerted on the fitted model by each observation.

The paper is organized as follows: the next section describes our robust procedure for estimation of the Pareto II distribution's parameters. In Sect. 3 we illustrate some results based on asymptotic confidence intervals obtained for pseudo-random realizations from the Pareto II distribution. In particular, we show that applying our algorithm to Pareto II-type data contaminated by values from a normal distribution leads to a correct identification of the Paretian subset. Sect. 4 summarizes and discusses further research.

[1] The Pareto II distribution is the second model Pareto proposed [13] to describe empirical income distributions. Its distribution function (CDF) is given by $\left[F(x) = 1 - \left(1 + \frac{x-\mu}{\sigma}\right)^{-\alpha}\right]$.

2 Robust Stepwise Fitting of the Pareto II Distribution

The traditional approach to heavy tail fitting [14] relies on the hypothesis that the largest observations form the hard core of some Paretian distribution. Therefore, they are used as the starting group to be enlarged successively by the sequential inclusion of smaller observations, until the desired stability in the parameter estimates is reached. Of course, the focus on extreme values is fully motivated by the interest in the occurrence of rare (extremely below or above the average) events, so that to treat them simply as outliers would be self-defeating. Still, we question the robustness of this procedure, arguing that including in the initial subset extreme values, which actually are by definition atypical observations, does not provide an ordering of the data according to the model.

On the contrary, in the wake of the Forward Search, our approach starts by finding a *Basic Subset* (BSB), defined here as the group of observations[2] that shows the best fit of the Pareto II distribution independently of their position along the ranked observations. The goodness-of-fit is assessed using a robust version of the Pearson statistic, i.e. summarizing the difference between theoretical and actual frequencies along cells by the median. In this paper, the units added at each step of the Forward Search are cells of observations rather than single observations, due to the nature of the Pearson statistic. Thus, also the BSB is constructed on a cell basis.

Once the BSB has been identified, i.e. once we have identified the most "Paretian" observations among the data, we proceed to an adjustment of the estimates based on the inclusion of the remaining observations according to a goodness-of-fit criterion. In fact, at each iteration of the Forward Search, the algorithm first adds the single cell of observations best ranked by the Pearson statistic at the preceding step, and secondly performs Maximum Likelihood Estimation (MLE) on the newfound BSB, that grows with the iterations. In this way, we order all cells up to the last one according to the feasibility of their inclusion in the search.

The first issue we have to face is to formalize the method for selecting the p cells composing the BSB.[3]

Let $X = \{x_1, x_2, \ldots, x_n\}$ be the vector of obervations partitioned in r cells,[4] denoted by $\{c_1, c_2, \ldots, c_r\}$, and let i_j denote the group of observations falling in cell c_j. Now, if r is moderate (e.g. not greater than 100), the choice of the initial subset can be performed by exhaustive enumeration of all $\binom{r}{p}$ distinct p-tuples $S^{(p)} \equiv \{i_1, \ldots, i_p\}$. If $\binom{r}{p}$ is too large, we use instead some arbitrary large number of combinations, say k with $k < \binom{r}{p}$.

[2] For the sake of brevity, we refer to the data undergoing the procedure as observations, wether they are realizations of a random variable observed on a specific statistical unit or pseudo-random realizations of the Pareto II-type.

[3] At this stage we do not assign any particular vale to p in order to maintain a general approach.

[4] Cells are chosen to have equal probabilities under the hypothesized distribution [12].

Selection of the BSB is realized through the following steps.

1. Estimation of the unknown vector of parameters, $\hat{\theta}$, whose components are $\hat{\mu}$, $\hat{\alpha}$ and $\hat{\sigma}$, is carried out separately on the basis of the observations belonging to each $S_j^{(p)}$, for $j = 1, 2, \ldots, k$. For this purpose the Log-likelihood function of the Pareto II distribution [1] is maximized through the constrained optimization routine collection `nlminb` provided by the *R* software.
2. Calculation of the goodness-of-fit Pearson statistic between the overall distribution and the model defined by each $S_j^{(p)}$ is performed on single cells. In particular, let the residual $e_{i,S_j^{(p)}}$ be defined as the relative difference between actual and theoretical frequencies for cell c_i, $i = 1, 2, \ldots, r$, i.e.

$$e_{i,S_j^{(p)}} = [\left(f_a^i - f_{tj}^i\right)^2 / f_{tj}^i]$$

with $j = 1, 2, \ldots, k$. Note that for one cell we have k residuals, each corresponding to a Pareto II model estimated from a different $S_j^{(p)}$.
3. The entire set of $S_j^{(p)}$ combinations is ranked from smallest (rank 1) to largest (rank k) according to the median value of the residuals over all cells, denoted by $e_{med,S_j^{(p)}}$
4. The BSB, $S_*^{(p)}$, is identified as the top-ranked combination, i.e. the one satisfying

$$e_{med,S_*^{(p)}} = \min_{j=1,\ldots,k} [e_{med,S_j^{(p)}}]. \tag{1}$$

Therefore, the BSB is defined as the subset of cells that provides the model best fitting the whole set of observations. Once selected, the BSB is then augmented by one cell at each iteration until all cells are included in the estimation subset.

In the second phase we use a similar procedure to the above. However, since the focus is now on establishing a model-driven order of inclusion for observations, the goodness-of-fit ranking is carried out on single cells and not, as above, on the basis of a summary measure.

In the classical formulation of the forward search for regression, given a subset of dimension $m \geq p$, say $S_*^{(m)}$, we move to dimension $m + 1$ by selecting the $m + 1$ observations with the smallest squared least squares residuals, the observations being chosen by ordering all squared residuals $e^2_{i,S_*^{(m)}}$, $i = 1, \ldots, r$. In most moves from m to $m + 1$ just one new observation joins the subset. It may also happen that two or more observations join $S_*^{(m)}$ as one or more leave. However, our experience is that such an event is quite unusual, only occurring when the search includes one observation that belongs to a cluster of outliers. At the next step the remaining outliers in the cluster seem less outlying and so several may be included at once. Of course, several other observations then have to leave the subset.

Here we observe the same pattern, although not with reference to single observations but to single cells, in that only one cell enters the subset at each step. Let

$\{c'_1, c'_2, \ldots, c'_p, c'_{p+1}, \ldots, c'_r\}$ be all cells reordered such that the first p cells belong to $S_*^{(p)}$. We denote by c'_i, $i = p+1, p+2, \ldots, r$ all remaining cells. In order to enlarge the BSB we proceed along the following path:

1. MLE of the parameters $\hat{\theta}$ based on all observations belonging to $S_*^{(p)}$
2. Calculation of the residual $e_{i,S_j^{(p)}}$ separately for each c'_i, $i = p+1, p+2, \ldots, r$
3. Ordering of the cells not yet included in the estimation subset on the basis of the residual in (2)
4. Addition of the top-ranked cell, c'_*, to $S_*^{(p)}$

In the next iteration, when searching for the $p+2$-th cell to be included, the estimation will be performed on the elements of $S_i^{(p+1)}$=$S_*^{(p)} \cup c'_*$ and so on for all cells.

3 Forward Chi-Square Test

The stepwise procedure presented in the previous section represents an attempt to integrate the processes of estimation and goodness-of-fit in turn in order to assess iteratively which observations have to be considered as realizations of the theoretical model. So far, in fact, we have defined firstly a robust group of observations to serve as a start-up subset for estimation and secondly an ordering of cells according to the estimated model, but we have not yet specified a criterion to mark the cells not consistent with the model itself.

In this section, we apply the procedure of Sect. 2 both to data in large part consistent with the Pareto II distribution[5] and to Pareto II-type data contaminated by values from a normal distribution. Furthermore, using Monte Carlo confidence intervals for the χ^2 statistic [17] built on pseudo-random realizations of Pareto II distribution, we show that the application of our algorithm to the former dataset imply the recognition of almost all data as Pareto II distributed, while application of the same to the latter dataset leads to a correct isolation of the Paretian subset and, as a byproduct, of the outliers.

Firstly we perform the forward search over the empirical dataset [6].

The Forward Search approach typically uses the smallest possible subset to start the estimation with (composed for example by $m = p$ observations, equal to the number of parameters in the regression model). We fix the dimension of the BSB in $p = 4$ cells, which is the minimum number of cells required in order to obtain

[5] Total Assets for Italian firms of the Chemical sector are extracted from the AIDA database, processed and managed by Bureau van Dijck Electronic Publishing. AIDA tracks accounts and activities for 500,000 Italian companies with sales greater than 500,000 Euros, plus ownership and management for the top 20,000 companies.

correct critical points for the Pearson statistic.[6] As about the total number of cells to divide the observations in, we exceed the usual recommended number $r = 2n^{2/5}$ and fix it to $r = 140$ cells in order to reduce the probability of observations conforming to different distributions to occupy the same cell. Also, since in this case $\binom{r}{p}$ exceed 15 millions of combinations, to determine the composition of the BSB we use 1,000 combinations of cells. At each step of the search, the algorithm produces the χ^2 statistics of goodness-of-fit and stores them in a vector of length $r - p = 136$ denoted by $\boldsymbol{\chi}^2_e$, where e stands for empirical.

In a second time, we repeat the above pattern on a Paretian dataset contaminated by positive random draws from a normal distribution. Using estimates from Italian chemical firms, with fixed $\hat{\mu}$, $\hat{\alpha}$ and $\hat{\sigma}$, we start by generating a dataset of $n = 1,344$ Pareto II realizations, contaminated in a second stage by a cluster of normal distributed outliers. We call outliers the normal realizations because they are not consistent with the null hypothesis of Pareto II distribution. Of course, it is important to maintain paretian and normal observations in separate cells. We now dispose of a second vector of χ^2 statistics, $\boldsymbol{\chi}^2_c$, where c stands for contaminated.

Finally, the forward search is carried over 100,000 sets of $n = 1,344$ pseudo-random realizations from the Pareto II distribution estimated on empirical data. We thus obtain a matrix $\boldsymbol{\chi}^2_s$ sized $136 \times 100,000$ to generate Monte-Carlo confidence intervals for the chi-square statistic. In this way we obtain Monte Carlo envelopes useful to highlight whether possible rejection of the Pareto II hypothesis depends on particular cells or it is diffused throughout the data.

In either case, a graph is drawn to represent chi-square values, both the empirical and the simulated ones, associated to the different steps of the forward search. In particular, $\boldsymbol{\chi}^2_e$ and the envelope given by the 5,25,50,75 and 95-th percentiles of $\boldsymbol{\chi}^2_s$ rows are plotted to get a forward chi-square graphical test (see Fig. 1).

The rationale behind this approach is that when a cell not consistent with the Pareto II hypothesis is added to the estimation subset, the chi-square statistic assessing the goodness-of-fit of the model with data in hand should overcome the boundaries of the Monte Carlo chi-square confidence interval, which are parameter independent.

The left panel in Fig. 1 refers to the original dataset and shows that the χ^2 statistic, when calculated on data largely consistent with the Pareto II distribution, lies inside the confidence envelopes in correspondence of all but 2 cells. The right panel, on the other hand, where the Pareto II realizations contaminated by normal realizations are examined, shows that the null hypothesis is rejected starting from the introduction of a number of cells slightly larger than $r/2$. In this way, the suggested procedure enables us to discriminate between these two cases and, more, to identify the observations deriving from alternative distributions.

[6] If raw data MLE are available, the correct critical points for the Pearson statistic fall between $\chi^2(M - k - 1)$ critical points and those of $\chi^2(M - 1)$, where k is the number of parameters to be estimated [5]. Therefore, in the Pareto II case, 4 is the minimum number of cells required.

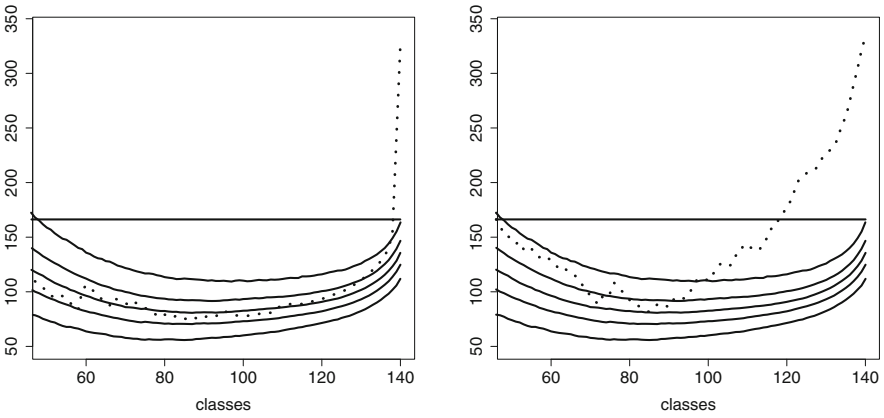


Fig. 1 Forward χ^2 test. The dotted line represents the χ^2 statistic for empirical (left panel)/contaminated (right panel) observations. Solid lines represent the 5,25,50,75 and 95th percentiles of simulated χ^2 statistics. On the x-axis cells are reordered according to their inclusion in the forward search

4 Concluding Remarks

The main contribution of this paper is to develop a robust estimation procedure for the Pareto II distribution through the forward search. To our knowledge, this represents the first attempt to extend the forward search routine to estimation and fitting of a distributional form rather than of parameters in a specified model. Furthermore, this procedure allows us to avoid two common practices in distribution modelling. The first is to reject a given model tested over the whole sample with no preliminary analysis of single firm effect. The second concerns the use of sequential algorithms which, in order to capture the extreme realizations behaviour, could miss an overall awareness of the data.

In the suggested approach the introduction of extreme (influential) observations is signaled by sharp changes in the curves that monitor parameter estimates, or any other statistics at every step. In this context, the robustness of the method does not derive from the choice of a particular estimator with a high breakdown point, but from the progressive inclusion of the observations into a subset which, in the first steps, is outlier free. As a bonus of the suggested procedure, the cells of observations can be naturally ordered according to the specified null model and it is possible to know which and how many of them are compatible with a particular specification. Furthermore, the suggested approach enables us to analyze the inferential effect of the atypical observations (outliers) on the results of statistical analyses. Finally, given that the sequence of subsets during the search are not random subsamples of the data, but contain the cells which give the best fit, we expect that the χ^2 test in the central part of the search will not have a χ^2 distribution. The graphical superimposition of simulated χ^2 statistics indicates when the rejection of the Pareto II hypothesis depends on particular cells.

The choice of modelling Pareto II distribution realizations is motivated here by our previous effort to assess the behaviour of firms belonging to the Italian stock exchange. Further research could extend the forward search routine to robust estimation of the parameters which index other heavy tailed distributions, and in particular could prove useful in studying the distribution of economic growth rates, both on the micro and on the macro side. Detection of heavy-taildness in broadly studied time series data, such as the GDP time series (for both the USA and Italy, see [8]), could widen the study of possible generating processes which lead to non-normally distributed growth rates [4]. Another possible line of development requires a modification of the cell definition in order to minimize the probability of mixing realizations from different distributions in the same cell. Concluding, a systematic comparison of the proposed procedure with sequential methods over different data sources would allow a better understanding of the pro and cons from both perspectives.

References

1. Arnold, B.: Pareto Distributions. International Co-operative Publishing House, Fairland, ME (1983)
2. Atkinson, A.C., Riani, M.: Robust Diagnostic Regression Analysis. Springer, New York, NY (2000)
3. Cabras, S., Morales, J.: Extreme value analysis within a parametric outlier detection framework. Appl. Stoch. Model. Bus. Ind. **23**(2), 157 (2007)
4. Castaldi, C., Dosi, G.: Technical change and economic growth: some lessons from secular patterns and some conjectures on the current impact of ICT. LEM Papers Series 2007/14, Laboratory of Economics and Management (LEM), Sant'Anna School of Advanced Studies, Pisa, Italy (2007). URL http://ideas.repec.org/p/ssa/lemwps/2003-02.html
5. Chernoff, H., Lehmann, E.: The use of maximum likelihood estimates in χ^2 tests for goodness of fit. Ann. Math. Stat. **25**(3), 579–586 (1954)
6. Corbellini, A., Crosato, L., Ganugi, P., Mazzoli, M.: Robust stepwise fitting of the Pareto II distribution: theoretical and computational aspects. In: Skiadas, C.E. (ed.) Advances in Data Analysis, pp. 33–41. Springer-Birkhauser, Boston, MA (2009)
7. Drees, H., Kaufmann, E.: Selecting the optimal sample fraction in univariate extreme value estimation. Stoch. Process. Appl. **75**(2), 149–172 (1998)
8. Fagiolo, G., Napoletano, M., Roventini, A.: Are output growth-rate distributions fat-tailed? Some evidence from OECD countries. J. Appl. Econ. **23**(5) (2008)
9. Kleiber, C., Kotz, S.: Statistical Size Distributions in Economics and Actuarial Sciences. Wiley, New York, NY (2003)
10. Lux, T.: The stable Paretian hypothesis and the frequency of large returns: an examination of major German stocks. Appl. Financ. Econ. **6**(6), 463–475 (1996)
11. Lux, T.: The limiting extremal behaviour of speculative returns: an analysis of intra-daily data from the Frankfurt Stock Exchange. Appl. Financ. Econ. **11**(3), 299–315 (2001)
12. Mann, H., Wald, A.: On the choice of the number of class intervals in the application of the chi square test. Ann. Math. Stat. **13**(3), 306–317 (1942)
13. Pareto, V.: Cours d'Economie Politique. Swisse, Lausanne (1897)
14. Rachev, S.T., Fabozzi, F., Menn, C.: Fat-Tailed and Skewed Asset Return Distributions: Implications for Risk Management, Portfolio Selection, and Option Pricing. Wiley, New York, NY (2005)
15. Reiss, R., Thomas, M.: Statistical Analysis of Extreme Values. Birkhauser Verlag, Basel (1997)

16. Resnick, S.: Modeling Data Networks. Chapman & Hall/CRC, Boca Raton, FL (2003)
17. Robert, C.P., Casella, G.: Monte Carlo Statistical Methods. Springer, New York (2005)
18. Silverberg, G., Verspagen, B.: The size distribution of innovations revisited: An application of extreme value statistics to citation and value measures of patent significance. J. Econ. **139**(2), 318–339 (2007)

Bootstrap and Nonparametric Predictors to Impute Missing Data

Agostino Di Ciaccio

Abstract A new nonparametric technique to impute missing data is proposed in order to obtain a completed data-matrix, capable of producing a degree of reliability for the imputations. Without taking into account strong assumptions, we introduce multiple imputations using bootstrap and nonparametric predictors. It is shown that, in this manner, we can obtain better imputations than with other known methods producing a more reliable completed data-matrix. Using two simulations, we show that the proposed technique can be generalized to consider non-monotone patterns of missing data with interesting results.

1 Introduction

Following Little and Rubin [7] we can distinguish between basically three missing data mechanisms. Data are said to be "missing at random" (MAR) if the mechanism resulting in its omission is independent of its (unobserved) value. If its omission is also independent of the observed values, than the missingness process is said to be "missing completely at random" (MCAR). Finally, if the missingness process depends on the unobserved values, it is said to be "missing not at random" (MNAR). The MCAR data can be handled quite easily while the MNAR data remains hard to analyse. The literature has focused mainly on the MAR mechanism, which can be considered realistic in many situations. In this paper we propose a non-parametric method to analyse MAR data, comparing it to the well-known Multiple Imputation technique using an extensive simulation approach. The proposed method can analyse non-monotone patterns of missing data, i.e. missing values can be observed for any variable and unit. Moreover, it is able to compute a reliable complete data-matrix without strong assumptions on the distribution of data.

A. Di Ciaccio (✉)
Department of Statistics, Probability and Applied Statistics, Sapienza University of Rome, Rome, Italy
e-mail: agostino.diciaccio@uniroma1.it

B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-13312-1_20,

2 Multiple Imputation

The term Multiple Imputation (MI) is usually used to indicate a technique proposed by Rubin [11]. With this technique, several values are imputed for each missing value so that the data-matrix is completed, say, m times. The imputations are obtained, ideally, drawing from the Bayesian posterior predictive distribution, that is, the distribution of the missing data given the observed data with the parameters integrated out using the prior distribution. Indicating by θ the parameter to estimate, we obtain m estimators $\hat{\theta}_j$ and m corresponding variance estimators $\sigma^2_{\hat{\theta}_j}$. Averaging $\hat{\theta}_j$ we obtain $\hat{\theta}$ as the point estimator of θ. Let

(average within imputation variance) (*between-imputation variance)*

$$\bar{\sigma}^2_{\hat{\theta}} = m^{-1} \sum_{j=1}^{m} \sigma^2_{\hat{\theta}_j} \qquad \tilde{\sigma}^2_{\hat{\theta}} = (m-1)^{-1} \sum_{j=1}^{m} \left(\hat{\theta}_j - \hat{\theta}\right)^2 \tag{1}$$

The variance of the estimator $\hat{\theta}$ is estimated by

$$\hat{\sigma}^2_{\hat{\theta}} = \bar{\sigma}^2_{\hat{\theta}} + \left(1 + m^{-1}\right) \tilde{\sigma}^2_{\hat{\theta}} \tag{2}$$

This formula is the famous "Rubin's Rule" [11]. It is usually assumed that

$$\hat{\sigma}^{-1}_{\hat{\theta}} \left(\hat{\theta} - \theta\right) \sim t_v \tag{3}$$

which allows us to calculate confidence intervals and tests [11].

The validity of this approach depends on some assumptions. The imputations have to be drawn from what Rubin calls a *proper* multiple imputation procedure: if the multiple imputations are proper, then $\hat{\theta}$ is a consistent, asymptotically normal estimator and $\hat{\sigma}^2_{\hat{\theta}}$, given by (2), is a weakly unbiased estimator of its asymptotic variance in sufficiently regular models. This technique shows good properties with multinormal data in the presence of a large sample, but in many cases the MI assumptions will not be completely satisfied and so we should be cautious in trusting the obtained confidence intervals and tests [9]. Moreover, MI does not give us a completed data-matrix, which is convenient to have in many cases. Other problems for MI are:

- Difficulties in analysing a large number of variables
- Difficulties in analysing mixed measurement level data
- Multinormality may be non-realistic
- It cannot consider constraints on the imputations
- It cannot consider bounds or complex survey designs

However, single imputations cannot reflect the uncertainty for the predictions of the unknown missing values and consequently the variances of the parameter estimates will be biased downward.

Moreover, MI is capable to analyse non-monotone patterns of missing data, i.e. the most common situation in which missing values are observed for several variables without any kind of order or pattern. If we can assume multivariate normality, having a monotone missing pattern, we can apply parametric methods such as the Regression method [11, 12] or the predictive mean matching method [13]. Having a non-monotone pattern of missing data, we should consider a more complex approach which requires an iterative procedure. Obtaining a complete matrix is not the aim of MI, but by calculating an adequate number of complete matrices (>20), we can obtain an estimation of each missing value by the mean of the imputed values.

As MI needs the posterior predictive distribution, we are required to know the multivariate distribution of data, which is usually unknown and often far from a multivariate Normal (e.g. when there are mixed measurement level variables). It has been noted that parametric assumptions, such as multivariate normality, seem to be much more sensitive in missing-data problems [5]. To overcome this difficulty, recent extensions of the MI approach consider mixtures of Gaussian distributions [17, 4].

On the other hand, for Non-Normal data, non-parametric approaches appear to be more appropriate in estimating a complete data-matrix and in assessing the accuracy of an estimator in a missing data situation.

3 Bootstrap and Missing Data Imputation

Given a sample x, a population parameter θ and an unbiased estimator $\hat{\theta}$, fixing randomly a number of missing data in the sample, we define $\hat{\theta}_{imp}$ the estimator of θ based on the completed data, where missing have been imputed by some procedure.

The bias of $\hat{\theta}_{imp}$ is

$$E_{x,m}\left(\hat{\theta}_{imp}-\theta\right)=E_x\left(\hat{\theta}-\theta\right)+E_{x,m}\left(\hat{\theta}_{imp}-\hat{\theta}\right)=E_{x,m}\left(\hat{\theta}_{imp}-\hat{\theta}\right) \quad (4)$$

where $E_{x,m}(.)$ indicate the expectation with respect to the sample ($\mathbf{x}$) and the missing data (m) draw. Considering the MSE of $\hat{\theta}_{imp}$, from expression (4), after some steps, we obtain:

$$E_{x,m}\left(\hat{\theta}_{imp}-\theta\right)^2\approx E_x\left(\hat{\theta}-\theta\right)^2+E_{x,m}\left(\hat{\theta}_{imp}-\bar{\theta}_{imp}\right)^2+E_x\left(\bar{\theta}_{imp}-\hat{\theta}\right)^2 \quad (5)$$

where we have indicated with $\bar{\theta}_{imp}$ the mean of $\hat{\theta}_{imp}$.

The first term on the right hand is independent of the missing data. The second and third terms on the right hand are, respectively, the variance of $\hat{\theta}_{imp}$ and its

squared bias. If we impute all missing values with the true values, these two terms are zero. In general we should choose an imputation method which produces low bias and variance terms. Instead of obtaining multiple imputations through a multivariate Bayesian predictive distribution, we could consider other ways of obtaining several imputations for each missing value. For example, Rubin [11] considered an alternative to MI based on Nearest Neighbours, using only two imputed data-matrices. Van Buuren et al. [16] proposed the Multivariate Imputation by Chained Equations (MICE) which requires specifying a conditional distribution for the missing data in each incomplete variable. Under the assumption that a corresponding multivariate distribution exists, MICE constructs a Gibbs sampler to generate multiple imputations. This approach requires us to specify only the conditional distributions and then iterates over all conditionally specified imputation models. This implies several advantages with respect to MI: the univariate problems are simpler than multivariate ones and it is possible to consider mixed measurement variables, bounds, constraints between variables, interactions and so on. A similar approach is used by Sequential Regression for Multiple Imputations, [10], implemented by IVEware software. In both metods the estimation algorithm proceeds iteratively:

Step 1: select the variable with smallest amount of missing values then use an appropriate regressive model (linear regression, logistic regression, Poisson loglinear) applied to the complete data to obtain an estimation of missing values. The variable is then considered complete and the process continues with the next variable.

Step 2: each variable with missing data is regressed using all the other variables with completed data. Each regression updates the imputed values of the variable including a random noise. This process continue for a predeterminated number of rounds.

The main problem with these methods is that they may not converge to a distribution if the separate models are not compatible with a multivariate distribution, though Van Buuren et al. [15] showed, by simulation studies, that reasonable imputations were obtained even when the separate models were incompatible. Other interesting proposals include the use of decision trees [8, 2]. This approach has the advantage of being able to analyse variables with different measurement levels using a single model.

Here we propose a method based on regression trees and Bootstrap Aggregating [1], considering a data-matrix with missing data for one or more variables (qualitative or quantitative). We are mainly interested in obtaining an estimated complete data-matrix, but our proposal also proved useful in making inferences on population parameters.

The use of Bootstrap for missing data problems is not new [5, 6, 14]. The use of Bagging with non-parametric predictors was recently considered in Di Ciaccio and Vallely [3].

Having just a single quantitative variable with missing data, we can consider all the other variables as covariates and, by using an adequate prediction model estimated on the complete data, we can estimate the missing values. This procedure can be repeated several times on different bootstrap samples obtaining an empirical

distribution for each missing value. If we are using a Regression Tree model, we know from the literature that aggregating these predicted values (usually by the mean) we can obtain a reliable estimation of the missing values. Moreover, using the obtained empirical distributions, we can calculate confidence intervals for a parameter of the population. Di Ciaccio and Vallely [3] used this nonparametric approach to impute missing data for only one quantitative variable Y. Considering a non-linear relationship between Y and the other variables, they showed that an approach based on Regression Trees with Bagging gives better results than MI. In particular, it was shown that RTB had, in most cases, lower variability than MI due to missingness. Moreover, they compared RTB and MI with respect to the width and coverage of estimated confidence intervals of the Y mean, obtaining good performance for RTB. In this paper we propose an extension of that method to analyse non-monotone patterns of missing data for two or more variables.

4 Bagged Trees to Predict Missing Data

Using an approach similar to MICE or IVEware, but considering a non-parametric approach, we could initialize, for example by the mean, the missing values in the data-matrix, and then, iteratively until convergence, consider each variable with missing data as the target variable. This method is illustrated in Table 1. The index we used to evaluate the convergence of the procedure is:

$$\delta = \frac{1}{n_{mis}} \sum_{j=1}^{J} \frac{1}{\sigma_j^2} \sum_{i \in M_j} \left(\hat{x}_{ij}^{t} - \hat{x}_{ij}^{t-1} \right)^2 \tag{6}$$

where n_{mis} is the total number of missing values, σ_j^2 the variance of the j-th variable, M_j the subset of indices in which the j-th variable is missing. The iterative process stops anyway when the maximum number of iterations T is achieved.

Indicate by $\hat{x}_{ij}^{(t)}$ the estimated value, obtained by the model at step t, of the missing value x_{ij}. In the update step 8 use the formula:

$$\hat{x}_{ij}^{t} = \gamma \hat{x}_{ij}^{t-1} + (1 - \gamma) \hat{x}_{ij}^{(t)} \tag{7}$$

with $0 < \gamma < 1$. This procedure does not introduce explicitly a noise, as in the Chained Equations approach. Instead, it tries to achieve the best estimation of the missing values. To evaluate the proposed algorithm we considered two simulations which analyse two opposite cases:

1. We generated 300 random populations from multinormal distributions with 5 independent quantitative variables.
2. We generated 300 random populations from non-normal data with 5 not independent quantitative variables

From each population we extracted one sample of 120 units, then we randomly drew 30% of the missing values on the first two variables $X_1$and X_2 giving a larger probability to units having X_3 higher than the median. The predictor model used in step 6, is a Regression Tree with Bagging (RTB), with 20 bootstrap replications, which allows us to calculate confidence intervals for the parameters of interest [3]. In particular, we fixed T=50 and $\gamma = 0.9$. For a comparison, we applied MI with 25 imputations to the same data and also the simplest univariate approach which imputes the mean computed on complete data (MEAN).

We evaluated:

1. The difference between the imputed and the real data values, by the weighted sum of square:

$$SSEW = \frac{1}{n_{mis}} \sum_{j=1}^{J} \frac{1}{\sigma_j^2} \sum_{i \in M_j} \left(\hat{x}_{ij} - x_{ij}\right)^2 \tag{8}$$

   This index measures the reliability of the final completed data matrix.
2. The difference between the true mean and the mean obtained on imputed data, for the variables $X_1$and X_2.
3. The difference between correlation coefficients on the population and on imputed samples (in particular, we compared r_{13}; r_{14}; r_{15}).

From the results of the simulations, shown in Table 1 and Table 2, we can deduce that having independent variables, also in the case of multinormal data, both MI and RTB do not work better than MEAN, which shows a good performance. Indeed, both RTB and MI iteratively try to emphasize the relation between each variable and the others, while the MEAN procedure is coherent with the independence of the vari-

Table 1 The algorithm

1. For each variable: initialize missing data with the mean
2. Iterate ($t \leq T$)
3. Iterate (j=1 to num. of variables)
4. Set the variable j as the target variable
5. Select cases which do not have missing value for variable j
6. Estimate the predictor model
7. Estimate missing values of variable j by the predictor model
8. Update missing values of variable j. Go to step (3).
9. If $\delta < 0.001$ or $t = T$ then STOP else go to step (2).

Table 2 Results of imputation on multinormal data with independent variables: number of samples favourable to each method

	RTB	MI	Mean
SSEW	37	0	263
Difference on the mean X1	91	93	116
Difference on correlation coefficients	30	0	270

Table 3 Results of imputation on non-normal data with non-independent variables: number of samples favourable to each method

	RTB	MI	Mean
SSEW	266	34	0
Difference on the mean X1	93	146	61
Difference on correlation coefficients	138	104	58

ables. In the opposite case, with non-normal non-independent data (Table 3), RTB appears preferable: it is definitely better when considering SSEW, i.e. the difference between real and imputed data, and is also better when considering the estimated and the true correlation coefficients. MI appears better than RTB only for the estimation of the $X1$ mean. This last result can be explained looking at (5). If $\hat{\theta}_{imp}$ is the sample mean, we do not need the estimates of missing values to be close to the corresponding true values. In fact, exchanging the missing estimates inside of each variable we would obtain the same estimate of $\hat{\theta}_{imp}$. Moreover, distribution of $\hat{\theta}_{imp}$ will be close to a Normal and the hypotheses of MI allow to make variance of $\hat{\theta}_{imp}$ smaller. In Fig. 1 we show the values of SSEW for the simulation with multinormal data and independent variables (each point is a distinct population). The values for RTB and MI are depicted on the graph, while the values of MEAN constitutes the abscissa. Points above the bisector have an SSEW larger then MEAN: we can see that MI has the SSEW values higher than MEAN and also than RTB. Concluding, the results of

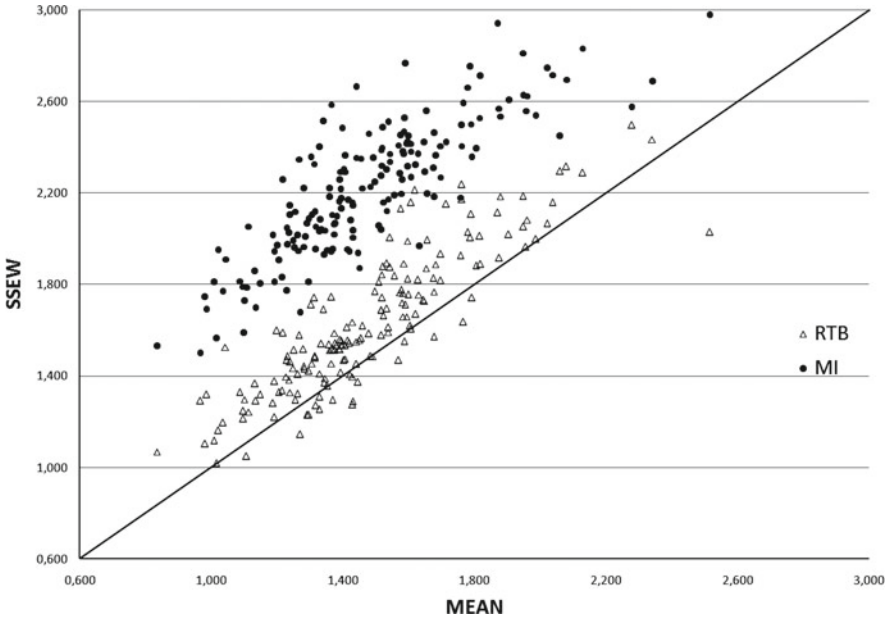


Fig. 1 Multinormal data with independent variables: SSEW for RTB, MI with respect to MEAN

the two simulations suggest that before choosing the imputation method, we should investigate the interdependence of the variables and evaluate the assumptions of multinormality. It is also evident the good performance obtained by the proposed method.

References

1. Breiman, L.: Bagging predictors. Mach. Learn. **24**, 123–140 (1996)
2. Conversano, C., Siciliano, R.: Incremental Tree-based missing data imputation with lexicographic ordering. In: Minotte, M., Swzychak, A. (eds.) Interface 2003 Proceedings, Interface Foundation of North America, Washington, DC (2003)
3. Di Ciaccio, A., Vallely, T.: Use of non-parametric methods for the imputation of missing data. A comparison based on extensive Montecarlo simulations. In: S.Co.2007, Venice. http://venus.unive.it/sco2007/ocs/papers.php (2007)
4. Di Zio, M., Guarnera, U., Luzi, O.: Imputation through finite Gaussian mixture models. Comput. Stat. Data Anal. **51**, 5305–5316 (2007)
5. Efron, B.: Missing data, imputation, and the bootstrap. J. Am. Stat. Assoc. **89**(426), 463–475 (1994)
6. Fay, E.R.: Alternative paradigms for the analysis of imputed survey data. J. Am. Stat. Assoc. **91**(434), 490–498 (1996)
7. Little, R., Rubin, D.: Statistical Analysis with Missing Data. Wiley, New York, NY (1987)
8. Mesa, D., Tsai, P., Chambers, R.L.: Using tree-based models for missing data imputation: an evaluation using UK census data. Research Note, Department of Social Statistics, University of Southampton, London (2000)
9. Nielsen, S.F.: Proper and improper multiple imputation. Intern. Stat. Rev. **71**(3), 593–607 (2003)
10. Raghunathan, T.E., Lepkowski, J.M., Van Hoewyk, J., Solenberger, P.: A multivariate technique for multiply imputing missing values using sequence of regression models. Surv. Methodol. **27**(1), 85–95 (2001)
11. Rubin, D.B.: Multiple Imputation for Nonresponse in Surveys. Wiley, New York, NY (1987)
12. Schafer, J.L., Schenker, N.: Inference with imputed conditional means. J. Am. Stat. Assoc. **95**(449), 144–154 (2000)
13. Schenker, N., Taylor, J.M.G.: Partially parametric techniques for multiple imputation. Comput. Stat. Data Anal. **22**, 425–446 (1996)
14. Shao J., Sitter R.R.: Bootstrap for imputed survey data. J. Am. Stat. Assoc. **91**(435), 1278–1288 (1996)
15. Van Buuren, S., Brand, J.P.L., Groothuis-Oudshoorn, C.G.M.: Fully conditional specification in multivariate imputation. J. Stat. Comput. Simul. **76**(12), 1049–1064 (2006)
16. Van Buuren, S., Oudshoorn, C.G.M.: Multivariate imputation by chained equations: MICE V1.0 User's manual. Report G/VGZ/00.038. Leiden, TNO Preventie en Gezondheid (2000)
17. Zhang, J., Everson, R.: Bayesian estimation and classification with incomplete data using mixture models. Proceedings of the 2004 International Conference on Machine Learning and Applications, Louisville, KY, USA, pp. 296–303 (2004)

On the Use of Boosting Procedures to Predict the Risk of Default

Giovanna Menardi, Federico Tedeschi and Nicola Torelli

Abstract Statistical models have been widely applied with the aim of evaluating the risk of default of enterprises. However, a typical problem is that the occurrence of the default event is rare, and this class imbalance strongly affects the performance of traditional classifiers. Boosting is a general class of methods which iteratively enforces the accuracy of any weak learner, but it suffers from some drawbacks in presence of unbalanced classes. Performance of standard boosting procedures to deal with unbalanced classes is discussed and a new algorithm is proposed.

1 Introduction

Credit risk models and methods aim at finding rules to measure the risk associated with credit applications or to separate defaulter credit applicants from non-defaulter ones. The decision about giving credit to an applicant may determine profits or costs to the lenders according to how "good" or "bad" his/her subsequent behavior will be. Moreover the recent supervisory regulation of the New Basel Capital Accord [2] imposes the minimum consistency of capital required to internationally active banks as proportional to the credit risk and determines the prerequisites for an internal rating based approach. Since this may determine lower capital requirements in comparison with external rating information, it has strongly affected the internal banking processes of modeling and measuring the credit risk. These reasons make clear the need of developing accurate models and methods to help lenders in their decision.

The statistical approach to this problem is mainly based on classification algorithms (see, e.g., [12]) which aim at finding what would have been the best rule to apply on a sample of previous applicants. The advantage of these procedures is that the subsequent behavior of these applicants is known. Our interest is on methods for separating defaulter enterprises from non-defaulter ones. In this context, a typical problem occurs because the proportion of defaulter firms is very close to zero,

G. Menardi (✉)
Department of Economics and Statistics, University of Trieste, Trieste, Italy
e-mail: giovanna.menardi@econ.units.it

B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-13312-1_21,

leading to a strong imbalance between the two classes. As a consequence the performance of the estimated models might be significantly affected since the classifiers learn from the most prevalent class and tend to ignore the rare examples [6].

In this work we evaluate the opportunity of using boosting methods to predict the default event. The potential of boosting techniques is often limited when applied to imbalanced data sets. In this paper an adjustment to prevent these drawbacks is proposed and applied to some real data sets.

2 Boosting Overview

Boosting refers to a general class of iterative methods to produce accurate prediction rules by combining any weak classifier. In its most popular formulation called AdaBoost (see, e.g., [10]), it takes as an argument a training sample $\mathcal{T} = (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$ where each $\mathbf{x}_i$ belongs to some domain $\mathcal{X} \in \mathbb{R}^p$, and each label y_i varies in $\mathcal{Y} = \{-1, 1\}$. At each iteration t a distribution of weights D_t is set over the training set and a weak classifier is built on the training set according to D_t. The algorithm starts by setting all weights at the same value, but at any successive round, the relative weights of misclassified examples are increased. The procedure stops after a fixed number of iterations. All the weak classifiers contribute to the prediction of the trained and new unlabeled examples in proportion to their accuracy. The pseudo code is reported in Fig. 1.

The algorithm is designed to maintain a distribution of weights over the training set in order to force the weak learner to focus on the hardest to classify examples. The weight update factor is based on the accuracy of the weak learner (α_t is a decreasing function of ε_t) and it is chosen in order to minimize an upper bound to the misclassification error rate on the training set. This bound decreases when any of the weak learner is improved. This way to *adapt* to the error rate of the weak classifiers makes the basis for the name of the algorithm "AdaBoost".

It has been observed that the weak learner error ε_t tends to increase with the number of iterations. [11] admit that it is an increasing function of t, "possibly

Given: $(x_1; y_1), \ldots, (x_m; y_m)$, $x_i \in X$, $y_i \in Y = \{-1, 1\}$.
Initialize $D_1(i) = 1/m$. For $t = 1 \ldots T$:

1. Train weak classifier using distribution D_t.
2. Get weak hypothesis $h_t : X \to \{-1, 1\}$ with error $\varepsilon_t = \sum_{i:h_t(x_i) \neq y_i} D_t(x_i)$
3. Choose $\alpha_t = \frac{1}{2} \log(\frac{1-\varepsilon_t}{\varepsilon_t})$
4. Update:
$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \cdot = \begin{cases} e^{-\alpha_t} & \text{if instance } i \text{ is correctly classified} \\ e^{\alpha_t} & \text{if instance } i \text{ is } \underline{\text{not}} \text{ correctly classified} \end{cases}$$

where Z_t is a normalization factor (chosen so that $\sum_{i=1}^{m} D_{t+1} = 1$).

Output the final hypothesis: $\mathscr{H}(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$.

Fig. 1 Pseudo code for the boosting algorithm AdaBoost

even converging to 1/2". They state that "characterizing the conditions under which the increase is slow is an open problem" and observe that, when it increases too quickly, Adaboost may perform poorly. Being γ_t the accuracy of the weak learner t ($\gamma_t = 1/2 - \varepsilon_t$), an upper bound to the training error is given by: $exp(-2\sum_{t=1}^{T}\gamma_t^2)$. This implies that convergence to 0 of the training error depends on the behavior of ε_t. However, not only the training accuracy of a weak learner is increased by boosting algorithms, but also a characterization of the generalization error is possible. This is given in [11] in terms of *margin*, defined as proportional to $y\sum_{t=1}^{T}\alpha_t h_t(\mathbf{x})$. The margin takes values in $[-1, 1]$, being positive if and only if the example is correctly classified by $\mathcal{H}$ and its magnitude (considered as a measure of confidence in the prediction) increases with the number of iterations. Moreover, larger margins in the training set translate in a lower upper bound for the test set error. Since this bound is entirely independent from the number of boosting iterations T, this might be interpreted as a capability of boosting to avoid overfitting. Hence, the better performance on the training set translates in a smaller generalization error rate.

2.1 Boosting in Presence of Unbalanced Classes

Three main classes of boosting algorithms have been applied to unbalanced class problems:

1. The algorithms of the first category, to which AdaBoost belongs, treat all correctly and incorrectly classified examples equally by increasing the weights of false positive or false negative examples in the same proportions. Prediction of new unlabeled examples is an average of the single classifications, weighted according to the overall accuracy of the single classifiers.
2. A more suitable category of boosting algorithms in presence of rare classes is the one to which AdaCost belongs [4], where a different cost of misclassification is given to the training examples. In a class imbalance framework it corresponds to give a higher misclassification cost to the rare class examples.
3. The third family of boosting procedures is expressly conceived to take into account rare classes. RareBoost [7] is the most significant representative of this family. Here, the weights on the training examples are increased (decreased) by giving a different treatment to false (true) positive and negative predictions. The final classification considers both the positive and the negative accuracy of the single classifiers. However, RareBoost is based on the assumption that the weighted true positive rate is greater than the weighted false positive rate. In the presence of a class imbalance problem such a constraint is rarely satisfied, being the small class associated with both poor recall and precision values.

In this work we have addressed the issue of analyzing the behavior of AdaBoost. Since Adaboost increases the weight of misclassified observations, when it is run on unbalanced training sets it has the interesting feature of raising the average weight of the instances of the rare class in the first iterations. This process goes on as

long as the accuracy of the weak learners is affected mostly from their performance on the majority class. It is: $\sum_{y_i=1} D_t(i) < \frac{1}{2} \Rightarrow \sum_{y_i=1} D_{t+1}(i) > \sum_{y_i=1} D_t(i)$. This implies that, in a given number of iterations, the weak learner places essentially half of the global weight to the units of the rare class.

Despite this suitable behavior, our experience suggests that two drawbacks have to be taken in consideration when applying boosting to an imbalance class problem:

1. Since each weak learner is evaluated basing on its accuracy, and since (as seen above) the error ε_t is an increasing function of t, the final classifier will be mostly affected by the first weak learners.
2. For large training sets, the skewer the distribution of the classes is, the quicker ε_t increases and in a few iterations it converges to $1/2$, leading the algorithm to a standstill.

While the first issue leads to a better performance on the frequent class than in the rare one (and some adjustments have been proposed to address this problem), the second point may affect the classifier ability to discriminate between the two classes. Basically, fewer iterations are used in order to build the final classifier and therefore, the feature of boosting to improve on single classifiers is weakened.

Moreover, given the resistance of boosting to overfit, the performance loss observed in case of class imbalance also translates in a higher generalization error.

3 The ROSEBoost Approach for Dealing with Class Imbalance

Although AdaBoost performs quite accurately even in presence of rare classes, a modification aimed at avoiding the ε_t convergence to $1/2$ would help us to make the most of boosting. A possible solution would consist in including a stochastic component in the algorithm, for allowing it to run for the desired number of iterations without coming to the standstill and possibly going on in learning from the errors.

In our previous work [9], we addressed the issue of classification in presence of rare classes and propose an effective strategy to balance the distribution of the labels, based on the smoothed bootstrap generation of synthetic examples. This strategy, called ROSE (Random Over-Sampling Examples) is here used in conjunction with AdaBoost in order to jitter the data and possibly avoid the convergence of ε_t to $1/2$.

ROSE produces an augmented sample of data (especially belonging to the rare class) by simulating new examples from an estimate of the conditional density f of the two classes. The procedure for generating one new example from the class $y \in \mathcal{Y}$ consists in two steps:

1. select $\mathbf{x}_i \in \{(\mathbf{x}_j, y_j) : y_j = y\}$ with probability $P(\mathbf{x}_i)$
2. sample $\mathbf{x}$ from $K_{H_y}(\mathbf{x}_i)$, with K_{H_y} a probability distribution centered at $\mathbf{x}_i$ and H_y a matrix of scale parameters. $K_{H_y}(\mathbf{x}_i)$ is usually chosen in the set of the symmetric distributions (e.g. K is a Gaussian distribution) and it is an estimate of the local density of $\mathbf{x}_i$.

Essentially, ROSE selects observed data belonging to the class y and generates new examples in its neighborhood, where the width of the neighborhood is determined by H_y. It is worthwhile to note that:

$$\hat{f}(\mathbf{x}|y) = \sum_{i:y_i=y} Pr(\mathbf{x}_i) Pr(x|\mathbf{x}_i) = \sum_{i:y_i=y} Pr(\mathbf{x}_i) K_{H_y}(\mathbf{x} - \mathbf{x}_i).$$

Hence, if the uniform distribution is set over the $\mathbf{x}_i$ ROSE generates new examples from the kernel density estimate of $f(\mathbf{x}|y)$, $y \in \mathcal{Y}$, while non-uniform $P(\mathbf{x}_i)$ correspond to the generation of new data from a weighted kernel estimate of the conditional densities of the class y. Instead, the selection of $\mathbf{x}_i$ from $\mathbf{x}_1, \ldots, \mathbf{x}_m$ in the first step of ROSE would entail the generation of new data from $\hat{f}(\mathbf{x})$.

This idea, combined with AdaBoost, gives rise to ROSEBoost, which may be thought of as a stochastic version of AdaBoost (see the pseudo code in Fig. 2). Both the algorithms start with an input training set $\mathcal{T}$ and an uniform distribution of weights D_t, and at each successive iteration a weak classifier is built based on the distribution of weights. The update mechanism of D_t is such that the weights of the misclassified examples is increased while the weights of the well labeled examples is decreased. Moreover, in both algorithms each weak learner contributes to the combined classifier according to its accuracy.

However, while AdaBoost estimates the classifiers on different weighted versions of the observed data, ROSEBoost uses, at each iteration, a new training set $\mathcal{T}^*$, generated according to ROSE, from the weighted kernel estimate of the density underlying the observed data. The key feature of ROSEBoost is that the training set $\mathcal{T}^*$ used at each round is generated by giving to each $(\mathbf{x}_i, y_i) \in \mathcal{T}$ a probability of being selected in the first step of ROSE which is proportional to its weight. In this way ROSE generates new data mainly from the hardest to classify examples.

Given: N, $\mathcal{T} = (x_1; y_1), \ldots, (x_m; y_m)$, $x_i \in X$, $y_i \in Y = \{-1, 1\}$.
Initialize $D_1(i) = 1/m$. For $t = 1 \ldots T$:

1. draw a sample $\mathcal{T}^* = (x_1^*; y_1^*), \ldots, (x_N^*; y_N^*)$ from $\hat{f}(x) = \sum_{i:y_i=+1} D_t(x_i) K_{H_+}(x - x_i) + \sum_{i:y_i=-1} D_t(x_i) K_{H_-}(x - x_i)$
2. Train a weak learner on $\mathcal{T}^*$
3. Get weak hypothesis h_t with error $\varepsilon_t = \sum_{i=1}^{N} I[h_t(x*_i) \neq y*_i]$
4. Choose $\alpha_t = \frac{1}{2} \log(\frac{1-\varepsilon_t}{\varepsilon_t})$
5. Get a prediction on $\mathcal{T}$ by using h_t
6. Update
$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \cdot = \begin{cases} e^{-\alpha_t} & \text{if instance } i \text{ is correctly classified} \\ e^{\alpha_t} & \text{if instance } i \text{ is } \underline{\text{not}} \text{ correctly classified} \end{cases}$$
where Z_t is a normalization factor (chosen so that $\sum_{i=1}^{m} D_{t+1} = 1$).

Output the final hypothesis: $\mathcal{H}(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$.

Fig. 2 Pseudocode for the boosting algorithm ROSEBoost

The idea of including a random element in boosting to generate new data has been already put forward by [8] and [3]. In [8] the use of AdaBoost in conjunction with Over/Under-Sampling and Jittering of the data (JOUS-Boost) is proposed with the aim of classification with unequal costs or, equivalently, at quantiles other than 1/2. The authors perform oversampling of the rare (or higher misclassification cost) class and add independent uniform noise to the training data for which over-sampling creates replications. [3] propose to include in a boosting algorithm their SMOTE procedure of generating new sinthetic examples. The new examples are generated according to the characteristics of the $k-$nearest neighbors of the rare events.

ROSE and its boosting version ROSEBoost extend this idea by placing it on a sounder theoretical basis which finds its justification in kernel methods. Moreover, their flexibility allows to include many of the existing procedures (as traditional oversampling and JousBoost) as a special case and hence provides an unified framework to data augmentation methods aimed at facing the class imbalance problem.

4 Some Real Data Applications

We have applied the methods discussed in the previous sections to some simulated and real data sets to compare their accuracy in a class imbalance framework. For brevity, the results deriving from three real data applications only are reported.

The first considered data set has been built by merging data from the Infocamere archive and the Business Register, with the aim of discriminating the defaulter and non defaulter firms. It consists of vital statistics (e.g. changes of the legal status, occurrence of a corporate merger or breakup, number of employees), balance sheet items and financial ratios of all the commercial companies located in a North Eastern province of Italy. The occurrence of a bankruptcy condition is considered as the default event. This data set is a notable example of classification in presence of rare classes, amounting the proportion of defaulter firms to less than the 7‰.

The second data set, known as the "German credit data", has been taken from the UCI Machine Learning Repository [1]. In its numerical formulation it includes 24 attributes observed on a set of 1,000 customers to be classified as good or bad credit risks. Since the original data set shows a balanced distribution of the two classes of customers, 500 examples have been drawn from the provided data by giving to the bad customers a probability of being selected amounting to the 2%. The remaining data have been used for testing the accuracy of the classification.

The third data set has been also drawn from the UCI Machine Learning Repository. It is known as the "Breast Cancer Wisconsin (Prognostic) Data Set" and represents 32 follow-up attributes observed on 196 breast cancer cases. The response variable is the possible recurrence of the cancer. A training set amounting to the 50% of the observed cases has been selected by giving to the positive examples a probability of selection amounting to the 3%, in order to get an unbalanced distribution of the responses. The remaining cases have been used as testing examples.

Four classification strategies have been run on the described data sets: a classification tree, AdaBoost, AdaCost and ROSEBoost. The three boosting algorithms

Table 1 AUC values of the classification procedures on test sets drawn from the Infocamere archive, German credit data and Breast cancer data respectively. The AUC values are averages obtained by running the classifiers on 100 randomly selected samples

	Defaulter firms data	German credit data	Breast cancer data
Classification tree	0.626	0.573	0.539
AdaBoost	0.766	0.656	0.568
AdaCost	0.759	0.623	0.556
ROSEBoost	0.815	0.663	0.613

have been trained by combining stumps for 300 iterations. AdaCost has been performed by giving to the rare examples a misclassification cost ten times higher than the negative examples. In order to choose the parameters H_+ and H_- in RoseBoost, the asymptotically optimal smoothing matrix for the normal distribution has been used for both the classes. The performance of the classifiers has been evaluated by measuring the area under the ROC curve [5] on a test set (AUC). Results are in Table 1. The classification trees perform just slightly better than the random choice. The increased accuracy of AdaBoost is notable, but even if the improvements are not remarkable in all the considered data sets, the application of ROSEBoost results in larger areas under the ROC curves. AdaCost performs better than classification trees but cannot even compete with AdaBoost.

5 Discussion and Concluding Remarks

In this work we have provided a deeper insight to the boosting behavior in dealing with rare classes. It has arisen that, even if in the last few years several boosting algorithms have focused this issue, there is still the lack of an univocally accepted approach to the problem. The boosting algorithms specifically conceived to perform classification in presence of rare events are based on some strong assumptions rarely satisfied, while methods taking account for higher misclassification costs for the rare examples turn out not to outperform the standard AdaBoost. Moreover, it has been shown that, although AdaBoost usually can improve the accuracy of a weak learner even in presence of rare classes, in such situations it suffers from some drawbacks that prevent it from making the most. For these reasons we have proposed ROSEBoost, a new boosting algorithm based on the idea of generating new artificial cases from the local density of the observed data. It has been shown that this idea has a sounder theoretical justification because it corresponds to the simulation of synthetic training examples from the kernel density estimate of the data. It should be noticed that we do not perform the classification by using density estimation (which would entail a further complication in our task) but we exploit the good properties of the kernel methods to enlarge the rare class (or, in general the hardest to classify examples) by sampling the observed data without producing ties in the training set.

A deeper investigation about the properties of ROSEBoost is necessary, and the choice of the optimal smoothing matrices H_y of the kernel density estimates has to be addressed in the context of data generation. In the described applications the asymptotically optimal smoothing matrix for the normal distribution, has been chosen as a rule of thumb, conditionally to the class label. More sophisticated methods for selecting the smoothing matrices can be used, by addressing the choice for our problem: this would entail, for instance, the use of some optimality criterion at each iteration of the boosting process. However, ROSEBoost turns out to be not only a good competitor of AdaBoost in classification problems with unbalanced data, but the results deriving from the described applications have shown that it tends to overperform the standard boosting methods. Moreover the generation of unobserved data might help in increasing the ability of generalization of the classifier.

References

1. Asuncion, A., Newman, D.J.: UCI Machine Learning Repository. Irvine, CA: University of California, http://www.ics.uci.edu/mlearn/MLRepository.html (2007)
2. Basel Committee on Banking Supervision: International Convergence of Capital Measurement and Capital Standards, a Revised Framework, Bank for International Settlements. (2004). Updated in November 2005, www.bis.org/bcbs/
3. Chawla, N.V., Lazarevic, A., Hall, L.O., Bowyer, K.W.: Smoteboost: improving prediction of the minority class in boosting. In Proceedings of the Seventh European Conference on Principles and Practice of Knowledge Discovery in Databases, pp. 107–119 (2003)
4. Fan, W., Stolfo, S.J., Zhang, J., Chan, P.K.: AdaCost: Misclassification cost-sensitive boosting. In: Machine Learning: Proceedings of the 16th International Conference, Bled, Slovenia (1999)
5. Fawcett, T.: An introduction to ROC analysis. Pattern Recognit. Lett. **27**, 861–874 (2006)
6. Japkowicz, N., Stephen, S.: The class imbalance problem: a systematic study. Intell. Data Anal. J. **6**. (2002)
7. Joshi, M.V., Kumar, V., Agarwal, R.C.: Evaluating boosting algorithms to classify rare classes: comparison and improvements. Technical Report. RC-22147, IBM Research Division (2001)
8. Mease, D. Wyner, A.J., Buja, A.: Boosted classification trees and class probability/quantile estimation. J. Mach. Learn. Res. **8**, 409–439 (2007)
9. Menardi, G., Statistical issues emerging in modeling unbalanced data sets. In Proceedings of the 16th European Young Statisticians Meetings, Bucharest, Romania (2009)
10. Schapire, R.E., Freund Y., Bartlett, P., Sun Lee, W.: Boosting the margin: a new explanation for the effectiveness of voting methods. Ann. Stat. **26**(5), 1651–1686 (1998)
11. Shapire, R.E.: The boosting approach to machine learning: an overview. In: Denison, D.D., Hansen, M.H., Holmes, C., Mallick, B., Yu, B. (eds.) Nonlinear Estimation and Classification, Springer, Heidelberg (2003)
12. Thomas, L.C., Edelman, D.B., Crook, J.N.(eds.): Credit Scoring and Its Applications. Siam, Philadelphia, PA (2002)

Part V
Categorical Data and Latent Class Approach

Assessing Similarity of Rating Distributions by Kullback-Leibler Divergence

Marcella Corduas

Abstract A mixture model for ordinal data modelling (denoted CUB) has been recently proposed in literature. Specifically, ordinal data are represented by means of a discrete random variable which is a mixture of a Uniform and shifted Binomial random variables. This article proposes a testing procedure based on the Kullback-Leibler divergence in order to compare CUB models and detect similarities in the structure of judgements that raters express on set of items.

1 Introduction

There are many research areas where interest is in the measurement of perceived attributes of a given object or phenomenon. This happens, for instance, in medicine when perceived chronic pain levels are analyzed, in business economics when customers satisfaction is considered or in psychology and sociology for the analysis of human behaviours. The best known statistical models for describing preferences, ratings or, in general, ordinal data have been developed using the Generalized Linear Models approach (see, amongst others, [1, 11, 12]).

Alternatively, a statistical model, namely CUB, was proposed by D'Elia and Piccolo [5], Piccolo [14], in order to describe the probability distribution of the random variable generating the observed ordinal data. The model arises from a conceptual description of the psychological mechanism running the individual's choices in a rating process. Specifically, two components of this process are identified. The first one relates to the uncertainty that each judge conveys to his/her final judgement when his/her opinion has to be summarized by means of a grading scale. As a matter of fact, extreme feelings of liking/disliking towards the item object of evaluation probably originate sharper opinions and, then, less uncertainty in the selection of the corresponding extreme scores, whereas, fuzzy opinions are likely to originate intermediate scores which are selected with larger uncertainty. The second component,

M. Corduas (✉)
Dipartimento di Scienze Statistiche, Universitá di Napoli Federico II, Via L. Rodinò 22, 80138 Napoli, Italy
e-mail: corduas@unina.it

B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-13312-1_22,

instead, is connected to the fundamental personal feeling of liking/disliking that the rater has for the item.

Both these features of the rater's choice are taken into account by defining a mixture of distributions: a discrete uniform and a shifted binomial distribution.

The CUB model has proved to be effective in numerous real applications arising in various fields such as social analysis, medicine, sensometrics and linguistics.

In complex surveys where several items are investigated, the comparison among the satisfaction level associated to each item is usually summarized by means of some indices such as the average or mode of the observed rating distributions which fail to produce a clear picture of evaluations or opinions expressed by respondents.

In this article, the use of Kullback-Liebler (KL) divergence is proposed in order to detect significant similarities and differences in the overall judgements expressed by raters and modelled by means of CUB models. Specifically, a testing procedure based on KL divergence is discussed and a clustering technique is presented. The proposed method is finally illustrated by means of a real data set from a survey on the perception of work riskiness in an industrial plant.

2 The CUB Model

The preference or score that a subject expresses describes a random variable R such that:

$$P(R = r) = \pi \binom{m-1}{r-1} (1-\xi)^{r-1} \xi^{m-r} + (1-\pi)\frac{1}{m}, \quad r = 1, 2, \ldots, m \quad (1)$$

where $\xi \in [0, 1]$, $\pi \in (0, 1]$ and m is the number of grades for evaluating an item. For a given $m > 3$, then, R is a Mixture of a Uniform and a (shifted) Binomial distribution.

The parameter π determines the role of *uncertainty* in the final judgment: the lower the weight $(1 - \pi)$ the smaller the contribution of the Uniform distribution in the mixture. On the other hand, the parameter ξ characterizes the shifted Binomial distribution and, therefore, depending on the meaning of the highest score (positive or negative judgment) it denotes the strength of "liking" (or "disliking") expressed by raters with respect to the item.

In a further extension of the model the influence of external factors in the final judgement is considered [14, 15]. Specifically, two relations, which connect the model parameters to significant covariates by means of a logistic link function, are added to (1). The following discussion, however, will focus on models without covariates.

In presence of experiments involving the judgement of several items, a graph of the CUB estimated coefficients, $\hat{\pi}$ and $\hat{\xi}$, in the parameter space has been often used in order to assess how close the models are [6]. However, this representation may result in misleading interpretations of data since the user tends to assess the closeness of two (estimated) CUB distributions in terms of the Euclidean distance

between the corresponding estimated parameters. As a matter of fact, the role of CUB coefficients is very different in determining the shape of the estimated distribution [13] and the dissimilarity between two CUB distributions cannot be explained by the simple Euclidean distance between the related parameter estimates.

3 Assessing Similarity of CUB Models

Thus, in order to perform the comparison of CUB models in a convincing manner, we propose using the KL divergence measure. For this purpose, we recall a general result derived by Kupperman [9]. Consider two discrete populations each characterized by a probability distribution function having the same functional form $p(x, \boldsymbol{\theta}_i)$ with unspecified vector parameters $\boldsymbol{\theta}_i$, $i = 1, 2$. Also assume that $p(x, \boldsymbol{\theta}_i) > 0, \forall x$. Suppose that we have two samples of N_1 and N_2 observations randomly drawn from the specified i-th population, respectively, and we wish to decide if they were in fact generated from the same population. In order to test the hypothesis $H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$ against $H_1 : \theta_{1j} \neq \theta_{2j}$, the KL divergence statistic is defined by:

$$\hat{J} = \frac{N_1 N_2}{N_1 + N_2} \left[\sum_x (p(x, \boldsymbol{\theta}_1) - p(x, \boldsymbol{\theta}_2)) \ln \frac{p(x, \boldsymbol{\theta}_1)}{p(x, \boldsymbol{\theta}_2)} \right]_{\boldsymbol{\theta}_1 = \hat{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_2 = \hat{\boldsymbol{\theta}}_2} \tag{2}$$

where the vector parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ have been replaced by the maximum likelihood estimators. Then, it can be shown that $\hat{J}$ is asymptotically distributed as a χ^2_g random variable when the null hypothesis is true, being g the dimension of the vector parameter [8]. In the case under investigation, $g = 2$.

A strategy for comparing and grouping CUB models is as follows. First, the KL divergence is evaluated for each couple of models and a binary matrix is built by setting the (i, j)th entry equal to 0, when the hypothesis of homogeneity of the i-th and j-th models is rejected, and 1 otherwise.

Secondly, this matrix is rearranged into an approximate block diagonal form. A clearly defined (unit) triangle immediately under the diagonal will indicate a cluster of items for which the judgements expressed by respondents, summarized by means of the CUB distributions, are similar. The presence of any zero value in such a triangle indicates that the cluster may be elongated or constituted by other well separated small clusters.

Several algorithms were proposed in literature for this aim (see for instance, Climer and Zhang [4] and references reported therein). In particular, in the rest of this article, we will refer to the BEA algorithm (that is the bond energy algorithm by [2, 10]). This procedure operates on an $M \times N$ matrix $\mathbf{A}$ of nonnegative entries and changes the arrangement of the rows and columns of $\mathbf{A}$ in order to maximize the expression:

$$ME = \sum_{j=1}^{M} \sum_{k=1}^{N} a_{j,k}[a_{j,k-1} + a_{j,k+1} + a_{j-1,k} + a_{j+1,k}], \tag{3}$$

where the maximization is over all $N!M!$ possible arrays that can be obtained from permuting $\mathbf{A}$ (with the convention that $a_{0,k} = a_{M+1,k} = a_{j,0} = a_{j,N+1} = 0$). The idea is that large values will be drawn to other large values (and vice versa small values to other small values) so as to increase the overall sum of the products. In general, the method has an additivity property so that the optimization of ME can be performed in two steps. Nevertheless, in the specific case that we are considering, since the binary matrix is symmetric, the same optimal ordering must hold for both rows and columns; hence it is only necessary to compute this ordering once. [1]

Finally, groups of items for which the raters express similar structure of judgements can be recognized as unit blocks along the diagonal of the reordered matrix.

4 An Application

The proposed procedure is illustrated by means of a case study concerning the perception of 348 employees about 10 types of causes and contributory factors of accidents at work (structural collapse, contact with electrical appliances, contact with moving machinery/part, eye contact, vehicle contact, fire or explosions, slip and falls, strain, contact with sharp edges, hits) and 5 aspects of risk perception (injury seriousness, frequency, fear of exposure, own ability to control or avoid risk, and training) from a survey in a printing and publishing plant. The judgements are expressed using a 7 point Likert scale where 7 relates to the highest perceived risk.

The data set was analyzed by Cerchiello et al. [3] who illustrated the risk perception by means of the plot of the estimated CUB model coefficients for each risk factor. Fig. 1 illustrates this type of plot for two categories: "Accidents Frequency" and "Ability to control/avoid hazards" and the graph of the related estimated CUB distribution. Note that although R is a discrete random variables, the CUB distributions are represented by lines in order to enhance the distribution shape. In both cases, the related CUB models are characterized by a rather low $\hat{\pi}$ coefficient which implies a fairly large uncertainty component. This remark applies to most of the items object of this study. Moreover, the graphs in the lower panels enhance that there are some hazards ("structural collapse", "contact with electrical appliances", "vehicle contact", "fire or explosions", "strain", "contact with sharp edges") which workers firmly believe to be able to control or avoid. The CUB distributions of the items are well separated. The grouping which can be detected in the graph representing the estimated coefficients in the parameter space is confirmed by the corresponding graph of the estimated CUB distributions. Instead, in case of the "Accidents Frequency" the interpretation of the above mentioned graphs is less clear. As a matter of fact, the closeness of some points in the parameter space cannot be easily recognized by the similarity of the distributions.

[1] The algorithm was implemented using GAUSS 8.0 system by Aptech Inc.; a routine is also available in the multivariate data analysis package of R (multiv).

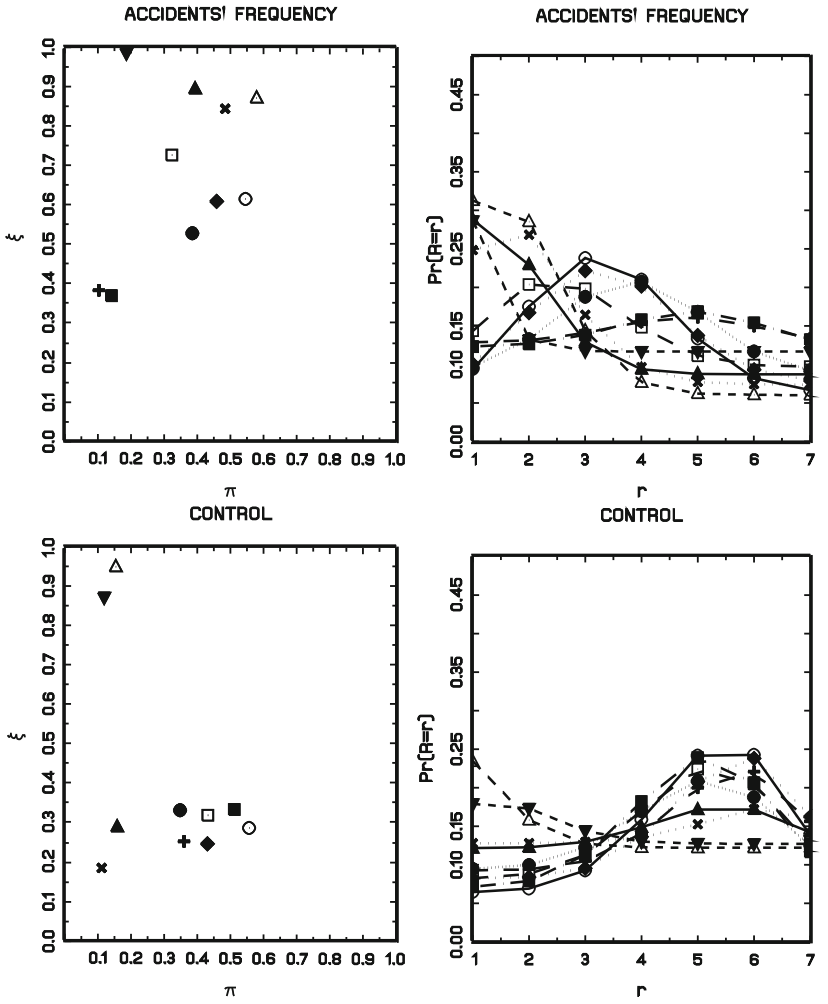


Fig. 1 Model coefficients (*left panel*) and estimated CUB distributions (*right panel*)
*L*egend: (O) structural collapse, (□) contact with electrical appliances, (×) contact with moving machinery/part, (△) eye contact, (●) vehicle contact, (■) fire or explosions, (▲) slip and falls, (+) strain, (♦) contact with sharp edges, (▼) hits.

Then, the proposed BEA technique is applied to all the observed items in order to detect groups of items which generate a similar overall perception of risk among workers. The significance level is set to 5%. Hereinafter the causes of injuries are numbered sequentially for each perceived risk, so that the first factor (injury seriousness) related to the various causes are numbered from 1 to 10; the second factor from 11 to 20 and so on.

The procedure identifies the following groups: $G_1 = (13, 14, 24, 27, 30)$, $G_2 = (10, 20, 34, 40, 46, 47, 50)$, $G_3 = (6, 12, 22, 23, 26, 28, 42, 43, 44, 45)$,

$G_4 = (11, 19)$, $G_5 = (16, 21, 5)$, $G_6 = (2, 8, 15, 25, 41)$, $G_7 = (29, 33, 37)$, $G_8 = (31, 32, 38, 39)$, $G_9 = (1, 9)$. The remaining items are initially isolated, but allowing for elongated clusters leads to further agglomerations: (35,36) with G_8, (17) with G_1, (7) with G_3 and, finally, only the following detached items are left: (3),(4),(18),(48),(49).

The CUB distributions of the clustered items are illustrated in Fig. 2 On one hand, workers are very uncertain in rating risk factors related to some items. The clusters G5, G2, G6, G7 show rather flat distributions with a modest dominance of low (G2), high (G7) or middle (G6) rates. Also, we unexpectedly find that "training" that workers have undergone for preventing any of the considered hazards is not perceived as adequate (as regards this aspect, most CUB distributions related to various accident causes belong to G1, G2, G3). On the other hand, workers have a clear and precise opinion about the "injury seriousness" derived from "structural collapses" and "contact with sharp edges" (G9) and, as mentioned above, they believe that "their own ability to control" is sufficient for reducing risks in relation to most hazards (G8) with the exception of "contact with eyes" and "hits" (G2). This result is consistent with usual findings in the printing publishing industry (see, for instance, Healy [7]) where fingers and hands lacerations, fractures and dislocations are the most frequent injuries because of the wide use of manual work and the proximity of workers with machineries, and heavy paper and ink rollers.

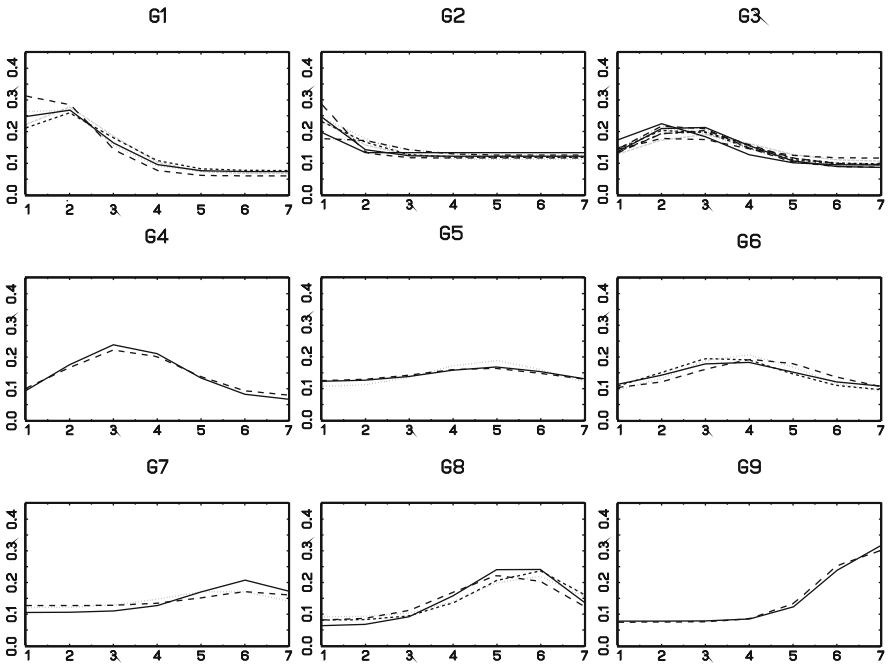


Fig. 2 Clustered items by KL divergence

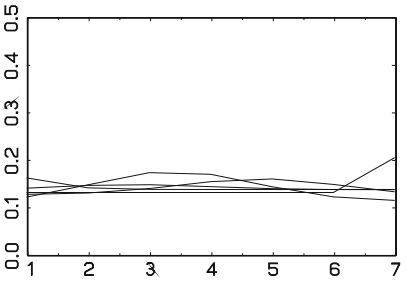


Fig. 3 Isolated elements: (3),(4),(18),(48),(49)

Moreover, workers do not show agreement of opinions on the "fear of exposure" to the various hazards in their firm. Most of the related ratings distributions belong to G1, G3 and G5, showing either uniform or negative skewed distribution. The isolated elements (Fig. 3) show rather flat distributions which again confirm that a part from few items for which workers have a clear and marked mental image, they find rather difficult to rate risk factors. From these findings, for instance, decision makers could plan some firm policy in order to increase workers' awareness about risk factors. Then, the proposed technique could provide useful results by comparing the distributions of ratings on a certain item observed in two time points: before and after the implementation of such policy. Finally, in Fig. 4, the estimated coefficients of the clustered CUB models are represented in the parameter space. The shape of clusters is generally stretched along the horizontal axis, confirming the substantially different role of the two parameters.

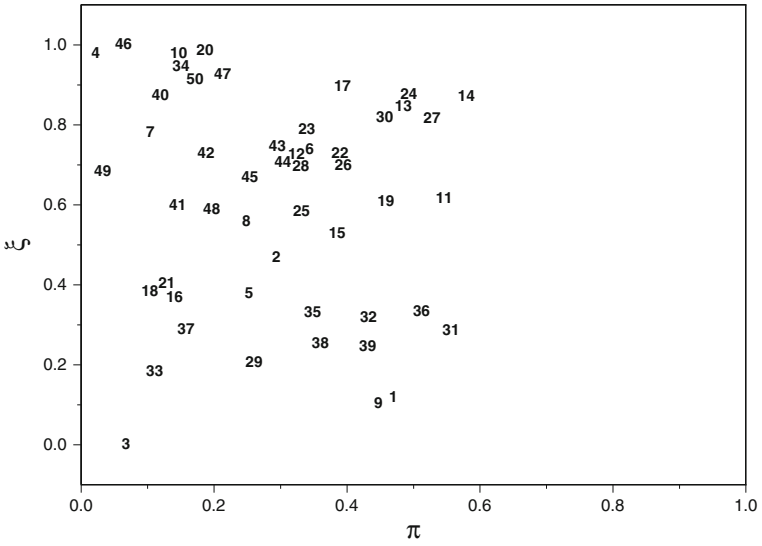


Fig. 4 Clusters of estimated CUB models in the parameter space

5 Final Remarks

The proposed technique helps the identification of similarities in the behaviour of groups of judges when they are asked to express their ratings on a set of items. Specifically, the technique is able to discriminate the different patterns of the scores distributions with respect to skewness, kurtosis, mode. Moreover, it helps to cluster items with respect to the overall ratings that the subjects express and it effectively overcomes the shortcomings of the coefficients plot.

Acknowledgments This research was partly funded by the MIUR-PRIN 2006 grant (Project on: "Stima e verifica di modelli statistici per l'analisi della soddisfazione degli studenti universitari") and CFEPSR (Portici, NA).

References

1. Agresti, A.: Categorical Data Analysis, 2nd ed. Wiley, New York, NY (2002)
2. Arabie, P., Hubert, L.J.: The bond energy algorithm revisited. IEEE Trans. Syst. Man Cybern. **20**, 268–274 (1990)
3. Cerchiello P., Iannario M., Piccolo D.: Assessing risk perception by means of ordinal models. In: Perna C., et al. (eds.) Mathematical and Statistical Methods for Insurance and Finance, pp. 65–73. Verlag, Berlin (2010)
4. Climer, S., Zhang, W.: Rearrangement clustering: Pitfalls, remedies and applications. J. Mach. Learn. **7**, 919–943 (2006)
5. D'Elia, A., Piccolo, D.: A mixture model for preference data analysis. Comput. Stat. Data Anal. **49**, 917–934 (2005)
6. D'Elia, A., Piccolo, D.: Uno studio sulla percezione delle emergenze metropolitane: un approccio modellistico. Quad. Stat. **7**, 121–161 (2005)
7. Healy, N.: Analysis of RIDDOR machinery accidents in the UK printing and publishing industries 2003–2004. Health Safety Laboratory. HSL/2006/83, Buxton, Derbyshire, UK (2006)
8. Kullback, S.: Information Theory and Statistics. Dover Publications, New York, NY (1959)
9. Kupperman, M.: Further Applications of Information Theory to Multivariate Analysis and Statistical Inference. George Washington University, Washington, DC (1957)
10. McCormick, W.T., Schweitzer, P.J., White, T.W.: Problem decomposition and data reorganization by a clustering technique. Oper. Res. **20**, 993–1009 (1972)
11. McCullagh, P., Nelder, J.A.: Generalized Linear Models. Chapman & Hall, London (1989)
12. McCullagh, P.: Regression models for ordinal data. J. R. Stat. Soc. Ser. B **42**, 109–142 (1980)
13. Piccolo, D., D'Elia, A.: A new approach for modelling consumers' preferences. Food Qual. Prefer. **19**, 247–259 (2008)
14. Piccolo, D.: Observed information matrix for MUB models. Quad. Stat. **8**, 33–78 (2006)
15. Piccolo, D.: On the moments of a mixture of uniform and shifted Binomial random variables. Quad. Stat. **5**, 86–104 (2003)

Sector Classification in Stock Markets: A Latent Class Approach

Michele Costa and Luca De Angelis

Abstract Stock indices related to specific economic sectors play a major role in portfolio diversification. Notwithstanding its importance, the traditional sector classification shows several flaws and it may not be able to properly discriminate the risk-return profile of financial assets. We propose a latent class approach in order to correctly classify the stock companies into homogenous groups under risk-return profile and to obtain sector indices which are consistent with the standard portfolio theory. Our results allow to introduce a methodological dimension in the stock's classification and to improve the reliability of sector portfolio diversification.

1 Introduction

Stock indices related to specific economic sectors play a major role in financial markets because they represent a main reference in portfolio diversification.

The purpose of this paper is to introduce a new sector classification, obtained by exploiting the potential of latent class (LC) models for classifying stock companies into homogenous groups under risk-return profile. The underlying hypothesis is that stocks belonging to the same sector are homogeneous, or, at least, that sectors characterize and influence the stock dynamics in a relevant way. In this framework, different sectors should be characterized by different risk and return levels. Moreover, sectors should be affected by the economic cycle thus introducing a distinction between pro-cyclic and anti-cyclic sectors. In order to achieve these goals, it is essential that the assignment of a single stock to a sector happens following a correct and strict methodological process.

The traditional sector classification shows several flaws on which it is urgent to suggest effective solutions. First, the traditional sector classification turns out to be strongly static, since it is rarely updated from the moment of a company IPO on the stock market. Second, stock companies frequently operate in different sectors, while sector classification considers only the main business. Finally, product sector

M. Costa (✉)
Dipartimento di Scienze Statistiche, Università di Bologna, Bologna, Italy
e-mail: michele.costa@unibo.it

B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-13312-1_23,

could not represent the relevant classification element in order to discriminate the risk-return profile.

The traditional observable sector classification is compared to the new classification, non-observable and achieved in LC field. The new proposal allows to both introduce a methodological dimension in the stocks classification and improve the investment opportunities.

In Sect. 2 we briefly introduce the LC model, while in Sect. 3 are illustrated the specification and the estimation of the LC model which better explains the associations among the variables. In Sect. 4 we explore the new stocks classification obtained by the estimated LC model and we show how this new classification is consistent with the standard portfolio theory and it improves the financial performance. Sect. 5 concludes.

2 Methodology

LC models provide an excellent framework in order to develop a new stock's classification because they require the use of manifest categorical variables and allow to explain the associations among the observed indicators through a set of latent categorical variables.

Standard portfolio theory evaluates stock risk-return profile on the basis of two latent variables, risk and expected return, which are usually approximated by means of two continuous variables, that is the standard deviation (S) and the mean (M) of the past observed returns. However, the characteristic of both S and M to be simple approximations, likely different from the true measures of the expected return and the risk, is frequently neglected. In order to stress the importance of this point, and also with the purpose of achieving a greater flexibility in the stock's classification, we propose to express S and M in the form of categorical ordinal variables. By using categorical ordinal variables we are able to go over the punctual value rigidity, thus avoiding a possible improper ranking imposed by the observed values. More specifically, we propose to reclassify stock's mean return punctual values into a few (two, three) categories, where the classes indicate low and high (or low, medium, and high) mean return. For example, let's consider a stock characterized by a low mean return, e.g. $M = 0.01$, and therefore classified in category 1 of the categorical indicator for the mean. It seems quite intuitive to deduce that the expected return of this stock will assume a low value, while it seems more difficult to affirm that the expected return will be exactly equal to 0.01. This new classification, based on ordinal variables, suggests a more flexible ranking and hence a likely closer correspondence between mean indicator and expected return.

An analogous procedure can be applied also to risk indicators. A further improvement in risk evaluation can be obtained by taking into account also extreme values, which frequently strongly characterize the return distributions. To this purpose, we propose to include among the manifest variables also the first percentile (P) of the return distributions.

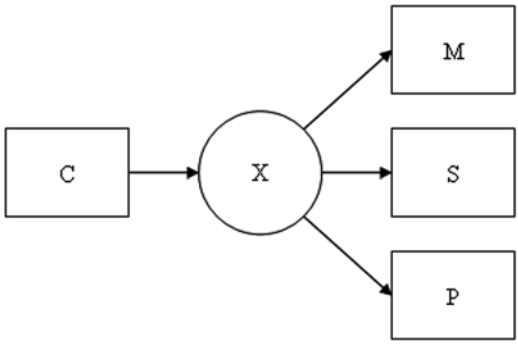


Fig. 1 LC model graphical representation

In the following we adopt the simplest possible specification, where indicator variables M, S, and P (with indices m, s, and p, respectively) have two categories, low and high, defined on the basis of the median. In addition we use the traditional sector C (with index c) as covariate.

The LC model is based on the local independence assumption which implies that all the relationship among the indicator variables are explained by the latent variable X. As shown in Fig. 1, the covariate C directly influences the latent variable but not the indicators.

A LC model with such configuration is specified as

$$\pi_{mspc} = \sum_{x=1}^{K} \pi_{xmspc}$$

where K denotes the number of latent classes and

$$\pi_{xmspc} = \pi_c \pi_{x|c} \pi_{msp|x} = \pi_c \pi_{x|c} \pi_{m|x} \pi_{s|x} \pi_{p|x}. \tag{1}$$

As classical latent class theory specifies, π_{xmspc} is the proportion of units in the five-way contingency table, $\pi_{x|c}$ is the probability of belonging to latent class x (given the covariate c), $\pi_{msp|x}$ is the probability of having a particular observed response pattern (m, s, p) given $X = x$, and π_c is the probability of each traditional sector. The other π parameters are conditional response probabilities: for instance, $\pi_{m|x}$ is the probability of being in category m of variable M, given that one belongs to latent class x. Model in Eq. (1) is known as the classical parameterization of the unrestricted LC model introduced by Lazarsfeld [3] with external variables [2].

One important goal of LC analysis is to determine the smallest number of latent classes K which is sufficient to explain the associations observed among the indicator variables. K is determined by comparing the log-likelihood ratio chi-square statistic L^2 of models with different number of latent classes and the Akaike information criterion [1]. The determination of the number of latent classes is a significant step in our work because it represents the number of sectors in which the new classification is constituted. Furthermore, a test for the choice of K greatly improves

the determination of the number of sectors, moving this decision from a subjective ground to a methodologically correct framework.

The last step of LC analysis is to classify units into the appropriate latent class. Units are assigned to the class for which the posterior membership probability is the highest. This approach is usually known as LC cluster model because the goal of classification into K homogenous groups is identical to that of cluster analysis [4].

Finally, in order to compare the latent class methodology with respect to some more traditional clustering approach, we refer to the K-means technique, which can be viewed as a particular case of LC model [5]. Since the K-means approach does not provide diagnostic statistics able to indicate the number of clusters, this choice has to be made in advance. In order to ensure a greater comparability, in the following we suggest for the K-means method the same number of clusters as in the LC model.

3 Model Estimation

We estimate LC models with a different number of latent classes in order to determine the smallest number which is able to account for the relationship observed among the indicator variables by using Latent Gold computer program [6].

We analyze a data set concerning the monthly return distribution from January 2002 to December 2007 of 136 stocks quoted at the Italian stock market. The selected stocks belong to 5 of the 10 sectors of the Global Industry Classification Standard (GICS): energy, consumer discretionary, utilities, finance, and materials.

The analysis typically starts by fitting the 1-class baseline model, which implies mutual independence among the variables. If the baseline model provides an adequate fit to the data, no LC analysis is needed, since there is no association among the variables to be explained.

The results of the different LC models are reported in Table 1. According to the L^2 statistic ($L^2 = 110.63$, $df = 32$, $p < .01$), the 1-class model must be rejected, thus indicating that the amount of association between the observed variables is too large to be explained without involving a latent variable with, at least, 2 classes.

The 2-class model provides a significant reduction of L^2 (63% from the baseline model). However, this statistic is still too high ($L^2 = 40.65$, $df = 27$, $p < .05$). Adding a third class to the model provides a further reduction in L^2 (a 74% reduction over the baseline model) and also provides an adequate overall fit. Table (1) shows

Table 1 Results from LC models with different number of classes

Model	LL	NPar	L^2	df	p-value	AIC(LL)
1-class	−282.789	3	110.628	32	$1.4E-10$	571.579
2-class	−247.798	8	40.646	27	0.04	511.596
3-class	−242.125	13	29.300	22	0.14	510.251
4-class	−239.544	18	24.137	17	0.12	515.088

Table 2 3-class unrestricted LC model results, conditional probabilities and indicator means

Indicator		Class 1 0.4332	Class 2 0.2904	Class 3 0.2763
	$\pi_{m=low\|x}$	0.5752	0.0280	0.9053
M	$\pi_{m=high\|x}$	0.4248	0.9720	0.0947
	mean	1.4248	1.9720	1.0947
	$\pi_{s=low\|x}$	0.0089	0.7675	0.9872
S	$\pi_{s=high\|x}$	0.9911	0.2325	0.0128
	mean	1.9911	1.2325	1.0128
	$\pi_{p=low\|x}$	0.0555	0.9573	0.7146
P	$\pi_{p=high\|x}$	0.9445	0.0427	0.2854
	mean	1.9445	1.0427	1.2854

that according to the AIC statistic, which takes parsimony into account, the 3-class model is also preferred over the 4-class model.

Also a further diagnostic statistic, the bivariate residual χ^2-based test [4], confirms the choice of the 3-class model.

Table 2 reports maximum likelihood estimation results for the 3-class model. The parameter estimates show that the three classes have quite similar probabilities: 43% of the stocks are estimated to be in Class 1 ($\pi_{x=1} = 0.43$), 29% in Class 2 ($\pi_{x=2} = 0.29$), and the remaining 28% in Class 3 ($\pi_{x=3} = 0.28$).

The characteristics of the three classes can be determined on the basis of the means of the indicators for each latent variable. The main feature which characterises the first class is the high risk: Class 1 has the highest values of indicators S and P (their means are 1.99 and 1.94, respectively) and M mean equal to 1.42. The second class is characterized by low risk and high return: Class 2 shows the lowest P mean (1.04), S mean equal to 1.23, and the highest M mean (1.97). According to its values of S and P means (1.01 and 1.29 respectively), the third class is characterized by low risk and the lowest value of M mean (1.09).

The conditional probabilities $\pi_{m|x}$, $\pi_{s|x}$, and $\pi_{p|x}$ in Table (2) underline that the characteristics of the three classes are quite well defined under the stock risk-return profile: Class 2 is the latent class which allows the best investment opportunities, Class 1 is the most risky, and Class 3 is defined by low risk but also low return.

In order to stress the advantages of our proposal, we develop the analysis also by referring to the traditional K-means technique, performed by using the original values of mean, standard deviation and first percentile.

According to K-means (Table 3 and Fig. 2), stocks are classified into three quite heterogeneous clusters: Cluster 1 includes 52 stocks and it is characterized by the highest mean and the lowest standard deviation and 1st percentile, Cluster 2 is com-

Table 3 Results from the K-means method

Cluster	Mean	Std.Dev.	1st Perc.	Size	Class 1	Class 2	Class 3
1	0.926	5.925	−11.379	52	0	32	20
2	0.722	9.017	−18.931	65	39	11	15
3	0.408	12.251	−29.222	19	19	0	0

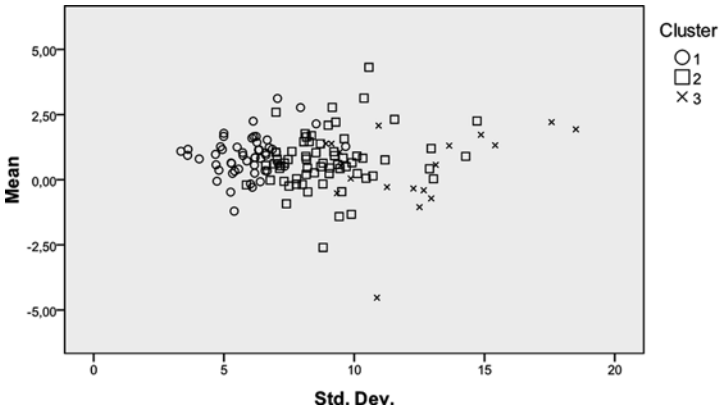


Fig. 2 K-means method results: mean and standard deviation of stock's returns in the three clusters

posed by 65 stocks and it assumes medium values of the three indicators, and Cluster 3, of only 19 stocks, is characterized by the lowest mean and the highest values of standard deviation and 1^{st} percentile. The stock's classification achieved with K-means shows some difference with respect to the classification obtained by using the 3-class LC model. The last three columns in Table (3) show how the stocks assigned to each cluster are classified into the latent classes. Cluster 1 is composed by the stocks allocated into Class 2 and Class 3 of the LC classification. Into Cluster 2 are allocated stocks from all the three classes, in majority from Class 1. Finally, Cluster 3 contains the stocks originally assigned to the first latent class. Analyzing intersections and differences between the K-means clusters and the latent classes, it highlights that LC model is able to define more homogenous sectors under the risk-return profile.

4 The New Stock's Classification

The LC model estimation allows to assign each stock to one of the three classes, thus obtaining the new classification. The new sectors are constituted by 58, 43, and 35 stocks respectively. For each traditional sector, Table 4 shows the weight of the latent classes. It can be observed that the majority (62.9%) of Consumer Discretionary stocks is classified into Class 1, while none of the utility stocks is assigned to the first latent class. In all other cases traditional sectors quite equally contribute to

Table 4 Allocation of the traditional sectors into the latent classes

Traditional Sector	Class 1	Class 2	Class 3
Energy	0.3902	0.4020	0.2078
Consumer Discretionary	0.6293	0.1353	0.2354
Finance	0.3517	0.3745	0.2738
Utilities	0.0002	0.4754	0.5243
Materials	0.4257	0.3139	0.2604

Table 5 Sector indices mean, standard deviation, percentiles, and Sharpe ratio

Index	Class 1	Class 2	Class 3	Energy	Cons. Disc.	Finance	Utilities	Materials
Mean	0.65	1.39	0.25	2.26	0.62	0.82	0.75	0.73
St.Dev.	5.66	3.64	3.73	6.05	4.92	4.54	4.43	4.70
1st Perc.	−14.2	−7.1	−9.5	−13.3	−12.3	−10.9	−11.7	−10.7
5th Perc.	−9.7	−5.6	−6.1	−8.0	−8.7	−6.7	−7.9	−8.2
Sharpe	0.07	0.32	0.01	0.33	0.08	0.13	0.12	0.11

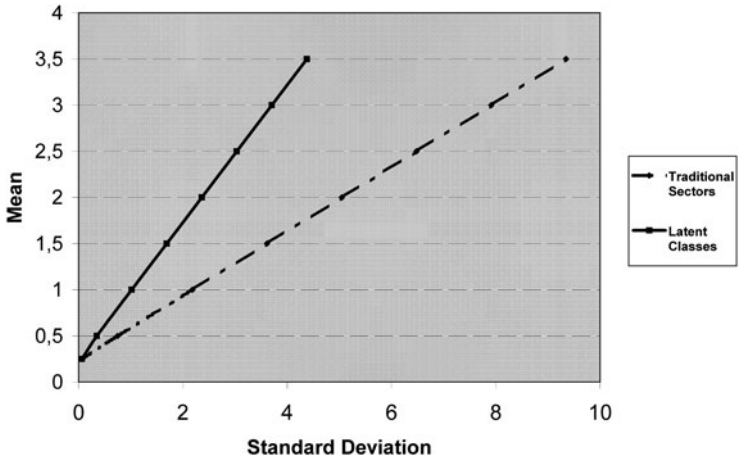


Fig. 3 Efficient frontiers calculated on traditional sectors and latent classes

the definition of all the new classes. We interpret this behavior as evidence that traditional sectors are not consistent under the stock risk-return profile.

Furthermore, in order to evaluate and compare the different performances, we calculate equal-weighted indices for each of both the traditional and the new sectors. Table 5 reports mean, standard deviation, 1st and 5th percentile, and Sharpe ratio of these indices. According to Sharpe ratio [8], which measures the excess return (with respect to 3-month Italian Treasury Bill) per unit of risk, Classes 2 performs better than all of the analyzed traditional sectors, except for Energy. On the contrary, Class 1 and Class 3 perform the worst.

In the framework of the standard portfolio theory [7], we finally compare the efficient frontier based on the traditional sectors to the efficient frontier related to the new classification. As shown in Fig. 3, the latter (solid line) performs much better than the one calculated on traditional sectors (dashed line): for a given level of return mean, it indicates a much lower standard deviation.

5 Conclusions

Our work shows how LC models represent an appropriate method in order to classify stocks into homogenous groups under risk-return profiles. We find evidence of a three-class latent model which allows to obtain a new stocks classification. Our proposal allows to overcome some problems related to traditional sector clas-

sification and to indicate a methodologically correct solution. Finally, our results are consistent with the standard portfolio theory and provide more efficient portfolio allocations than traditional sector classification, thus giving new and improved investment opportunities.

References

1. Akaike, H.: A new look at the statistical model identification. IEEE Trans. Automat. Contr. **19**(6), 716–723 (1974)
2. Hagenaars, J.A.: Categorical Longitudinal Data – Loglinear Analysis of Panel, Trend and Cohort Data. Sage, Newbury Park, CA (1990)
3. Lazarsfeld, P.F., Henry, N.W.: Latent Structure Analysis. Houghton Mill, Boston, MA (1968)
4. Magidson, J., Vermunt, J.K.: Latent class factor and cluster models, Bi-plots and related graphics displays. Sociol. Methodol. **31**, 223–264 (2001)
5. Magidson, J., Vermunt, J.K.: Latent class models for clustering: A comparison with K-means. Can. J. Mark. Res. **20**, 37–44 (2002)
6. Magidson, J., Vermunt, J.K.: Latent GOLD 4.0 User's Guide. Statistical Innovations Inc., Belmont, MA (2005)
7. Markowitz, H.: Portfolio selection. J. Finance **7**(1), 77–91 (1952)
8. Sharpe, W.F.: The Sharpe Ratio. J. Portf. Manag. **21**(1), 49–58 (1994)

Partitioning the Geometric Variability in Multivariate Analysis and Contingency Tables

Carles M. Cuadras and Daniel Cuadras

Abstract Most methods of multivariate analysis obtain and interpret an appropriate decomposition of the variability. In canonical variate analysis, multidimensional scaling and correspondence analysis, the variability of the data is measured in terms of distances. Then the geometric variability (inertia) plays an important role. We present a unified approach for describing four methods for representing categorical data in a contingency table. We define the generalized Pearson contingency coefficient and show situations where this measure can be different from the geometric variability.

1 Introduction

A common practice in multivariate analysis is to obtain an appropriate decomposition of the variability. When the variability can be summarized in a covariance matrix Σ, the total variance $\mathrm{tr}(\Sigma)$ and the generalized variance $|\Sigma|$, are two general measures of dispersion depending on the eigenvalues of Σ.

Often the variability is related to a distance measure. A clear example is the sample variance of n univariate observations, which can be expressed as the average of unidimensional Euclidean distances between n^2 pair of observations. In some multivariate methods such as canonical variate analysis, multidimensional scaling and correspondence analysis, the variability of the data is measured in terms of distances. In these methods it is natural to consider the geometric variability (GV) as a measure of dispersion. We also define the generalized Pearson contingency coefficient (GPC). Although GV is in general equivalent to GPC, we show situations where both measures are essentially different.

C.M. Cuadras (✉)
Department of Statistics, University of Barcelona, Barcelona, Spain
e-mail: ccuadras@ub.edu

B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-13312-1_24,

2 Geometric Variability

2.1 Finite Set

Let $\Omega = \{\omega_1, \ldots, \omega_g\}$ be a set with g objects, δ a distance function on Ω providing the $g \times g$ distance matrix $\Delta_g = (\delta_{ij})$, where $\delta_{ij} = \delta(\omega_i, \omega_j)$. Let $w = (w_1, \ldots, w_g)'$ a weight vector such that $w'1 = \sum_{i=1}^{g} w_i = 1$, with $w_i \geq 0$ and 1 the column vector of ones. The geometric variability (GV) of Ω with respect to δ is defined by

$$V_\delta = \frac{1}{2} \sum_{i,j=1}^{g} w_i \delta_{ij}^2 w_j = \frac{1}{2} w' \Delta_g^{(2)} w,$$

where $\Delta_g^{(2)} = (\delta_{ij}^2)$.

The GV as one half the average of distances, was considered by Light and Margolin [15] in categorical data analysis, by Rao [16] in studying the quadratic entropy and by Cuadras et al. [7] in distance-based discriminant analysis.

The GV is usually related to the problem of displaying the elements of Ω as points in Euclidean space of low dimension. Assuming $(I_g - 1w') \left(-\frac{1}{2}\Delta_g^{(2)}\right) (I_g - w1')$ s.d.p., the weighted metric MDS solution finds the spectral decomposition

$$D_w^{1/2} (I_g - 1w') \left(-\frac{1}{2}\Delta_g^{(2)}\right) (I_g - w1') D_w^{1/2} = U \Lambda^2 U', \tag{1}$$

where $D_w = \text{diag}(w)$. Matrix $X = D_w^{-1/2} U \Lambda$ contains the principal coordinates of Ω.

Let $G = XX'$ and d the column vector with the diagonal entries in G. Then $\Delta_g^{(2)} = d1' + 1d' - 2G$. Since $w'X = 0$ and $w'1 = 1$, we have $d'w = \text{tr}(D_w^{1/2} G D_w^{1/2}) = \text{tr}(U \Lambda^2 U') = \text{tr}(\Lambda^2)$. Thus the geometric variability (also called inertia) is $V_\delta = \sum_{i=1}^{K} \lambda_i^2$.

2.2 Random Vector

Let $\mathbf{X}$ be a random vector with pdf $f(\mathbf{x})$, with respect to a suitable measure, and support S. Let $\delta(\mathbf{x}, \mathbf{y})$ be a distance function between the observations of $\mathbf{X}$. The GV of $\mathbf{X}$ with respect to δ is defined by

$$V_\delta(\mathbf{X}) = \frac{1}{2} \int_{S \times S} \delta^2(\mathbf{x}, \mathbf{y}) f(\mathbf{x}) f(\mathbf{y}) d\mathbf{x} d\mathbf{y}.$$

Let us suppose that there exists a representation $\psi : S \to L$ of S in a Euclidean (or separable Hilbert) space L with inner product $\langle \cdot, \cdot \rangle$ and related norm $|| \cdot ||$, such

that $\delta^2(\mathbf{x}, \mathbf{y}) = ||\psi(\mathbf{x}) - \psi(\mathbf{y})||^2$. Then $V_\delta(\mathbf{X}) = E||\psi(\mathbf{X})||^2 - ||E(\psi(\mathbf{X})||^2$, which is formally similar to the variance. Related to $V_\delta(\mathbf{X})$ is the proximity function from an observation $\mathbf{x}$ to the population represented by $\mathbf{X}$

$$\phi_\delta^2(\mathbf{x}) = \int_S \delta^2(\mathbf{x}, \mathbf{y}) f(\mathbf{y}) d\mathbf{y} - V_\delta(\mathbf{X}).$$

If we transform the distance $\widetilde{\delta^2}(\omega, \omega') = a\delta^2(\omega, \omega') + b$, if $\omega \neq \omega'$, then $V_{\widetilde{\delta}}(\mathbf{X}) = aV_\delta(\mathbf{X}) + b/2$ and $\phi_{\widetilde{\delta}}^2(\mathbf{x}) = a\phi_\delta^2(\mathbf{x}) + b/2$. Thus we can consider suitable choices of a, b and generate the probability density $f_\delta(\mathbf{x}) = \exp(-\phi_\delta^2(\mathbf{x}))$. Then

$$I(f||f_\delta) = V_\delta(\mathbf{X}) - H(f) \geq 0,$$

where $I(f||f_\delta)$ is the Kullback-Leibler divergence and $H(f)$ is the Shannon entropy. Therefore $H(f)$ is the lower bound for the GV of a random vector. It can be proved that GV reaches $H(f)$ only if $f_\delta(\mathbf{x})$ coincides wit the true density $f(\mathbf{x})$. See Cuadras et al. [4], [7].

2.3 Mixtures

Now suppose that $f(\mathbf{x}) = w_1 f_1(\mathbf{x}) + \cdots + w_g f_g(\mathbf{x})$ is the mixture of g densities with the same support S. Assume that the above representation $\psi : S \to L$ exists. Then the GV with respect to a distance δ is given by

$$V_\delta(\mathbf{X}) = V(\mu_1, \ldots, \mu_g) + \sum_{i=1}^{g} w_i V_i,$$

where $V(\mu_1, \ldots, \mu_g) = \frac{1}{2}\sum_{i,j=1}^{g} w_i \delta^2(\mu_i, \mu_j) w_j = \sum_{i=1}^{g} w_i \delta^2(\mu_i, \mu)$, with $\mu_i = E_i(\psi(\mathbf{X}))$, $\delta^2(\mu_i, \mu_j) = ||\mu_i - \mu_j||^2$, $\mu = w_1\mu_1 + \cdots + w_g\mu_g$, and $V_i = \frac{1}{2}\int_{S\times S} \delta^2(\mathbf{x}, \mathbf{y}) f_i(\mathbf{x}) f_i(\mathbf{y}) d\mathbf{x}d\mathbf{y}$. We can interpret $V_\delta(\mathbf{X})$ as the total GV, which splits into two parts: between and within groups GV, as it is shown in the next section.

3 Distance-Based Analysis of Variance

Suppose $g \geq 2$ independent data sets of sizes $n_1, \ldots, n_g$ coming from the populations $\Pi_1, \ldots, \Pi_g$. To test $H_0 : \Pi_1 = \cdots = \Pi_g$ let us assume that, by means of a distance function δ between observations, we can obtain the intra-distance matrices $\Delta_{11}, \ldots, \Delta_{gg}$, and the inter-distance matrices $\Delta_{12}, \ldots, \Delta_{g-1g}$. The overall distance matrix is Δ, and the GV can be defined for Δ and for each group separately. By taking principal coordinates, we can obtain the following $p \times p$ matrices:

$$\begin{aligned}\mathbb{T} &= \sum_{k,h=1}^{g} \sum_{i,i'=1}^{n_k,n_h} (\mathbf{x}_{ki} - \mathbf{x}_{hi'})(\mathbf{x}_{ki} - \mathbf{x}_{hi'})',\\ \mathbb{B} &= \sum_{k,h=1}^{g} n_k n_h (\overline{\mathbf{x}}_k - \overline{\mathbf{x}}_h)(\overline{\mathbf{x}}_k - \overline{\mathbf{x}}_h)',\\ \mathbb{W}_k &= \sum_{i,i'=1}^{n_k} (\mathbf{x}_{ki} - \mathbf{x}_{ki'})(\mathbf{x}_{ki} - \mathbf{x}_{ki'})',\end{aligned}$$

which satisfy $\mathbb{T} = \mathbb{B} + n\sum_{k=1}^{g} n_k^{-1}\mathbb{W}_k$. Hence $\mathrm{tr}(\mathbb{T}) = \mathrm{tr}(\mathbb{B}) + n\sum_{k=1}^{g} n_k^{-1}\mathrm{tr}(\mathbb{W}_k)$, where $n = \sum n_i$. Thus, for general multivariate data and working only with distances, it is possible to decompose GV as follows:

$$V_\delta(\text{total}) = V_\delta(\text{between}) + n^{-1}\sum_{i=1}^{g} n_i V_\delta(\text{within } i). \tag{2}$$

For testing H_0 we can use the statistic $\gamma = V_\delta(\text{between})/V_\delta(\text{total})$, see [3].

An early example is Light and Margolin [15]. They proposed an analysis of variance for categorical data (CATANOVA). Let $N = (n_{ij})$ be an $I \times J$ contingency table and $P = n^{-1}N$ the correspondence matrix, where $n = \sum_{ij} n_{ij}$. Let $K = \min\{I, J\}$ and $r = P1$, $D_r = \mathrm{diag}(r)$, $c = P'1$, $D_c = \mathrm{diag}(c)$, the vectors and diagonal matrices with the marginal frequencies of P. The marginals counts are $n_{i\cdot}$ and $n_{\cdot j}$. CATANOVA deals with categorical data in I groups and J categories, and uses the distance $\delta_{ij} = 1$ if i and j are the same category, 0 otherwise. With the marginal vector $(n_{\cdot 1}, \ldots, n_{\cdot J})$ and $(n_{i1}, \ldots, n_{iJ})$ for each group, the above GV (total, between and within) are the total, between and within groups sum of squares:

$$\begin{aligned}V_\delta(\text{total}) &= TSS = \tfrac{n}{2} - \tfrac{1}{2n}\sum_{j=1}^{J} n_{\cdot j}^2,\\ V_\delta(\text{between}) &= BSS = \tfrac{1}{2}\left(\sum_{i=1}^{I}\tfrac{1}{n_{i\cdot}}\sum_{j=1}^{J} n_{ij}^2\right) - \tfrac{1}{2n}\sum_{j=1}^{J} n_{\cdot j}^2,\\ V_\delta(\text{within}) &= WSS = \tfrac{n}{2} - \tfrac{1}{2}\sum_{i=1}^{I}\tfrac{1}{n_{i\cdot}}\sum_{j=1}^{J} n_{ij}^2.\end{aligned}$$

Then $TSS = BSS + WSS$ and a test is proposed based on $R^2 = BSS/TSS$. See [15] for details.

4 Contingency Tables

4.1 General Approach

There are several methods for visualizing the rows and columns of a contingency table. We present a general approach, which includes correspondence analysis (CA). In these methods, besides the GV, it is also used the generalized Pearson contingency coefficient.

Given $N = (n_{ij})$, we consider r, D_r, P, etc., defined above. In order to represent the rows and columns of N, Goodman [9] introduced the generalized nonindependence analysis (GNA) by means of the SVD:

$$D_r^{1/2}(I - 1r')(R[D_r^{-1}PD_c^{-1}])(I - c1')D_c^{1/2} = U\Lambda V', \tag{3}$$

where $R(x)$, for $x > 0$, is any monotonically increasing function and $R(M)$ is applied term by term. The principal coordinates for rows and columns are given by $A = D_r^{-1/2}U\Lambda$, $B = D_c^{-1/2}V\Lambda$. Clearly GNA reduces to CA when $R(x) = 1$.

A suitable choice of $R(x)$ is the Box-Cox transformation $R(x) = (x^\alpha - 1)/\alpha$ if $x > 0$, $R(x) = \ln(x)$ if $\alpha = 0$. With this function, let us consider the following SVD depending on three parameters:

$$D_r^{1/2}(I - \gamma 1r')\left(\frac{1}{\alpha}\left[\left(D_r^{-1}PD_c^{-1}\right)^\alpha - 11'\right]\right)D_c^\beta = U\Lambda V', \tag{4}$$

where $M^\alpha = (m_{ij}^\alpha)$. Note that $(I - c1')$ in (3) is missing in (4), see below.

The principal coordinates for the I rows and the standard coordinates for the J columns of N are given by $A = D_r^{-1/2}U\Lambda$ and $B_* = D_c^{-\beta}V$, respectively. B_* is used in the sense that A, B_* reconstitute the model: $(I - \gamma 1r')\left(\frac{1}{\alpha}[(D_r^{-1}PD_c^{-1})^\alpha - 11']\right) = AB_*'$. However, different weights are used for the column graphical display, for instance, $B = D_c^\beta V\Lambda$.

Implicit with this (row) representation is the squared distance between rows

$$\delta_{ii'}^2 = \sum_{j=1}^{J}\left[\left(\frac{p_{ij}}{r_i c_j}\right)^\alpha - \left(\frac{p_{i'j}}{r_{i'} c_j}\right)^\alpha\right]^2 c_j^{2\beta}. \tag{5}$$

The first principal coordinates account for a relative high percentage of inertia. This parametric approach has been explored by Cuadras and Cuadras [5] and Greenacre [11]. See also [6]. Here we use Greenacre's α parametrization.

The GV (for representing rows) is one half average of the distances weighted by the row marginal frequencies: $GV = \frac{1}{2}r'\Delta^{(2)}r$, where $\Delta^{(2)} = (\delta_{ii'}^2)$ is the $I \times I$ matrix of squared parametric distances (5).

For measuring the dispersion in model (4), let us introduce the generalized Pearson contingency coefficient

$$\phi^2(\alpha, \beta) = \sum_{i=1}^{I}\sum_{j=1}^{J}\left[\left(\frac{p_{ij}}{r_i c_j}\right)^\alpha - 1\right]^2 r_i c_j^{2\beta}.$$

Note that $GV = \phi^2(\alpha, \beta) = 0$ under statistical independence between rows and columns. In general $GV \neq \phi^2(\alpha, \beta)$. The unified approach for all methods (centered and uncentered) discussed below, are given in Table 1.

Two remarks: 1) From $(I - 1r')(D_r^{-1}PD_c^{-1} - 11') = D_r^{-1}PD_c^{-1} - 11'$, the centered ($\gamma = 1$) and uncentered ($\gamma = 0$) solutions coincide in CA and also in NSCA. 2) In order to give a weighted MDS approach compatible with (1), we mainly consider generalized versions without right-centering, i.e., without post-multiplying

Table 1 Four methods for representing rows and columns in a contingency table

Method	Inertia (centered) $\gamma = 1$ $GV = \sum \lambda_i^2$	Inertia (uncent.) $\gamma = 0$ $\phi^2(\alpha, \beta) = \sum \lambda_i^2$
CA $\alpha = 1, \beta = 1/2$ Pearson-Benzécri	$GV = \sum_{i,j} \left(\frac{p_{ij}}{r_i c_j} - 1 \right)^2 r_i c_j$	$\phi^2(1, 1/2) = GV$
HD $\alpha = \beta = 1/2$ Matusita-Rao	$GV = 1 - \sum_j (\sum_i \sqrt{p_{ij} r_i})^2$	$\phi^2(1/2, 1/2) = 2(1 - \sum_{i,j} \sqrt{p_{ij} r_i c_j})$
LR $\alpha = 0, \beta = 1/2$ Aitchison-Greenacre	$GV = \sum_{i,j} c_j r_i (\ln(p_{ij}/r_i))^2$ $- \sum_j c_j (\sum_{i=1}^I r_i \ln(p_{ij}/r_i))^2$	$\phi^2(0, 1/2) = \sum_{i,j} \left(\ln \frac{p_{ij}}{r_i c_j} \right)^2 r_i c_j.$
NSCA $\alpha = \beta = 1$ Lauro-D'Ambra	$GV = \sum_{i,j} \left(\frac{p_{ij}}{r_i} - c_j \right)^2 r_i$	$\phi^2(1, 1) = GV$

$\left(\frac{1}{\alpha} [(D_r^{-1} P D_c^{-1})^\alpha - 11'] \right)$ by $(I - c1')$. In fact, we can display columns in the same graph of rows without applying this post-multiplication. To do this, compute the SVD $(H_I Q)'(H_I A) = RDS'$, with D diagonal and H_I the unweighted $I \times I$ centering matrix. Then $(H_I Q) = (H_I A)RS$ and if we take principal coordinates $H_I A$ for the rows, and identify each column as the dummy row profile $(0, \ldots, 0, 1, 0, \ldots, 0)$, then the centered projection $B = H_J RS'$ provides standard coordinates for the columns [5], [6]. See also [8].

We next describe five methods for representing contingency tables: correspondence analysis (CA), Hellinger distance analysis (HD), non-symmetrical correspondence analysis (NSCA), the log-ratio alternative (LR), which only can be used for positive frequencies, and double-centered LR.

4.2 Correspondence Analysis (Centered = Uncentered)

$$D_r^{1/2}(D_r^{-1} P D_c^{-1} - 11')D_c^{1/2} = U \Lambda V' \qquad (\alpha = 1, \beta = 1/2).$$

1. Chi-square distance between rows: $\delta^2_{ii'} = \sum_{j=1}^J \left(\frac{p_{ij}}{r_i} - \frac{p_{i'j}}{r_{i'}} \right)^2 \frac{1}{c_j}$
2. To represent rows and columns: $A = D_r^{-1/2} U \Lambda, \quad B = D_c^{-1/2} V \Lambda.$
3. Decomposition of inertia: $\phi^2(1, 1/2) = GV$ (see Table 1).

This method can be justified under many different perspectives, see [10]. For a cumulative frequency approach, see [2].

4.3 Hellinger Distance Analysis (Centered and Uncentered)

C.: $D_r^{1/2}(I - 1r')(D_r^{-1/2} P^{1/2} D_c^{-1/2} - 11')D_c^{1/2} = U \Lambda V' \quad (\alpha = \beta = 1/2, \gamma = 1),$
U.: $D_r^{1/2}(D_r^{-1/2} P^{1/2} D_c^{-1/2} - 11')D_c^{1/2} = U \Lambda V' \quad (\alpha = \beta = 1/2, \gamma = 0).$

1. Hellinger distance between rows: $\delta^2_{ii'} = \sum_{j=1}^{J}(\sqrt{p_{ij}/r_i} - \sqrt{p_{i'j}/r_{i'}})^2$
2. To represent rows and columns: $A = D_r^{-1/2}U\Lambda, \quad B_* = D_c^{-1/2}V.$
3. Decomposition of inertia: $\phi^2(1/2, 1/2) \neq GV$ (see Table 1).

Note that $\sum_{i,j}\sqrt{p_{ij}r_ic_j})$ is the so-called affinity coefficient and that $GV < \phi^2(1/2, 1/2)$. See [17].

4.4 Nonsymmetrical CA (Centered = Uncentered)

$$D_r^{1/2}(D_r^{-1}PD_c^{-1} - 11')D_c = U\Lambda V' \qquad (\alpha = \beta = 1).$$

1. Distance between rows: $\delta^2_{ii'} = \sum_{j=1}^{J}\left(\frac{p_{ij}}{r_i} - \frac{p_{i'j}}{r_{i'}}\right)^2$
2. To represent rows and columns: $A = D_r^{-1/2}U\Lambda, \quad B = V\Lambda.$
3. Decomposition of inertia: $\phi^2(1, 1) = GV$ (see Table 1).

GV is related to the Goodman-Kruskal coefficient $\tau = \left[\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\frac{p_{ij}}{r_i} - c_j\right)^2 r_i\right]/t$, where $t = 1 - \sum_{i=1}^{I} r_i^2$. The numerator of τ represents the overall predictability of the columns given the rows. This solution is also related to BSS, WSS and TSS, see Sect. 3. Then $R^2 = BSS/TSS$ is a measure of association which coincides with τ. See [12, 13].

4.5 Log-Ratio Analysis (Centered and Uncentered)

$$\begin{aligned}
&\text{C.: } D_r^{1/2}(I - 1r')\ln(D_r^{-1}PD_c^{-1})D_c^{1/2} = U\Lambda V' \quad (\alpha = 0, \beta = 1/2, \gamma = 1),\\
&\text{U.: } D_r^{1/2}\ln(D_r^{-1}PD_c^{-1})D_c^{1/2} = U\Lambda V' \qquad\quad (\alpha = 0, \beta = 1/2, \gamma = 0).
\end{aligned}$$

1. Log-ratio distance between rows: $\delta^2_{ii'} = \sum_{j=1}^{J} c_j\left(\ln\frac{p_{ij}}{r_i} - \ln\frac{p_{i'j}}{r_{i'}}\right)^2$
2. To represent rows and columns: $A = D_r^{-1/2}U\Lambda, B_* = D_c^{-1/2}V\Lambda.$
3. Decomposition of inertia: $\phi^2(0, 1/2) \neq GV$ (see Table 1).

Note that $GV < \phi^2(0, 1/2)$.

4.6 Double Centered Log-Ratio Analysis

In LR analysis Lewi [14] and Greenacre [11] considered the weighted double-centered solution

$$D_r^{1/2}(I - 1r')\ln(D_r^{-1}PD_c^{-1})(I - 1c')'D_c^{1/2} = U\Lambda V',$$

called spectral map. The unweighted double-centered solution, called variation diagram, was considered by Aitchison and Greenacre, [1]. In this solution the role of rows and columns is symmetric and the distance is unweighted.

References

1. Aitchison, J., Greenacre, M.J.: Biplots of compositional data. Appl. Stat. **51**, 375–392 (2002)
2. Cuadras, C.M.: Correspondence analysis and diagonal expansions in terms of distribution functions. J. Stat. Plan. Inference **103**, 137–150 (2002)
3. Cuadras, C.M.: Distance-based multisample tests for multivariate data. In: Arnold, B.C. et al. (eds.) Advances in Mathematical and Statistical Modeling, pp. 61–71. Birkhauser, Boston, MA (2008)
4. Cuadras, C.M., Atkinson, R.A., Fortiana, J.: Probability densities from distances and discriminant analysis. Stat. Probab. Lett. **33**, 405–411 (1997)
5. Cuadras, C.M., Cuadras, D.: A parametric approach to correspondence analysis. Linear Algebra Appl. **417**, 64–74 (2006)
6. Cuadras, C.M., Cuadras, D., Greenacre, M.J.: A comparison of different methods for representing categorical data. Commun. Stat. Simul. Comp. **35**, 447–459 (2006)
7. Cuadras, C.M., Fortiana, J., Oliva, F.: The proximity of an individual to a population with applications in discriminant analysis. J. Classification **14**, 117–136 (1997)
8. Fichet, B., Gbegan, A.: Analyse factorielle des correspondances sur signes de présence-absence. In: Diday, E. et al. (eds.) Data Analysis and Informatics IV, pp. 209–219. Elsevier, Amsterdam (1985)
9. Goodman, L.A.: Correspondence analysis, association analysis, and generalized nonindependence analysis of contingency tables: saturated and unsaturated models, and appropriate graphical displays. In: Cuadras, C.M., Rao, C.R. (eds.) Multivariate Analysis: Future Directions 2, pp. 265–294. Elsevier, Amsterdam (1993)
10. Greenacre, M.J.: Theory and Applications of Correspondence Analysis. Academic Press, London (1984)
11. Greenacre, M.J.: Power transformations in correspondence analysis. Comput. Stat. Data Anal. **53**, 3107–3116 (2008)
12. Kroonenberg, P.M., Lombardo, R.: Nonsymmetric correspondence analysis: a tool for analysing contingency tables with a dependence structure. Multivariate Behav. Res. **34**, 367–396 (1999)
13. Lauro, N., D'Ambra, L.: L'analyse non symetrique des correspondances. In: Diday, E. et al. (eds.) Data Analysis and Informatics III, pp. 433–446. North Holland, Amsterdam (1984)
14. Lewi P.J.: Spectral mapping, a technique for classifying biological activity profiles of chemical compounds. Arzneim. Forsch. (Drug Res.) **26**, 1295–1300 (1976)
15. Light, R.J., Margolin, B.H.: An analysis of variance for categorical data. J. Am. Stat. Assoc. **66**, 534–544 (1971)
16. Rao, C.R.: Diversity: Its measurement, decomposition, apportionment and analysis. Sankhya A **44**, 1–21 (1982)
17. Rao, C.R.: A review of canonical coordinates and an alternative to correspondence analysis using Hellinger distance. Qüestiió **19**, 23–63 (1995)

One-Dimensional Preference Data Imputation Through Transition Rules

Luigi Fabbris

Abstract Preferences may be elicited with methods based either on pair comparison between items or ordering/sorting one or more items out of the given set. In both cases, the multivariate analysis of preferences requires that preferability is expressed for all pairs of items, so that an irreducible dominance matrix can be defined and mathematically processed. In this paper we present, apply and evaluate a new transition rule for the estimation of empty cells of a dominance matrix. The method was applied to preference data on students' guidance services. The new methodology showed to be more reliable than other methods in the literature.

1 Preference Elicitation in Surveys

The preferences of respondents for items listed in a survey questionnaire may be elicited in several ways. The elicitation method is to be suited to the type of survey, type of population and communication medium. In general, computer assisted interviewing systems offer excellent opportunities of item pairing, multimedia question administration and experiment embedding.

An advisable data collection method is *pair comparison*, which consists in administering items pair wise and asking the eligible respondents to select the most adequate item within each administered pair. So, if the list includes q items, $q(q-1)/2$ distinct pairs of items are to be administered. Other popular methods are those of picking the k ($k \geq 1$) preferred items from the list or ordering the q items of the list. The former method is prone to response error as q diverges and k is small; the latter one is difficult to respondents and can plausibly be administered to motivated respondents at particular settings. Whatever the data collection method, scholars share the feeling that no more than 4 or 5 items can be administered for an aware choice, even less if items are long sentences, without risking refusal to collaboration and inconsistency in responses. The only method that does not suffer from

L. Fabbris (✉)
Statistics Department, University of Padua, Padua, Italy
e-mail: luigi.fabbris@unipd.it

B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-13312-1_25,

list's length is the one-by-one rating of items. Nevertheless, this method presents the inconveniency of low discriminatory power.

In the following we deal with the method of pair comparisons among q items [3]. If q is large, the number of pairs diverges and the pair comparison procedure is no longer viable. Hence, if we are stuck with the pair comparison procedure, more compact methods have to be used in order to overcome this difficulty and preserve the possibility of score estimation in a one-dimensional setting [4]. These methods are aimed at reducing the number of pairwise comparisons and applying transition rules for imputing the preferences ignored at the data collection level.

A transition rule is a function which derives the unknown quantitative preference p_{ij} from p_{ik} and p_{kj} ($i \neq j \neq k = 1, \ldots, q$), where p_{ij} is the preference rate for item i over item j and p_{ik} and p_{kj} denote analogous preference rates. A transition rule may be applied in contingent situations in which the collected responses are inadequate for estimating the preferences between all possible pairs of items. For instance, suppose that a list of q items (q is even, without loss of generality) is divided in two equal parts for facilitating the respondents to order items of or pick items from a list. Thus just half of the possible relationships between the q items, i.e. $q(q-1)/4$, can be estimated directly with the obtained responses. The other half, given by the unmatched pairs, is to be estimated indirectly.

Transitivity rules depend on the researcher's hypothesis about the underlying relationship between preferences. In Sect. 2 we present transition rules that are popular in the scaling literature and introduce a new rule for between-items preference estimation in case of incomplete data. This allows us to estimate the items' scores following the method of principal eigenvalue extraction of the dominance matrix [10, 11].

We applied some transitions rules to a sample of 1,526 Padua University students. The data collection was based on an anonymous self-completion questionnaire. The items were included into 10 sets of university services. Since, for each set, q was divisible by 4, we created 6 distinct questionnaires by pairing all possible subsets of two quarters of an item set and administered each questionnaire to a random sub-sample of students. Thus, all items matched to each other at least once and a dominance matrix could have been filled with the preference estimates. For practical reasons, one of the questionnaires was not administered; therefore some of the planned pairs did not match.

The estimation procedure will be applied to unpublished data on student's prefernces for before-university guidance services. We discuss also the possibility to estimate preferences between items belonging to non-observed pairs in a multivariate analysis of preference framework. Results are presented in Sect. 3 and discussed in Sect. 4.

2 One-Dimensional Preference Rating Method

Let us suppose that the observed preferences between all possible pairs of q items be ordered in a $(q \times q)$ skew symmetric matrix $\mathbf{P} = \{p_{ij} = 1 - p_{ji}\ (i \neq j = 1, \ldots, q)\}$

where $p_{ij} \geq 0$ denotes the probability for item i to be preferred to item j (heretofore $i > j$) by n judges and $p_{ii} = 0$. If no permutation of rows and columns is possible that reduces $\mathbf{P}$ into a block matrix where one or more off-diagonal blocks are zero, matrix $\mathbf{P}$ is "irreducible". We ignore the trivial case $p_{ij} = 0.5$ for all $i \neq j$.

The Perron-Frobenius theorem states that a positive irreducible matrix, such as $\mathbf{P}$, can be associated to

- an eigenvalue λ_1, which is the largest real and positive root of the characteristic equation

$$\mathbf{P}\,\mathbf{w} = \lambda \mathbf{w}, \tag{1}$$

  with the uniqueness constraint $\mathbf{w}'\mathbf{w} = 1$ [10],
- a unique, simple and positive eigenvector $\mathbf{w}$ with which a normalized score of entry i ($i = 1, \ldots, q$) can be calculated: $w_i^* = w_i / \sum_i^q w_i$ such that $\sum_i^q w_i^* = 1$. The scores define the relative intrinsic position of the q items in the $0 \div 1$ interval (see [6]).

According to Parker [9] inequality, the largest eigenvalue in modulus of a positive matrix $\mathbf{P}= \{p_{ij}\}$ must verify:

$$\lambda_1 \leq \max(p_{1+} + p_{+1}, \ldots, p_{q+} + p_{+q})/2, \tag{2}$$

where p_{i+} and p_{+j} stand for i-th row and j-th column sum of $\mathbf{P}$: $p_{i+} = \sum_j^q p_{ij}$ and $p_{+j} = \sum_i^q p_{ij}$. In a skew symmetric matrix, the upper bound of λ_1 equals $(q-1)/2$ and can be attained in the trivial case $p_{i+} = p_{+i} = 0.5$ for all elements [5].

If some cells of matrix $\mathbf{P}$ are either empty, or some estimates are unreliable because based on a limited sample size, we may estimate the preferences with a transition function. A transition function for p_{ij} is an increasing function of the two (positive) preferences p_{ik} and p_{kj} ($i \neq j \neq k = 1, \ldots, q \geq 3$).

The function is undefined if $p_{ik} = 0$ and $p_{kj} = 1$ because k is absolutely preferred to both i and j so that no preference between i and j can be inferred [7].

If the transition condition is verified for all distinct i, j, k ($i \neq j \neq k = 1, \ldots, q$) items, $\mathbf{P}$ is of unit rank and λ_1 gets its minimum value. If this condition is verified, we expect either that $p_{ik} \leq p_{jk}$ or $p_{ik} \geq p_{jk}$ for all $1 \leq k \leq q$ if p_{ij} is to be estimated. This condition is more restrictive than the general conditions given by Jech [7]. Besides, cardinal consistency is not guaranteed even if $p_{ik} \leq p_{jk}$ or $p_{ik} \geq p_{jk}$ for all $1 \leq k \leq q$.

We will verify the effectiveness of the so-called "weak" rule [2]:

$$\text{if } (p_{ik}, p_{kj}) > 0.5 \Rightarrow \hat{\pi}_{ij} > 0.5, \tag{3}$$

which generates purely a ranking and implies that the only estimate we can produce for π_{ij} with (3) is 0.5. We will also apply the "moderate" rule:

$$\text{if } (p_{ik}, p_{kj}) \geq 0.5 \Rightarrow \hat{\pi}_{ij} \geq \min(p_{ik}, p_{kj}), \tag{4}$$

which in practice is estimated as: $\hat{\pi}_{ij} = \min(p_{ik}, p_{kj})$. The "strong" transition rule is:

$$\text{if } (p_{ik}, p_{kj}) \geq 0.5 \Rightarrow \hat{\pi}_{ij} \geq \max(p_{ik}, p_{kj}), \tag{5}$$

thus, even in this case, we estimate p_{ij} with its limiting value: $\hat{\pi}_{ij} = \max(p_{ik}, p_{kj})$.

The rules are symmetric with respect to 0.5 that is the indifference measure of preference. The symmetric weak rule is formulated as: if $(p_{ik}, p_{kj}) < 0.5 \Rightarrow \hat{\pi}_{ij} < 0.5$ but the estimate is again 0.5. The symmetric moderate and strong inequalities are reversed, i.e. the moderate rule becomes: if $(p_{ik}, p_{kj}) \leq 0.5 \Rightarrow \hat{\pi}_{ij} \leq \max(p_{ik}, p_{kj})$ and the strong one: if $(p_{ik}, p_{kj}) \leq 0.5 \Rightarrow \hat{\pi}_{ij} \leq \min(p_{ik}, p_{kj})$.

We suggest the adoption of a new transition rule that generates estimates beyond the intervals defined by p_{ik} and p_{kj}. It is an extension of rule (5). The rule, which we name "linear", partitions the interval between 1 and $\max(p_{ik}, p_{kj})$ if $p_{ik} \geq p_{kj} \geq 0.5$ and the interval between 0 and $\min(p_{ik}, p_{kj})$ if $p_{ik} \leq p_{kj} \leq 0.5$. If p_{ik} and p_{kj} are larger than 0.5, the "linear" estimate goes beyond p_{ik} and approaches one, proportionally to the distance between p_{kj} and 0.5, the indifference condition. The larger this distance, the larger the difference between "strong" and "linear" estimates. If $p_{ik} \geq p_{kj} \geq 0.5$ the rule for $\pi_{ij} = \Pr(i > j)$ estimation is:

$$\hat{\pi}_{ij} = p_{ik} + (1 - p_{ik})\frac{p_{kj} - 0.5}{0.5}, \tag{6}$$

or equivalently:

$$\hat{\pi}_{ij} = p_{ik} + 2(1 - p_{ik})(p_{kj} - 0.5). \tag{7}$$

Symmetrically, if $p_{ik} \leq p_{kj} \leq 0.5$, the estimate lies between 0 and p_{ik} and is:

$$\hat{\pi}_{ij} = 2p_{ik}p_{kj}. \tag{8}$$

The linear transition rule is not eligible if either $p_{ik} > 0.5$ and $p_{kj} < 0.5$ or $p_{ik} < 0.5$ and $p_{kj} > 0.5$. These occurrences imply intransitivity and can be considered inconsistent expressions of preferences.

A dominance matrix whose λ_1 reaches its maximum value, $(q - 1)/2$, is fully intransitive. So, we can define an index of transitivity based on the main eigenvalue:

$$I_t' = \frac{(q - 1)/2 - \lambda_1}{q - 1/2} = 1 - \frac{2\lambda_1}{q - 1}, \tag{9}$$

that becomes larger as the proportion of transition occurrences diverges. Its minimum value equals 0. Another index may be based on inconsistent transitions, $n_{\bar{c}}$, counted over the whole $q(q - 1)$ preferences and adjusted for the number t of ties (i.e. number of times $p_{ij} = 0.5$ in the dominance matrix $\mathbf{P}$):

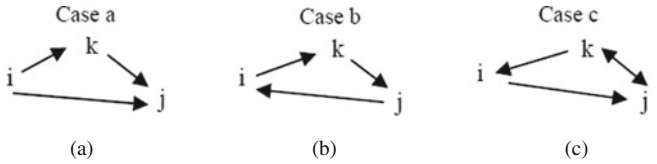


Fig. 1 Transitive **(a)**, intransitive **(b)** and tied **(c)** relations between units i, j, k

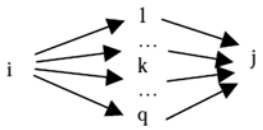


Fig. 2 Paths from i to j through the other q-2 units of the examined set

$$I_t'' = 1 - \frac{n_{\bar{c}}}{\max(n_{\bar{c}})} = 1 - \frac{n_{\bar{c}}}{q[(q-1)(q-2)-t]}, \tag{10}$$

which varies in the $0 \div 1$ interval and may be expressed in percent (Fig. 1).

The number of possible transitions is the number of possible paths through which p_{ij} can be reached. This number depends on q and t, being the number of all distinct paths through any k-th element ($k = 1, \ldots, q$) different from i-th and j-th (Fig. 2), ties excluded.

The conditions $p_{ik} \geq p_{kj} \geq 0.5$ and $p_{ik} \leq p_{kj} \leq 0.5$ imply that any linear algorithm based on the unconditional product of p_{ik} and p_{kj} is not viable. In fact, if both preference probabilities are larger than 0.5, we expect that p_{ij} be larger than the larger of the two probabilities (and of course larger than 0.5). This favours the "strong" transition rule with respect to the moderate one. Moreover, the weak transition rule is trivial because 0.5 represents the ignorance position about population preferences since the two conditions $k > j$ and $j > k$ are equivalent.

We compare the above-described rules also with another estimator based on the product of p_{ik} and p_{kj} [7]:

$$\hat{\pi}_{ij} = \frac{p_{ik}\, p_{kj}}{p_{ik}\, p_{kj} + p_{ki}\, p_{jk}}. \tag{11}$$

The estimate of π_{ij} is averaged over all c ($c = 1, \ldots, q-2$) consistent transition estimates, $\hat{\pi}_{ij}(k)$, so to improve the estimate's stability:

$$\bar{\hat{\pi}}_{ij} = \frac{1}{c} \sum_{k}^{c} \hat{\pi}_{ij}(k). \tag{12}$$

If all $\hat{\pi}_{ij}(k)$ are inconsistent, $\bar{\hat{\pi}}_{ij} = 0.5$. If we refer to a known p_{ij}, we can evaluate the unreliability of estimates with an absolute loss function (see also [8]:

$$L_1 = E|\bar{\hat{\pi}}_{ij} - p_{ij}|\,, \tag{13}$$

where $E(.)$ stands for the expected value of the argument, or with an Euclidean one:

$$L_2 = \sqrt{E(\tilde{\hat{\pi}}_{ij} - p_{ij})^2}\,. \tag{14}$$

3 An Application

Four cells of the dominance matrix were empty because of incompleteness of the data collection process (Table 1). There are 4 ways for estimating each of the missing preferences. For instance, for estimating p_{37}, we can "transit" through and then average 4 couples of the observed preferences: $p_{31} \cap p_{17}$; $p_{32} \cap p_{27}$; $p_{35} \cap p_{57}$; $p_{36} \cap p_{67}$. The first, the second and the fourth transitions are eligible; the third is a limiting case because $p_{57} = 0.5$.

The estimates in the four upper-triangle empty cells are computed with rules (4), (5), (6) and (11) and then averaged. The estimation through rule (3) gives always 0.5 and can be considered a trivial case. Symmetric preferences are one minus the given estimates. We can found 5 inconsistencies and 11 valid estimates. The average estimates of the four unknown preferences are presented in Table 2 (grey cells).

For measuring the reliability of the examined methods we simulated the absence of the expressed preferences and then compared the average of the transition-based estimates with the observed preferences through formulae (13) and (14). The estimates of preferences are presented in Table 2 and their evaluative statistics in Table 3.

Formulae (5), (6) and (11) give preference and efficiency estimates that are similar to each other and are different from those obtainable with formula (4). All this indicates a clear preferability for stronger transitivity rules if preferences are to be indirectly estimated. The linear and the ratio methods are the most reliable ones; the former performs slightly better than the latter. The transitivity occurrences are slightly more than one third for the three functions (5), (6) and (11), lower than that of the response-based matrix (0.463). Nevertheless, if we ignore ties (i.e. $p_{hk} = 0.5$ for any h and k), the three transition rules would perform somewhat worse than the moderate one (0.391) but better than directly observed preferences (0.349). Hence,

Table 1 Dominance matrix between 8 before-university guidance services according to Padua University (Italy) students, 2007

Item	1	2	3	4	5	6	7	8
1	0	0.461	0.275	0.488	0.488	0.617	0.558	0.847
2	0.539	0	0.353	0.568	0.500	0.622	0.597	0.891
3	0.725	0.647	0	0.714	0.566	0.717		
4	0.512	0.432	0.286	0	0.327	0.495		
5	0.512	0.500	0.434	0.673	0	0.578	0.500	0.830
6	0.383	0.378	0.283	0.505	0.422	0	0.532	0.847
7	0.442	0.403			0.500	0.468	0	0.804
8	0.153	0.109			0.170	0.153	0.196	0

Table 2 Estimates of the preferences between before-university guidance services at Padua University (1st within-cell value: rule (4); 2nd value rule: (5); 3rd value: rule (6), 4th value: rule (9))

Item	2	3	4	5	6	7	8
1	0.494	0.479	0.505	0.496	0.490	0.516	0.588
	0.460	0.358	0.617	0.449	0.488	0.553	0.826
	0.455	0.343	0.621	0.446	0.483	0.565	0.855
	0.454	0.339	0.621	0.445	0.483	0.568	0.868
2		0.500	0.503	0.500	0.520	0.524	0.565
		0.434	0.648	0.597	0.598	0.560	0.832
		0.434	0.649	0.597	0.612	0.579	0.854
		0.434	0.650	0.597	0.615	0.583	0.864
3			0.546	0.500	0.602	0.547	0.664
			0.679	0.647	0.650	0.664	0.854
			0.710	0.647	0.722	0.693	0.901
			0.719	0.647	0.734	0.702	0.918
4				0.498	0.512	0.420	0.512
				0.427	0.617	0.529	0.847
				0.425	0.626	0.448	0.851
				0.425	0.628	0.448	0.853
5					0.504	0.515	0.523
					0.569	0.578	0.847
					0.572	0.590	0.854
					0.573	0.592	0.876
6						0.500	0.532
						0.422	0.804
						0.422	0.817
						0.422	0.823
7							0.500
							0.830
							0.830
							0.830

Table 3 Distance between observed and estimated preferences in the dominance matrix **P** presented in Table 1 by type of transitivity rule and loss function

Loss function	Moderate	Strong	Linear	Product
Mean absolute distance	0.124	0.061	0.055	0.057
Euclidean distance	0.164	0.073	0.071	0.072
Transitivity index (formula 8)	0.391	0.357	0.369	0.369

measures of goodness of preference fitting and transitions' regularity are not necessarily coherent.

4 Conclusions

In this paper we put forward a new method for estimating preference data between couples of items, grounded on a transitivity rationale. The mathematical characteristics of the method do not allow us to state that our method is definitely superior

to others (see also the pertinent Arrow's impossibility theorems: [1]). That is why we applied several rules to a dataset and showed that our new method is fairly more reliable than other known methods for estimating preference probabilities. This brings us to state that our approach is consistent with the idea of modelling the observed preferences and guessing transition rules for the estimation of unmeasured preferences and even for a second stage "smoothing" of preferences.

Of course, more comparisons with other transition estimators could improve our considerations and other empirical analyses will help researchers to detect the empirical superiority of either methods.

Acknowledgments This work was supported by a grant from the Italian Ministry of Education, University and Research as well as the University of Padua (PRIN 2007).

References

1. Arrow, K.J.: Social Choice and Individual Values, 2nd ed. Wiley, New York, NY (1963)
2. Coombs, C.H.: A Theory of Data. Mathesis Press, Ann Arbor, MI (1976)
3. David, H.A.: The Method of Paired Comparisons. 2nd ed. Oxford University Press, New York, NY (1988)
4. Fabbris, L.: Dimensionality of scores obtained with a paired-comparison tournament system of questionnaire items. In: Palumbo, F. Lauro, C.N., Greenacre, M.J. (eds.) Data Analysis and Classification. Proceedings of the 6th Conference of the Classification and Data Analysis Group of the Società Italiana di Statistica, pp.155–162. Springer, Berlin Heidelberg (2010)
5. Genest, C. Lapointe, F., Drury, S.W.: On a proposal of Jensen for the analysis of ordinal pairwise preferences using Saaty's eigenvector scaling method. J. Math. Psychol. **37**(4), 575–610 (1993)
6. Horn, R.A. Johnson, C.R.: Matrix Analysis. Cambridge University Press, New York, NY (1985)
7. Jech, T.: A quantitative theory of preferences: some results on transition functions. Soc. Choice Welfare **6**, 301–314 (1989)
8. May, K.O.: Intransitivity, utility and the aggregation of preference patterns. Econometrica **22**(1), 1–13 (1954)
9. Parker, W.V.: The characteristic root of a matrix. Duke Math. J. **3**, 484–487 (1937)
10. Saaty, T.L.: A scaling method for priorities in hierarchical structures. J. Math. Psychol. **15**, 234–281 (1977)
11. Saaty, T.L.: Rank according to Perron: a new insight. Math. Mag. **60**(4), 211–213 (1987)

About a Type of Quasi Linear Estimating Equation Approach

Giulio D'Epifanio

Abstract In this work, a type of quasi-linear system is presented, which is able to identify the "true" value of parameter-profile in the setup of "generalized linear mixed models". A type of quasi-linearization of the link function is used, which would preserve basic sampling properties of conditioned moments of the random latent profile. Then, an approach is outlined in estimating. It uses a weighted quasi-linear estimating system which is exactly unbiased. Due to quasi-linearization, it might be solved by using easy-to-implement recursive procedures.

1 Introduction

It is well known that, except for the linear mixed models, the full maximum-likelihood(FML) function is analytically intractable within the setup of "generalized linear mixed models" (GLMM), whenever the system of random effects is multi-dimensional, because random effects enter the model non-linearly. Therefore in estimation, direct FML-based approaches, which typically implement Gauss-Hermite-like quadrature formulas, are burdensome although adaptive quadrature rules [9] would promise to overcome some problems. On the other hand, simulation-based approaches (like Gibbs sampling or MCMC) will be difficult to implement as unitary-task (e.g. see [12]) in general-purpose statistical packages. However, practitioners would feel trouble in specifying unambiguous stopping-rules and in interpreting results based on convergence-diagnostic tools. Thus, in literature (e.g. for a review, see [10, 11]), several approximate approaches were proposed and compared, which were based on Taylor's expansion of the link-function. Unfortunately, the comparative study of Rodriguez and Goldman [11] raised serious doubts, whenever the random-effects are multi-dimensional and sizeable, in using

G. D'Epifanio (✉)
Faculty of Political Science, University of Perugia, 06100 Perugia, Italy
e-mail: ggiulio@stat.unipg.it

B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-13312-1_26,

approximate approaches.[1] The basic problem was that their underlying estimating systems are structurally biased. Therefore, researches on general approaches in tackling both the "non linearity" and the "multi-dimensionality" should deserve further attention.

In this work, our primary purpose is to communicate a, perhaps little known, theoretical note on which estimation might be based. This note concerns a certain type of "quasi-linear" structural system, which might be used to identify the "true" value of a parameter-profile within the GLMM setup. This system uses certain quasi-linear operators which, in a certain sense, would mimic the (1st and 2nd order) conditional implicit full-moments of the latent profile of random parameters given data (formally, these would be Bayes rules). This note says that, using these quasi-linear operators, at the "true" value of parameters, the universal sampling properties of the (1st and 2nd order) full-conditional implicit moments would be exactly preserved. Thus, this note is a theoretical foundation on which exactly unbiased types of weighted quasi-linear-estimating-systems may be developed. In practice, using quasi-linearization, the structure of certain procedures and formulas might be generalized from the Gaussian-linear case to the very larger setup of GLMM. Furthermore, this structural-linearity might be useful in developing easy-to-implement and effective estimation procedures.

In Sect. 2, a reference model is presented in structural GLMM format. In Sect. 3, a type of quasi-linearization is proposed, certain quasi-linear-Bayes operators are defined and the main note is presented. In Sect. 4, statistical estimating is outlined. In Sect. 5, an empirical comparative study is reported.

2 The Reference Model in GLMM

Without a practical and theoretical loss of generality, in order to delineate certain objects and concepts, we re-formulate the "three-level logit-normal model" of Rodriguez and Goldman ([11], pp. 340–341) in general structural-GLMM format (1)-(2). There, compact notations are used, which are adapted to hierarchical structure. The realized observable full-outcome profile $x := (x_k, k := 1, \ldots, n)$ is partitioned on n top-level data-slices, which are structured on hierarchy of clusters so that $x_k := (x_{k[ji]},\ [ji] := 1, \ldots, p_k)$ denotes the k-th top-level data-slice, $x_{kj} := (x_{k[ji]},\ i := 1, \ldots, p_{kj})$ his sub data-slice kj. Here, $[ji]$ denotes the complex index such that $x_{k[ji]} := x_{kji}$, which runs over the p_k 1st-level units within top-level data-slice k; indexes j and i run, respectively, over the q_k 2nd-level units and the p_{kj} 1st-level units within sub-data-slice kj of data-slice k.

[1] Implementing the three-level logit-normal model over reference-data sets, this study shows that all approximate first order (the "marginal quasi-likelihood" and the "first order penalized quasi-likelihood") procedures have led to a substantial underestimation of the fixed and random effects. But, approximate second order procedures (the "second order marginal quasi-likelihood" and the "second order penalized quasi-likelihood") were numerically unstable, actually failing convergence. Even Gibbs sampling experienced problems concerning convergence.

$$X_{k[ji]} \mid \theta_{k[ji]}, 1 \overset{indep.}{\sim} Bin(\theta_{k[ji]}, 1), \tag{1}$$

$$\theta_{k[ji]}(\eta_k) := \exp(\eta_{k[ji]})(1 + \exp(\eta_{k[ji]}))^{-1},$$

$$\eta_k \mid m_k, \Sigma_k \overset{indep.}{\sim} \Phi_{p_k}(m_k(\beta), \Sigma_k(\tilde{\Sigma}_k)), m_k(\beta) := Z_k\beta, \tag{2}$$

$$Vec(\Sigma_k(\tilde{\Sigma}_k(\zeta)) := (W_k \otimes W_k)Vec(\tilde{\Sigma}_k(\zeta)) = (W_k \otimes W_k)\mathcal{M}_k\zeta$$

$$k := 1, \ldots, n, \quad [ji] := 1, \ldots, p_k.$$

In (1), the observable outcome profile x_k is a realization of a set of conditionally independent binary random outcomes $X_{k[ji]}$ from the binomial distribution $Bin(\theta_{k[ji]}, 1)$ (for a single observation) with parameter $\theta_{k[ji]} = Pr(X_{k[ji]} = 1)$. The latent profile $\eta_k := (\eta_{k[ji]}, \ [ji] := 1, \ldots, p_k)$ is a realization from the p_k-variate-normal $\Phi_{p_k}(m_k, \Sigma_k)$ with expectation m_k and matrix of variance-covariance Σ_k. In (2), structural parameters (m_k, Σ_k) are constrained, by means of specific parameters β and $\tilde{\Sigma}_k$, on a linear manifold whose points are represented by coordinate-parameters $\gamma := (\beta, \zeta)$. Here,[2] β is the vector of fixed effects, ζ is the parameter-profile which enters the (vectorized by using the standard Kronecker's product) variance-covariance matrix $Vec\ (\tilde{\Sigma}_k(\zeta)) := \mathcal{M}_k\zeta$ of random effects, where $\mathcal{M}_k$ is a proper matrix operator which is specific for the k-th top-level unit; Z_k and W_k denote proper design matrices.

---

[2] Recall the "three-level logit-normal model" ([11], pp. 340-341):

$$X_{kji}|\theta_{kji}, 1 \overset{indep.}{\sim} Bin(\theta_{kji}, 1), \quad logit(\theta_{kji}) = Z_{kji}\beta + \varepsilon^{\{2\}}_{kj} + \varepsilon^{\{3\}}_{k}.$$

Here, $\varepsilon^{\{2\}}_{kj}$ and $\varepsilon^{\{3\}}_{k}$ denote, respectively, 2nd and 3rd level (centered on zero) Gaussian random effects with variance, respectively, σ^2_2 and σ^2_3. Rewrite model in matrix form as follows:

$$logit(\theta_k) = Z_k\beta + \Gamma^{\{1\}}_k \begin{bmatrix} \varepsilon^{\{2\}}_{k1} \\ ..... \\ \varepsilon^{\{2\}}_{kq_k} \end{bmatrix} + \Gamma^{\{1\}}_k\Gamma^{\{2\}}_k\varepsilon^{\{3\}}_k = Z_k \begin{bmatrix} \beta^{\{1\}} \\ \beta^{\{2\}} \\ \beta^{\{3\}} \end{bmatrix} + W_k \begin{bmatrix} \begin{bmatrix} \varepsilon^{\{2\}}_{k1} \\ ..... \\ \varepsilon^{\{2\}}_{kq_k} \end{bmatrix} \\ \varepsilon^{\{3\}}_k \end{bmatrix}.$$

Here, the profile $\beta := (\beta^{\{1\}}, \beta^{\{2\}}, \beta^{\{3\}})$ of fixed effects is partitioned across levels; $Z_k := [Z^{\{1\}}_k, \Gamma^{\{1\}}_k Z^{\{2\}}_k, \Gamma^{\{1\}}_k\Gamma^{\{2\}}_k Z^{\{3\}}_k]$ and $W_k := (\Gamma^{\{1\}}_k, \Gamma^{\{1\}}_k\Gamma^{\{2\}}_k)$ are the design matrices. Note that $Z^{\{1\}}_k$, $Z^{\{2\}}_k$ and $Z^{\{3\}}_k$ are design sub-matrices (including level-specific covariates) which are specific, respectively, for the 1st, 2nd and 3rd level; $\Gamma^{\{1\}}_k$ and $\Gamma^{\{2\}}_k$ are grouping matrix operators, respectively, for the 1st and 2nd level units ($\Gamma^{\{1\}}_k$ groups 1st level units according to the 2nd level unit at which they belong; $\Gamma^{\{2\}}_k$ groups 2nd level units within their common 3rd level unit). Assume that random effects $(\varepsilon^{\{2\}}_{k1}, \ldots, \varepsilon^{\{2\}}_{kq_k})$ are independent and that, for any k and $j := 1, \ldots, q_k$, $\varepsilon^{\{2\}}_{kj}$ and $\varepsilon^{\{3\}}_k$ also are independent. Then, the variance-covariance matrix of system of random effect is

$$\tilde{\Sigma}_k(\zeta) := Var\left[\begin{bmatrix} \begin{bmatrix} \varepsilon^{\{2\}}_{k1} \\ ..... \\ \varepsilon^{\{2\}}_{kq_k} \end{bmatrix} \\ \varepsilon^{\{3\}}_k \end{bmatrix}\right] = \begin{bmatrix} diag(\underbrace{\sigma^2_2, \ldots, \sigma^2_2}_{q_k}) & 0 \\ 0 & \sigma^2_3 \end{bmatrix}, \text{ where } \zeta := (\sigma^2_2, \sigma^2_3).$$

Estimation concerns the full coordinate-parameter profile $\gamma := (\beta, \zeta)$. Identifiability of coordinate-parameters is assumed, at least locally, by supposing proper conditions on design matrices.

3 Quasi-Linearization and the Main Note

Let $M_k(m_k, \Sigma_k) := E(\Theta_k; m_k, \Sigma_k)$ and $S_k(m_k, \Sigma_k) := Var(\Theta_k; m_k, \Sigma_k)$ denote, respectively, the vector of expectations and the matrix of variance-covariance, which are associated to the random profile Θ_k which will take values θ_k in system (1). Let us define the following transformation:

$$\theta_k^{QL}(\eta_k; m_k, \Sigma_k) := M_k(m_k, \Sigma_k) + D\theta_k|_{\theta_k^{-1}(M_k(m_k, \Sigma_k)(\gamma))}(\eta_k - m_k) \tag{3}$$

Here, transformation (3) linearizes[3] $\theta_k(\eta_k)$ by using the tangent of function $\theta_k(\eta_k)$ at the point $\theta^{-1}(M_k(m_k, \Sigma_k))$, and then re-centering it at $M_k(m_k, \Sigma_k)$, rather than at m_k as usual Taylor expansion would do. By reversing now (3) with respect to the profile η_k, the *quasi-linearized(QL-) recovery* of η_k is provided by

$$\eta_k^{QL}(\theta_k; m_k, \Sigma_k) := (D\theta_k)^{-1}_{|_{M_k(m_k, \Sigma_k)}}(\theta_k - M_k(m_k, \Sigma_k)) + m_k \tag{4}$$

We will see later that, using QL-transform (4), some objects (expectations, variances, etc.) could be reported from the "natural" (but, "strongly non linear") scale of θ_k ($\theta_k \in [0, 1]$) to that ("quasi-linear") of η_k^{QL} ($\eta_k^{QL} \in [-\infty, +\infty]$), which would approximate the original ("linear") scale of profile η_k. Notice here centering: for any mixing distribution, which is associated to parameters (m_k, Σ_k) in system (1), $E(\eta_k^{QL}(\Theta_k); m_k, \Sigma_k) = E(\eta_k; m_k, \Sigma_k) = m_k$. Emphasize also the structural-generality of QL-transform: quasi-linearization (4) is compatible with any strictly monotone sufficiently regular function, whenever it was used in (1–2) alternatively to the logit-function.

Let us define now the following QL-objects:

$$\hat{x}_k(m_k, \Sigma_k) := (D\theta_k)^{-1}_{|_{M_k(m_k, \Sigma_k)}}(x_k - M_k(m_k, \Sigma_k)) + m_k,$$

$$\hat{\Sigma}_k(m_k, \Sigma_k) := Var(\eta_k^{QL}(\Theta_k); m_k, \Sigma_k) = (D\theta_k)^{-1}_{|_{M_k}} \cdot Var(\Theta_k; m_k, \Sigma_k) \cdot (D\theta_k)^{-1}_{|_{M_k}}$$

$$\hat{\Lambda}_k := (D\theta_k)^{-1}_{|_{M_k}} \cdot \Lambda_k \cdot (D\theta_k)^{-1}_{|_{M_k}}, \text{ where } \Lambda_k$$

$$:= \int Var(X_k \mid \theta_k(\eta_k)) \cdot \Phi_{p_k}(\eta_k; m_k, \Sigma_k) d\eta_k$$

$$\hat{R}_k := \hat{\Sigma}_k \cdot (\hat{\Lambda}_k + \hat{\Sigma}_k)^{-1}.$$

[3] here, $D\theta_k := \frac{\partial \theta_k(\eta_k)}{\partial \eta_k}$ is the derivative matrix of the (p_k-)vector-valued anti-link function $\theta_k(\eta_k)$, which is diagonal because $\theta_{ki}(\eta_{k1}, \ldots, \eta_{kp_k}) = \theta_{ki}(\eta_{ki}),\ i := 1, \ldots, p_k$.

Here, $\hat{x}_k(m_k, \Sigma_k)$ denotes the "artificially adjusted" data set, which is temporarily QL-recovered from the actual data set x_k, given mixing density at (m_k, Σ_k); $\hat{\Sigma}_k(m_k, \Sigma_k)$ denotes the variance-covariance matrix of η_k^{QL}.

Now, a notable consequence of quasi-linearization (4) is that expectation of random profile $\hat{X}_k(m_k, \Sigma_k)$ is exactly centered on profile m_k while its variance-covariance matrix is $(\hat{\Lambda}_k + \hat{\Sigma}_k)$ (for technical details, see [2, 3]. Finally, we could define now the "quasi-linear" operators $T_{x_k}^{QLB}$ and $V_{x_k}^{QLB}$ as follows:

$$T_{x_k}^{QLB}(m_k, \Sigma_k) := \hat{R}_k \cdot (\hat{x}_k - m_k) + m_k,$$
$$V_{x_k}^{QLB}(m_k, \Sigma_k) := \hat{\Lambda}_k(\hat{\Lambda}_k + \hat{\Sigma}_k)^{-1} \cdot \hat{\Sigma}_k + \Delta T_{x_k}^{QLB} \otimes (\Delta T_{x_k}^{QLB})^{tr}.$$

Here, $$\Delta T_{x_k}^{QLB} := T_{x_k}^{QLB}(m_k, \Sigma_k) - m_k = \hat{R}_k \cdot (\hat{x}_k - m_k).$$

By construction, quasi-linear operators $T_{x_k}^{QLB}$ and $\hat{\Lambda}_k(\hat{\Lambda}_k + \hat{\Sigma}_k)^{-1}\hat{\Sigma}_k$ have a structure which would imitate that, respectively, of the "linear Bayes expectation" and of the "linear Bayes variance-covariance matrix" (see [6]). Therefore, they might be conventionally referred as QL-Bayes rules. Thus, $\hat{R}_k$ would mimic the shrinkage factor. In particular, provided that responses X_k were Gaussian and the link-function was the identity-function, QL-operators $T_{x_k}^{QLB}$ and $V_{x_k}^{QLB}$ would exactly coincide with their full-Bayes counterparts. As a consequence of "quasi-linearization", we would emphasize now the following note.

The main note. *Let us assume model (1) and (2) and let $\gamma_0 := (\beta_0, \zeta_0)$ denote the "true" value-profile. Then, for any γ_0, the following system*

$$E_{(m_k, \Sigma_k)(\gamma_0)}[T_{X_k}^{QLB}(m_k, \Sigma_k)(\tilde{\gamma})] = m_k(\tilde{\gamma}), \tag{5}$$
$$E_{(m_k, \Sigma_k)(\gamma_0)}[Vec\ V_{X_k}^{QLB}(m_k, \Sigma_k)(\tilde{\gamma})] = Vec\ \hat{\Sigma}_k(m_k, \Sigma_k)(\tilde{\gamma}),$$
$$k := 1, \ldots, n$$

is solved, with respect to the unknown profile $\tilde{\gamma}$, by $\tilde{\gamma} = \gamma_0$ itself.

Here, $E_{(m_k, \Sigma_k)(\gamma_0)}[.]$ denotes expectation, over the sample space of X_k, in system (1)-(2), whenever the true-value profile γ was set at γ_0. This note (for technical details, see [2, 3] might be used to recursively characterize the "true" value-profile within the setup (1) and (2). Then, as a practical consequence of "quasi-linearity" in estimating, types of exactly unbiased weighted QL-estimating-systems may be developed, which would be relatively easy to solve.

4 Estimating

Without claim to completeness, we would like to outline here developments of quasi-linearization in estimating.

4.1 A General Quasi-Linear Estimating System

Let $\Delta_{x_k}(\gamma) := \begin{bmatrix} \Delta T_{x_k}^{QLB}(m_k, \Sigma_k)(\gamma) \\ \Delta Vec V_{x_k}^{QLB}(m_k, \Sigma_k)(\gamma) \end{bmatrix}$ $:= \begin{bmatrix} T_{x_k}^{QLB}(m_k, \Sigma_k)(\gamma) - m_k(\gamma) \\ Vec V_{x_k}^{QLB}(m_k, \Sigma_k)(\gamma) - Vec\hat{\Sigma}_k(m_k, \Sigma_k)(\gamma) \end{bmatrix}$ denote the vector of variations of QL-operators on the mixing density which is associated to the point (m_k, Σ_k) at coordinate $\gamma := (\beta, \zeta)$ of the linear manifold (2). Given any realization of profiles $X_1, \ldots, X_k, \ldots, X_n$ from stochastic system (1-2), consider the following QL-system:

$$\sum_{k=1}^{n} \left[\frac{\partial}{\partial \gamma}(m_k, Vec(\Sigma_k))(\gamma) \right]^{tr} \cdot [\mathcal{W}_k^{-1}((m_k, \Sigma_k)(\gamma)) \cdot \Delta_{x_k}(\gamma)] = \mathbf{0}. \quad (6)$$

This system would combine, using convenient weights, independent top-cluster-specific estimating-functions [1, 5], both for expectations and variance-covariances, which are specified according to working conjectures over space of constrained parameters. Here, $\mathcal{W}_k^{-1}((m_k, \Sigma_k)(\gamma))$ denotes a generic symmetric (positive definite or semi-definite) weighting matrix. Let h denote the dimension of full parameter profile γ, $\left[\frac{\partial}{\partial \gamma}(m_k, Vec(\Sigma_k))(\gamma) \right]$ denote the $((p_k + p_k \times p_k) \times h)$ matrix whose columns are the coordinate vectors which are tangent to linear manifold (2).

Mimicking the intrinsic-recursive characterization of the "true" profile γ_0 in the main note, a solution γ^* of (6) might be operationally characterized as follows: *search for that profile γ^*, which represents point $p(\gamma^*)$ on linear manifold (2), such that the full profile of weighted variations $(\mathcal{W}_k^{-1}\Delta_{x_k}(\gamma), k := 1, .., n)$ would be orthogonal to manifold (2)* at $p(\gamma^*)$.

As a consequence of the main note, we could realize that system (6) is exactly unbiased regardless of the weighting system. Therefore, recalling the general theory of estimating equation,[4] the main requirement would satisfied here in order to assure that there exists a sequence of solutions of (6) which converges (under proper conditions, in some statistical sense) to the true value γ_0, provided model (1–2) was correctly specified. Afterwards, we propose to estimate the value γ_0, of true profile in (1–2), by searching for $\gamma^* := (\beta^*, \zeta^*)$ values such that γ^* solves estimating system (6).

4.2 A Decomposable Estimating System

Consider now a weighting system such that, for generic top-level cluster k, the weighting-matrix has the following block partitioned structure:

[4] Technically, solutions of (6) would be M-estimators [7]. Then, under the conditions that were given in Huber ([7], pag. 131) there exists a sequence which is consistent and asymptotically normal.

$$\mathcal{W}_k(\gamma) := \begin{pmatrix} \mathcal{S}_k & \mathbf{0} \\ \mathbf{0} & \tilde{\mathcal{S}}_k \otimes \tilde{\mathcal{S}}_k \end{pmatrix}(\gamma). \tag{7}$$

Here, $\mathcal{S}_k$ and $\tilde{\mathcal{S}}_k$ denote symmetric, positive definite or positive semi-definite $(p_k \times p_k)$ matrices. Although in principle different choices are admissible,[5] this structure of weighting seems sufficiently general to adapt to concrete situations of interest. It seems conceptually difficult to conceive and justify mixed weights for expectations and variance-covariances specific estimating functions. Then, using weighting structure (7), system (6) might be separated in the following two inter-crossed estimating sub-systems:

$$\sum_{k=1}^{n} Z_k^{tr} \cdot \mathcal{S}_k(\beta, \zeta) \cdot \hat{R}_k(\beta, \zeta) \cdot (\hat{x}_k(\beta, \zeta) - Z_k\beta) = \mathbf{0}, \tag{8}$$

$$\sum_{k=1}^{n} \{(W_k \otimes W_k) \cdot \mathcal{M}_k)^{tr} \cdot (\tilde{\mathcal{S}}_k \otimes \tilde{\mathcal{S}}_k)(\beta, \zeta) \cdot Vec\{V_{x_k}^{QLB}(m_k, \Sigma_k) - \hat{\Sigma}_k(m_k, \Sigma_k)\}(\beta, \zeta) = \mathbf{0} \tag{9}$$

Consider now matrices $F_k(\beta, \zeta)$ and $G_k(\beta, \zeta)$ such that[6] $\hat{\Sigma}_k(\beta, \zeta) = F_k(\beta, \zeta) \cdot \Sigma_k(\zeta) \cdot G_k^{tr}(\beta, \zeta)$. Then, using some matricial algebra (for details, see [2, 3]), system (8) and (9) might be rewritten as the following fixed-point system:

$$\beta = \left\{ \sum_{k=1}^{n} Z_k^{tr} \cdot (\mathcal{S}_k \hat{R}_k)(\beta, \zeta) \cdot Z_k \right\}^{-1} \left\{ \sum_{k=1}^{n} Z_k^{tr} (\mathcal{S}_k \hat{R}_k)(\beta, \zeta) \cdot \hat{x}_k(\beta, \zeta) \right\} \tag{10}$$

$$\zeta = \left\{ \sum_{k=1}^{n} [\mathcal{M}_k^{tr} \cdot (W_k^{tr} \tilde{\mathcal{S}}_k G_k W_k) \otimes (W_k^{tr} \tilde{\mathcal{S}}_k \hat{R}_k F_k W_k)](\beta, \zeta) \cdot \mathcal{M}_k \right\}^{-1} (\beta, \zeta) \cdot \left\{ \sum_{k=1}^{n} \mathcal{M}_k^{tr} [(W_k^{tr} \tilde{\mathcal{S}}_k) \otimes (W_k^{tr} \tilde{\mathcal{S}}_k)] \cdot Vec\,(\Delta T_{x_k}^{QLB} \otimes (\Delta T_{x_k}^{QLB})^{tr}) \right\} (\beta, \zeta) \tag{11}$$

Therefore, estimations are implicitly identified by solutions of fixed-point system (10)-(11), which may be recursively solved.

---

[5] In practice, choice of weighting system should adhere to specific goals. For instance, in finite population sampling, weighting might be used to contrast non responding effects. However, from a theoretical perspective, although non relevant to consistency, weighting system may have a crucial role in matters that concern efficiency.

[6] For any (β, ζ), a matrix F_k should exist such that $G_k = F_k$ and $\hat{\Sigma}_k(\beta, \zeta) = F_k(\beta, \zeta) \cdot \Sigma_k(\zeta) \cdot F_k^{tr}(\beta, \zeta)$ because, by construction, both $\hat{\Sigma}_k$ and Σ_k are positive semi-definite symmetric matrices which have the same dimension and the same rank

4.3 Special Weighting and Approximate Asymptotic Formulas

As a special relevant case of weighting in (7), we propose[7] to set $\mathcal{S}_k = \tilde{\mathcal{S}}_k := \hat{\Sigma}_k^-$, where $\hat{\Sigma}_k^-$ is the Moore-Penrose generalized inverse matrix of $\hat{\Sigma}_k$. Then, Eq. (10) might be rewritten, provided proper conditions as:

$$\beta = \left\{ \sum_{k=1}^{n} Z_k^{tr} \cdot (\hat{\Lambda}_k + \hat{\Sigma}_k)^{-1}(\beta, \zeta) \cdot Z_k \right\}^{-1} \left\{ \sum_{k=1}^{n} Z_k^{tr} (\hat{\Lambda}_k + \hat{\Sigma}_k)^{-1}(\beta, \zeta) \cdot \hat{x}_k(\beta, \zeta) \right\}.$$

Let $\beta^*(x)$ denote a solution of the FP-equation above, supposing that $\zeta := \zeta_0$ was assigned. Suppose now that, within some proper hypothetical design while n increases, a sequence of solutions β^* exists which converges to the "true value" β_0 in probability. Then (although exact sampling properties of $\beta^*(X)$ over the sample space of random profile X would be difficult to establish) asymptotic properties might be inherited[8] by those of the following (hypothetical) "one-step ahead" iteration:

$$\hat{\beta}(x) = \left\{ \sum_{k=1}^{n} Z_k^{tr} \cdot (\hat{\Lambda}_k + \hat{\Sigma}_k)^{-1} \cdot Z_k \right\}^{-1} \left\{ \sum_{k=1}^{n} Z_k^{tr} \cdot (\hat{\Lambda}_k + \hat{\Sigma}_k)^{-1} \cdot \hat{x}_k(x) \right\} (\beta_0, \zeta_0). \tag{12}$$

Thus, the asymptotic variance-covariance matrix of $\beta^*(x)$ may be approximately evaluated by the following formula:

$$AsyVar[\beta^*(X)] \approx AsyVar[\hat{\beta}(X)] = \left\{ \sum_{k=1}^{n} Z_k^{tr} \cdot (\hat{\Lambda}_k + \hat{\Sigma}_k)^{-1}(\gamma^*) \cdot Z_k \right\}^{-1}, \tag{13}$$

which would provide also evaluation of standard errors for the fixed-effects. Here, QL-based formula (13) might have some practical interest due to the relative simplicity in implementing calculations. Notice that, if responses x_k were Gaussian and the link was the identity function, then $\beta^*(x)$ exactly would coincide with the usual (e.g. see [4]) FML estimate and (13) with the exact FML-matrix of variance-covariance.

---

[7] This choice would be the optimal, in the class of QL-linear estimating Eq. (8 and 9), [2, 3]

[8] We suppose here that subsequent iterations of updating Eq. (12) may be neglected, provided the sample size is sufficiently large. Here, to recall certain analogies, see McCullagh and Nelder ([8], pp. 328, pp. 347–348). Recalling sampling identities of $\hat{X}_k$, due to quasi-linearization, we could see that random vector $\hat{\beta}(X)$ is unbiased while its variance-covariance matrix has a typical sandwich-like structure. Thus, sequence of solutions $\beta^*(x)$ would be asymptotically unbiased with the asymptotical variance-covariance matrix provided by $Var_{(\beta_0,\zeta_0)}[\hat{\beta}(X)]$.

Table 1 "The modern prenatal care data set". Results by Rodriguez et al. [11] and added estimates from "EFP"

Estimating methods									
	Logit	MQL-1	MQL-2	PQL-1	PQL-2	PQL-B	ML: Maximum likelihood	Gibbs	EFP
Numeric convergence	yes	yes	(?)	yes	(?)	(?)	yes	(?)	yes
Effects									
Fixed effects									
Individual									
Child aged 3-4 years	−0.20	−0.17	−0.25	−0.22	−0.44	−0.81	−1.04	−1.33	−1.111
Mother aged ≥ 25 years	0.32	0.31	0.38	0.36	0.58	1.35	1.08	1.26	0.823
Birth order 2-3	−0.10	−0.10	−0.16	−0.13	−0.20	−0.49	−0.75	−1.00	−0.551
Birth order 4-6	−0.23	−0.23	−0.32	−0.26	−0.31	−0.97	−0.56	−0.49	−0.447
Birth order ≥ 7	−0.19	−0.28	−0.45	−0.30	−0.45	−1.08	−1.08	−1.21	−0.833
Family									
Indigenous, no Spanish	−0.84	−0.97	−1.02	−1.22	−2.18	−4.63	−5.60	−7.54	−5.210
Indigenous Spanish	−0.57	−0.56	−0.93	−0.67	−1.00	−2.54	−2.62	−4.00	−2.800
Mother's education primary	0.31	0.35	0.59	0.42	0.65	1.64	1.89	2.62	1.707
Mother's education secondary or better	1.01	0.90	1.06	0.98	1.93	3.81	3.61	5.68	3.622
Husband's education primary	0.18	0.22	0.32	0.25	0.30	0.95	0.96	1.11	0.805
Husband's education secondary or better	0.68	0.69	0.85	0.82	1.59	3.07	4.37	4.85	3.103
Husband's education missing	0.00	0.06	0.07	0.06	0.01	0.16	0.13	0.02	0.081
Husband's professional, sales, clerk	−0.32	−0.40	−0.49	−0.47	−0.64	−0.60	−0.62	−0.56	−0.486
Husband's agricultural self-employed	−0.54	−0.52	−0.66	−0.62	−0.86	−1.75	−1.77	−2.64	−1.636
Husband's agricultural employer	−0.70	−0.27	−0.33	−0.29	−0.25	−2.34	−2.67	−3.77	−2.303
Husband's skilled service	−0.37	−0.15	−0.19	−0.18	−0.05	−1.05	−0.80	−1.12	−0.696
Modern toilet in household	0.47	0.37	0.57	0.41	0.94	1.72	2.01	2.69	1.531

Table 1 (continued)

Estimating methods									
	Logit	MQL-1	MQL-2	PQL-1	PQL-2	PQL-B	ML: Maximum likelihood	Gibbs	EFP
Numeric convergence	yes	yes	(?)	yes	(?)	(?)	yes	(?)	yes
Television not watched daily	0.32	0.27	0.48	0.31	0.53	1.16	1.35	2.03	1.098
Television watched daily	0.47	0.33	0.41	0.39	0.67	1.55	1.51	2.05	1.235
Community									
Proportion indigenous, 1981	−0.90	−0.97	−1.61	−1.12	−2.05	−4.48	−5.01	−6.61	−3.688
Distance to nearest clinic	−0.01	−0.01	−0.01	−0.01	−0.02	−0.05	−0.05	−0.07	−0.055
Random effects									
Standard deviation σ									
Family	–	1.01	1.74	1.25	2.75	6.66	7.40	10.24	6.614
Community	–	0.79	1.23	0.86	1.71	3.48	3.74	5.40	2.917

5 A Comparative Study

A comparative study is presented in Table 1, which would complete that of Rodriguez et al. [11] over a reference data-set "the modern prenatal care". This data set is critical in that it exhibits high clustering effect due to relatively large random effects in small-sized groups. In Table 1, EFP ("Empirical Fixed Point") denotes a procedure which implemented recursive updating of fixed-point Equations (10 and 11), by using weighting (7) with $S_k := \tilde{S}_k := \hat{\Sigma}_k^-$. It used only matrix algebra and 1-dimensional quadrature-formula, although in principle 2-dimensional integrals would be necessary (regardless of whether or not dimension of the system of random effects is higher than two) for exact implementing of QL-objects. Simulation-based studies on asymptotics are reported in D'Epifanio [2].

6 Concluding Remarks

Based on structural sampling properties of quasi-linearized latent implicit moments, we have proposed an approach which uses a type of non-standard estimating system. It is flexible enough to be adapted to different criteria of weighting, robust with respect to the size of variances of random effects. Intrinsically by construction, it should be structurally robust (recalling M-estimators) with respect to the exact form of mixing density. Furthermore, it is easy to implement through effective numerical procedures, which do not need simulation based tools. Some formulas and algorithms extend, in their structure under certain specific choices of weighting, the usual one (which are been developed using maximum likelihood), from the Gaussian setup to the larger class of the "generalized linear mixed models".

References

1. Basawa I., Godambe V.P., Taylor R.L. (eds.): Selected Proceeding of the Symposium on Estimating Function, Athens. Lectures Note-Monograph Series, vol. 32. Institute of Mathematical Statistics, Hayward, CA (1997)
2. D'Epifanio, G.: An Implicit Estimating Equation Approach for Mixed Model. Applications to Highly Correlated Binary Responses, Departiment of Economy, Finance and Statistics, University of Perugia, http://www.ec.unipg.it/DEFS/uploads/highcorrefpfinallogo.pdf (2006)
3. D'Epifanio G.: Su una Procedura di Stima per Modelli Multilivello con Risposte Binarie Altamente Correlate. In: Liseo, B., Montanari, G.E.M., Torelli, N. (eds.) Metodi per l' Integrazione di Dati da Fonti Diverse, Franco Angeli, Milano (2006)
4. Diggle P.J., Liang K., Zeger, S.L.: Analysis of Longitudinal Data. Clarendon Press, Oxford (1994)
5. Godambe, V.P. (ed.): Estimating Functions. Oxford University Press, Oxford (1991)
6. Hartigan, J.A.: Linear Bayes methods. J. R. Stat. Soc. Ser. B **31**(3), 446–454 (1969)
7. Huber, P.J.: Robust Statistics. Wiley, New York, NY (1981)
8. McCullagh, P., Nelder, J.A.: Generalized Linear Model, 2nd ed. Chapman and Hall, London (1991)

9. Rabe-Hesketh, S., Skrondal, A., Pickles, A.: Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. J. Econ. **128**, 301–323 (2005)
10. Reza Fotouhi, A.: Comparison of estimation procedures for multilevel models. J. Stat. Softw. **8**(9), http://www.jstatsoft.org/v08/i09 (2003)
11. Rodriguez, G., Goldman, N.: Improved estimation procedure for multilevel models with binary response: a case-study. J. R. Stat. Soc. Ser. A **164**, 339–355 (2001)
12. Zeger, S.L., Karim, M.R.: Generalised linear models with random effects: a Gibbs sampling approach. J. Am. Stat. Assoc. **86**(413), 79–86 (1991)

Causal Inference Through Principal Stratification: A Special Type of Latent Class Modelling

Leonardo Grilli

Abstract Principal stratification is an increasingly adopted framework for drawing counterfactual causal inferences in complex situations. After outlining the framework, with special emphasis on the case of truncation by death, I describe an application of the methodology where the analysis is based on a parametric model with latent classes. Then, I discuss the special features of latent class models derived within the principal strata framework. I argue that the concept of principal stratification gives latent class models a solid theoretical basis and helps to solve some specification and fitting issues.

1 Introduction

Principal stratification is a conceptual framework developed in the setting of counterfactual causal inference to deal with situations where the causal path from the treatment to the outcome includes an intermediate variable that cannot be ignored [3]. Examples are *non-compliance* [1, 2, 10], where the intermediate variable is the compliance status, estimation of *direct effects* [14, 16], where the intermediate variable is a variable whose effect one wishes to control for, *surrogacy in clinical trials* [7, 8], where the intermediate variable is a surrogate endpoint, and *truncation by death* [4, 5, 18, 19], where the intermediate variable determines the existence of the outcome.

Basically, the problem with intermediate variables is that they are measured after treatment and thus they are not balanced among the treatment arms. Therefore, the conventional estimators of the causal effect of an intermediate variable are generally biased; moreover, conditioning on an intermediate variable may bias the estimators of other causal effects of interest.

The application described in the next section focuses on truncation by death, a case taking its name from the studies on the quality of life, where the outcome of interest does not exist for patients who died. The simplest approach is to carry out

L. Grilli (✉)
Department of Statistics "G. Parenti", University of Florence, Florence, Italy
e-mail: grilli@ds.unifi.it

B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-13312-1_27,

the analysis on the patients who survived, but this is likely to yield biased results since survival is a post-treatment variable and conditioning on it destroys the randomized structure of the experiment. Zhang and Rubin [18] noted that the same issue may arise in experiments for comparing educational programmes. In fact, they applied principal stratification to the hypothetical case of a randomized experiment concerning two high school educational programmes, where the intermediate variable is graduation and the outcome is the score on a final test. This is an instance of truncation by death since the outcome of interest exists only for students who graduated.

Later Grilli and Mealli [4, 5] used principal strata to tackle a case of truncation by death in the evaluation of the effectiveness of two degree programmes with respect to job opportunities, where the treatment is the degree programme (Economics vs Political Science), the intermediate variable is the graduation status (graduated within 9 years) and the outcome is the employment status (having a permanent job). This is another instance of truncation by death: since the aim is to assess the relative effectiveness of graduation in different degree programmes, the employment status is not defined for students who did not graduated.

A further application of principal stratification to deal with truncation by death is given by Zhang, Rubin and Mealli [19] in the context of the effectiveness of job-training programs: indeed, estimating the effects of training programs on wages is complicated by the fact that, even in a randomized experiment, wages are truncated by nonemployment, that is, they are only observed and well-defined for individuals who are employed.

The paper proceeds with a section illustrating principal stratification through an application to the effectiveness of degree programmes and a section discussing the latent class perspective of principal stratification.

2 Principal Stratification: Basic Ideas and an Application

The principal stratification framework requires a treatment with a finite number of levels and two post-treatment variables, namely an intermediate variable and an outcome. The nature of the intermediate variable determines the type of strata: a discrete intermediate variable implies discrete strata, while a continuous intermediate variable implies continuous strata. It will be clear that only discrete strata can be seen as latent classes. The simplest case of discrete principal strata arises when both the treatment and the intermediate variable are binary, implying four principal strata.

The principal stratification framework will be illustrated through the application of Grilli and Mealli [4, 5], who analyzed 1941 freshmen of the University of Florence: 1,068 enrolled in Economics and 873 in Political Science.

The *treatment* Z_i takes the value 1 if student i enrolled in Economics and 0 if enrolled in Political Science. Under the standard Stable Unit Treatment Value Assumption [3] (SUTVA), the post-treatment variables are defined as follows. The

intermediate variable $S_i(z_i)$ is 1 or 0 if student i *graduated* or *did not graduate* within 9 years when enrolled in degree programme z_i. The *outcome* $Y_i(z_i)$ is 1 or 0 if student i *had* or *did not have* a permanent job at the time of the interview (i.e. from one to two years after the degree) when enrolled in programme z_i and graduated.

Since for each individual the treatment assumes a single value, for every post-treatment variable only one of the two potential versions can be observed: $S_i^{obs} = S_i(Z_i)$ and $Y_i^{obs} = Y_i(Z_i)$. Since both the treatment and the intermediate variable are binary, there are four *principal strata*:

- GG (Graduated, Graduated) if $S_i(1) = 1$ and $S_i(0) = 1$;
- GN (Graduated, Not graduated) if $S_i(1) = 1$ and $S_i(0) = 0$;
- NG (Not graduated, Graduated) if $S_i(1) = 0$ and $S_i(0) = 1$;
- NN (Not graduated, Not graduated) if $S_i(1) = 0$ and $S_i(0) = 0$.

The principal stratum of individual i cannot be observed since either $Z_i = 0$ or $Z_i = 1$. The principal stratum is thus a latent class, denoted with a latent variable C_i taking values in the set $\{GG, GN, NG, NN\}$. The probability that an individual belongs to a given principal stratum can be estimated. A crucial feature is that, given the values of the treatment Z_i and the intermediate variable S_i^{obs}, some principal strata are ruled out: for example, a student who enrolled in Economics ($Z_i = 1$) and then graduated ($S_i^{obs} = 1$) can only belong to the strata GG and GN, so the strata NG and NN are inadmissible and their probability is null.

The key feature of the principal strata is that they are defined by the couple of potential values of the intermediate variable, so they are not affected by the treatment and thus can be seen as categories of an unobserved pre-treatment covariate.

The terms entering the causal effect of interest $Y_i(1) - Y_i(0)$ are both defined only in the GG stratum, i.e. students who would be able to graduate in both programmes. The estimand of main interest is thus the *Average Causal Effect* (ACE) *on employment in the GG stratum*, i.e. the difference between the probabilities of being employed for Economics and Political Science in the subset of students that would be able to graduate in any of the two degree programmes.

Grilli and Mealli [4, 5] included also some covariates $\mathbf{x}_i$. In general, covariates are important when the treatment is not randomized, since the *unconfoundedness assumption* required for the causal interpretation of the effect of the treatment is more reasonable if stated conditional on good covariates. Formally, the treatment is conditionally unconfounded when $Z_i \perp \{S_i(0), S_i(1), Y_i(0), Y_i(1)\}|\mathbf{x}_i$.

Under the assumptions of SUTVA and conditional unconfoundedness, the data generating process can be defined in terms of the following two sets of probabilities: (a) probabilities of the principal strata $\{\pi_{GG:i}, \pi_{GN:i}, \pi_{NG:i}, \pi_{NN:i}\}$, e.g. $\pi_{GG:i} = \Pr(C_i = GG|\mathbf{x}_i)$; (b) probabilities of the outcome conditional on the principal stratum $\{\gamma_{1,GG:i}, \gamma_{0,GG:i}, \gamma_{1,GN:i}, \gamma_{0,NG:i}\}$, where the number 0 or 1 in the subscript is the value of Z_i. For example, $\gamma_{0,GG:i} = \Pr(Y_i(0) = 1|C_i = GG, \mathbf{x}_i)$. Here the γ's for other combinations of programme and principal stratum, such as $Z_i = 1$ *and* $C_i = NG$, are not defined.

As in the majority of the applications with principal strata, the treatment and the intermediate variable are both binary, leading to four principal strata. However, while in many settings it is sensible to assume that certain strata are empty (e.g. the assumption of no defiers in an experiment with non-compliance), in the present context such assumptions are not plausible in the light of the symmetry of the two treatments, so all the strata are allowed to exist and thus every observed group is generated by a mixture of two distributions.

The principal stratification framework can be exploited to carry out a non-parametric analysis based on large-sample bounds [5] or to build a parametric model to be fitted with Bayesian or likelihood methods [4]. In the example, the likelihood is a product over four observable groups defined by Z_i and S_i^{obs}:

$$\prod_{i:\, Z_i=1,\, S_i^{obs}=0} \{\pi_{NG:i} + \pi_{NN:i}\} \times \prod_{i:\, Z_i=1,\, S_i^{obs}=1} \left\{\pi_{GG:i} B_{1,GG:i} + \pi_{GN:i} B_{1,GN:i}\right\} \times$$
$$\prod_{i:\, Z_i=0,\, S_i^{obs}=0} \{\pi_{GN:i} + \pi_{NN:i}\} \times \prod_{i:\, Z_i=0,\, S_i^{obs}=1} \left\{\pi_{GG:i} B_{0,GG:i} + \pi_{NG:i} B_{0,NG:i}\right\}$$

where the B's are the Bernoulli likelihoods for the γ's, for example $B_{1,GG:i}$ is $\left(\gamma_{1,GG:i}\right)^{Y_i^{obs}} \left(1-\gamma_{1,GG:i}\right)^{1-Y_i^{obs}}$.

The parametric model devised by Grilli and Mealli [4] is made of two components: a multinomial logit model for the probabilities of the principal strata conditional on the covariates (the π's) and a set of logit models for the probabilities of the outcome conditional on both the covariates and the principal stratum (the γ's). The model is thus a latent class model, but the principal stratification framework entails some peculiarities that make the analysis different from traditional latent class modelling.

3 Principal Stratification and Latent Class Modelling

In the previous section it has been shown that in the case of discrete principal strata the corresponding statistical model is a latent class (LC) model. Note that even if almost all applications assume discrete strata, the principal strata can also be continuous: For example, Jin and Rubin [6] tackled partial compliance by defining the strata as couples of proportion of compliance to drug and proportion of compliance to placebo.

The connection between principal stratification and LC modelling has been recognized in the case of *non-compliance*, with reference to the simple instance of a binary treatment and a binary compliance status (all-or-none compliance). In the notation of the previous section, the intermediate variable $S_i(z_i)$ is the compliance status under treatment z_i. The target quantity, called Complier Average Causal Effect (CACE), is the average difference $Y_i(1) - Y_i(0)$ for individuals in the principal

stratum of compliers, namely the individuals that comply with the treatment regardless of the assigned treatment [1].

Bengt Muthén described CACE modelling in terms of LC modelling in [11] and then implemented the idea in the M*plus* software, whose user's manual [12] reports a re-analysis of Little and Yau's data [9]. In M*plus* the class membership restrictions are handled by the so-called *training data*, i.e. an auxiliary dataset declaring, for each sample unit, which classes are admissible and which classes are not. The possibility to specify a CACE model as an LC model with restrictions is also noted by Vermunt and Magidson in the manual of the software Latent GOLD [17], where the class membership restrictions are inserted via the *Known Class* option.

The latent class perspective in CACE modelling was exploited also by Skrondal and Rabe-Hesketh in their book on Generalized Linear Latent Mixed Models [15], where they showed how a CACE model can be written as an LC model that fits the GLLAMM framework. Moreover, they re-analyzed Little and Yau's data using the Stata `gllamm` command [13].

The mentioned treatments of CACE via LC models are aimed at showing that causal inference can be carried out within a general statistical modelling framework based on latent variables. However, the implications of the connection have not been investigated. Moreover, there seems to be no discussion of the connection in the more general principal stratification framework, thus including topics such as direct effects and truncation by death.

Let us use the notation introduced in the previous section and let us denote the latent class corresponding to the principal stratum with $C_i = c$ for $c \in \mathcal{C}$. An LC model derived within a framework with discrete principal strata differs from a general LC model in several respects: (*a*) the number of classes (i.e. the cardinality of $\mathcal{C}$) and their meaning is determined a priori, as each class corresponds to a principal stratum; (*b*) an individual can only belong to a subset of latent classes, i.e. given the data the probabilities of belonging to certain classes are zero by assumption: $\exists c \in \mathcal{C}$ such that $\Pr(C_i = c | Z_i, S_i^{obs}, \mathbf{x}_i) = 0$. Truncation by death adds another peculiarity, namely: (*c*) latent class membership determines whether the outcome is defined or not (and its probability in case it is defined): $\exists c \in \mathcal{C}$ such that $Y_i(z_i)$ is not defined.

Feature *a* allows to avoid the tricky problem of a data-driven choice of the number of latent classes and the somewhat arbitrary exercise of attaching labels to the classes. Feature *b* makes estimation simpler with respect to a standard LC model with the same number of classes, since some components of the mixtures are ruled out by assumption. Feature *c* is specific to truncation by death in the principal strata framework and does not apply to standard LC models, where it is not conceivable to let the outcome be defined or not depending on the class.

As for model specification, principal stratification gives solid arguments to put restrictions on the latent classes based on substantive assumptions or on the design: for example, in experiments with non-compliance [1, 2] the latent class of *defiers* can be assumed to be empty based on considerations on the behaviour of the individuals, while the latent class of *always takers* is empty if the design prevents people assigned to control from taking the active treatment.

Last but not least, a latent class model with a structure derived within the principal strata framework guarantees that the model is consistent with the principles of counterfactual causal inference and thus the parameters refer to well-defined causal quantities.

References

1. Angrist, J.D., Imbens, G.W., Rubin, D.B.: Identification of causal effects using instrumental variables. J. Am. Stat. Assoc. **91**, 444–472 (1996)
2. Barnard, J., Frangakis, C.E., Hill, J.L., Rubin, D.B.: Principal stratification approach to broken randomized experiments: a case study of school choice vouchers in New York City. J. Am. Stat. Assoc. **98**, 299–323 (2003)
3. Frangakis, C.E., Rubin, D.B.: Principal stratification in causal inference, Biometrics **58**, 21–29 (2002)
4. Grilli, L., Mealli, F.: University studies and employment. An application of the principal strata approach to causal analysis. In: Fabbris, L. (ed.) Effectiveness of University Education in Italy, pp. 219–232. Physica-Verlag, Heidelberg (2007)
5. Grilli, L., Mealli, F.: Nonparametric bounds on the causal effect of university studies on job opportunities using principal stratification. J. Educ. Behav. Stat. **33**, 111–130 (2008)
6. Jin, H., Rubin, D.B.: Principal stratification for causal inference with extended partial compliance. J. Am. Stat. Assoc. **103**, 101–111 (2008)
7. Joffe, M.M., Greene, T.: Related causal frameworks for surrogate outcomes. Biometrics **65**, 530–538 (2009)
8. Li, Y., Taylor, J.M.G., Elliott, M.R.: A Bayesian approach to surrogacy assessment using principal stratification in clinical trials. Biometrics. **66**, 523–531 (2010)
9. Little, R.J., Yau, L.H.Y.: Statistical techniques for analyzing data from prevention trials: treatment of no-shows using Rubins causal model. Psychol. Methods **3**, 147–159 (1998)
10. Mattei, A., Mealli, F.: Application of the principal stratification approach to the Faenza randomized experiment on breast self-examination. Biometrics **63**, 437–446 (2007)
11. Muthén, B.O.: Beyond SEM: general latent variable modeling. Behaviormetrika **29**, 81–117 (2002)
12. Muthén, L.K., Muthén, B.O.: Mplus User's Guide, 5th ed. Muthén & Muthén, Los Angeles, CA (2007)
13. Rabe-Hesketh, S., Skrondal, A., Pickles, A.: GLLAMM Manual, U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 160. University of California, Berkeley, CA (2004)
14. Rubin, D.B.: Direct and indirect causal effects via potential outcomes. Scand. J. Stat. **31**, 161–170 (2004)
15. Skrondal, A., Rabe-Hesketh, S.: Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models. Chapman & Hall/CRC Press, Boca Raton, FL (2004)
16. VanderWeele, T.: Simple relations between principal stratification and direct and indirect effects. Stat. Probab. Lett. **78**, 2957–2962 (2008)
17. Vermunt, J.K., Magidson, J.: Technical Guide for Latent GOLD 4.0: Basic and Advanced. Statistical Innovations Inc., Belmont, MA (2005)
18. Zhang, J.L., Rubin, D.B.: Estimation of causal effects via principal stratification when some outcomes are truncated by 'death'. J. Educ. Behav. Stat. **28**, 353–368 (2003)
19. Zhang, J.L., Rubin, D.B., Mealli, F.: Likelihood-based analysis of causal effects of job-training programs using principal stratification. J. Am. Stat. Assoc. **104**, 166–176 (2009)

Scaling the Latent Variable Cultural Capital via Item Response Models and Latent Class Analysis

Isabella Sulis, Mariano Porcu, and Marco Pitzalis

Abstract One of the main tasks of an educational system is to enrich the *Cultural Capital* of its students. The *Cultural Capital* linked to social origins is considered crucial in determining students'social life and subsequent professional achievement. This work moves from an *ad hoc* survey carried out on a sample of students who enrolled or applied for an entrance test at the university. The *Cultural Capital* is treated as a latent variable which students are supposed to possess at a greater or lesser degree. Latent Class Analysis is adopted in order to provide a non arbitrary scaling of *Cultural Capital* and to sort out mutually exclusive classes of students. Moreover, Item Response Models are implemented to assess the calibration of the questionnaire as an instrument to measure the *Cultural Capital* of the surveyed population.

1 Introduction

This paper deals with the role played by *Cultural Capital* (CC) in shaping students'choices with respect to the transition from high school to university. Its main aim is to propose a way of quantifying the intangible construct CC via a survey questionnaire and to spot out differences in the amount of CC owned by clusters of students. This issue is investigated with an *ad hoc* survey carried on in 2007 at the University of Cagliari. According to Pierre Bourdieu's standpoint [5], we assume the CC as a strategic resource that involves the construction of individual habits linked to a defined position in a relational space. In Pierre Bourdieu's theory, the CC has three different forms: *embodied*, *objectified*, *institutionalized* [4]. We focus on the embodied form of CC which is the product of family socialization and cultural activities.

Hereafter, we will suppose that each individual possesses a basic amount of CC, namely the *inherited* CC (CC_{IH}). This basic amount of CC is measured considering the highest level of formal education reached by students'parents. It is called

I. Sulis (✉)
Dipartimento di Ricerche Economiche e Sociali, Università di Cagliari, Cagliari, Italy
e-mail: isulis@unica.it

B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-13312-1_28,

inherited because we suppose that it is an asset owned by students' parents and automatically transmitted to the family. This work focuses on two more sub-components of CC that can be considered the results of family and individual choices / actions / activities: (i) the *family made* CC (CC_{FM} – built up by positive actions made by students'families); (ii) the *pro-active* CC(CC_{PA} – built up by the students) or self constructed.

2 The Survey

In order to shed some light on the cultural characteristics of the population of university students a survey has been carried out at the University of Cagliari. The analysis has been run in a more extended research project aimed to investigate upon the transitions school-university in an isolated regional context such as Sardinia.

This work, specifically, aims to quantify the amount of CC owned by university students in order to investigate on factors which influence the educational achievement taking into account the social and geographic context in which students have grown up. A sample has been selected from the population of students who completed their secondary school schemes in 2006 and applied for an entrance test or directly enrolled at the University of Cagliari in the 2006–2007 academic year (69.3% of the population applied for an entrance test and the 31.7% directly enrolled to a faculty). A CATI survey was carried out in April–May 2007. The sampling rate has been set equal to about 10% of the overall population. The sample size is equal to 494 units. The 7.6% of the sample is composed by students who did not applied at

Table 1 Some descriptive statistics

Variables	Sample	Population	Variables	Sample	Population
School* (%)			Age		
Liceo	45.95	46.84	Mean	19.88	19.93
Not-Liceo	54.05	53.16	Median	19.28	19.37
Faculty (%)			SD	2.74	2.75
Economics	10.88	11.87	Final mark§		
Pharmacy	4.31	4.29	Mean	79.07	79.23
Law	12.73	11.79	Median	78.00	78.00
Engineering	16.22	17.90	SD	12.51	14.18
Literature	10.88	9.92	Sex (%)		
Foreign Languages	4.52	5.78	F	58.10	62.05
Medicine	3.29	4.80	M	41.90	37.95
Educational Science	6.16	7.77			
Sciences†	13.76	14.84			
Political Science	9.65	11.09			
None‡	7.60	-			

*The *Liceo* provides a classical education such as the old British *Grammar Schools*.
†Math, Physics, Biology, Chemistry, Natural Science, Computer Science.
‡The *sample* column contains 37 people who did not enrol after failing the admission tests.
§ At school graduation (in hundreds of pts.).

Table 2 Items contains and percentage of positive responses

Items	% Yes
CC_{FM}	
I_1 Student's parents belong to a cultural association	22.9
I_2 Student has attended non-school music classes	40.9
I_3 Student has attended non-school foreign language classes	36.7
I_4 Student's family has traveled for holidays	72.4
I_5 Student has visited cultural expositions with parents	10.9
I_6 Student's parents have used to buy non-school books as a gift	24.5
CC_{PA}	
I_7 The student has bought books as a gift	12.9
I_8 The student has bought non-school books for herself	38.9
I_9 The student has attended classical music live performances	2.4
I_{10} The student has attended pop music live performances	11.3
I_{11} The student has attended jazz music live performances	1.2
I_{12} The student belongs to a cultural association	22.1

the University of Cagliari after failing the admission test to a specific faculty. Some descriptive statistics are depicted in Table 1.

The *family made* and the *pro-active* sub-components are measured by *actions / activities* made by the students or by their families. These actions are described by the items listed in Table 2. For each of the 12 items the percentage of positive answers is reported in the last column. Item contents have been defined in order to build up a measurement instrument which made possible comparisons across students belonging to different faculties and to have a picture of their general level of CC at the moment they enter at the university. The selected *actions* require that students / families own a minimum amount of CC in order to be acted and the minimum threshold varies across them. Questionnaire items concern general habits that are not linked to university / school *curricula*. The *family made* sub-component loads all items which require that the family has activated a specific action in order to be positively answered by students (i.e. to enrol children to private language classes or music lecturers, to visit cultural exposition, or to travel for holiday, etc.). The *pro-active* sub-component is addressed to account for all the activities / habits that students practice without a direct involvement or support of the family but for a personal interest (i.e., to attend music live exhibition, to buy books, or to belong to a cultural association, etc.). In the following we focus on the analysis of these two sub-components and on the statistical methods useful to scale them.

3 Scaling the Cultural Capital via LCA

Latent Class Analysis (LCA) is applied in order to sort out a number R of mutually exclusive classes of individuals (the latent classes of the categorical latent variable) who are supposed to possess different amount of the latent variables [1, 3, 8] moving from individual responses to the manifest variables (cross classification of I polythomus or binary indicators). Individuals (students) are classified into clusters

based upon membership probabilities (posterior probabilities). Each latent class (LC) groups students who share the same level of *Cultural Capital* (with respect to the specific dimension defined by the set of items). The assumption of a basic latent class model is that responses $\mathbf{Y_j}(Y_{j1}, \ldots, Y_{jI})$ of individual j $(j = 1, \ldots, n)$ to a set of (binary or polythomous) manifest variables I $(i = 1, \ldots, I)$ are independent conditionally upon the latent classes $r = 1, \ldots, R$ of the categorical latent variable θ to which the individual belongs to. In the case of binary items, Y_{ij} is an indicator variable (which can assume values 0 or 1), π_{ir} the probability that an observation in class r answers positively to item i (the item-response probabilities conditional upon the latent class membership), γ_r the probability to belong to class r of the latent variable θ (the latent class membership); the probability to observe a specific response pattern given that the student is in the latent class r is defined as:

$$\pi_{y_{j1}\ldots y_{jI}|\theta_r} = \prod_{i=1}^{I} \pi_{ir}^{y_{ji}} (1 - \pi_{ir})^{(1-y_{ji})}. \tag{1}$$

The contribution of individual j to to the likelihood is obtained by the summing the Eq. 3 over the R latent classes. The LCA has been estimated adopting the poLCA package implemented in R by D. A. Linzer and J. Lewis [8] which uses the **EM** algorithm in order to maximizing the log-likelihood function.

Table 3 shows the LCA for the 2, 3, and 4 LCA models measures of fit. The analysis was carried out separately for each sub-component. The 3 class model was retained for both. Moving from the item response probability conditional upon the LC memberships the profile of each LC was sketched out and LCs were ordered according to the degree of CC owned by their members (moving from the *lowest* to the *highest* amount). The criteria adopted for sorting out classes is based on the item response probability conditional upon class membership: values of $\pi_{i|\theta_r}$ was used to sort out the LCs and to label them. Moreover, the rate of positive answers to each item (see Table 2) helps us to classify each item in the range among *easy* and *difficult*. According to the criteria used to sort out categories, the relation $C_1 < C_2 < C_3$ holds on both sub-components.

Looking at the rate of positive answers in Table 2 arises that item I_5 (to visit cultural expositions) contains information on the activity that requires students the highest level of *family made* CC in order to be made. The rate of positive responses is 11.3%. It is followed by I_1, I_6, I_3, I_2 and I_4, which have percentages equal to 22.9, 24.5, 36.7, 40.9 and 72.4%, respectively. It is interesting to highlight that the three activities with the lowest rates of positive responses are those which requires a direct involvement of students'parents in the action. Students clustered in C_2 show a slightly higher probability than students clustered in C_3 to answer positively to items I_1 and I_2. However, considering that in the remaining four items students classified in C_3 show higher probability of providing positive answers, we rank $C_2 < C_3$. Furthermore, students clustered in C_3 are those who possess an amount of *family made* CC sufficient to answer positively to item I_5.

Looking at the second component *pro-active* CC it seems straightforward to order $C_1 < C_2 < C_3$. The ranking of the items according to the rate of positive

Table 3 Models results and measures of fit

a) LCA model measures of fit CC_{FM}, CC_{PA}: 2,3,4 class model

Comp.	n^o of par.	CC_{FM}	CC_{PA}
$4CLA$:	27	BIC(4): 3,354 ,G^2(4): 18	BIC(4): 2,104, G^2(4): 20
$3CLA$:	20	BIC(3): 3,318 ,G^2(3): 24	BIC(3): 2,065, G^2(3): 24
$2CLA$:	13	BIC(2): 3,298 ,G^2(2): 48	BIC(2): 2,047, G^2(2): 49

b) Three class model results

	CC_{FM}							CC_{PA}					
		$Pr(Y_i = Yes)$							$Pr(Y_i = Yes)$				
Latent classes	$\hat{\gamma}_\theta^*$	I_1	I_2	I_3	I_4	I_5	I_6	$\hat{\gamma}_\theta^*$	I_7	I_8	I_9	I_{10}	I_{11}	I_{12}
C_1	47%	0.06	0.12	0.21	0.50	0.00	0.14	60%	0.00	0.00	0.01	0.09	0.03	0.17
C_2	42%	0.35	0.65	0.45	0.81	0.00	0.21	38%	0.30	0.95	0.03	0.08	0.00	0.29
C_3	11%	0.32	0.49	0.51	1.00	0.62	0.56	2%	0.44	0.83	0.27	1.00	0.30	0.27

* predicted class memberships (by modal posterior prob.)

answers is: I_{11}, I_9, I_{10}, I_7, I_{12}, I_8. Students in C_1 exhibit a probability close to 0 to score positively in four items out of six, whereas students in C_3 show the highest probabilities to score positively in four items out of six (I_7, I_9, I_{10}, I_{11}). The response pattern of the second class is in the middle. Predicted class membership (CM) vectors are [0.47, 0.42, 0.11] for the first sub-component and [0.60, 0.38, 0.02] for the second one.

On the basis of the *family made* sub-component, the first class (C_1) identifies students who received from their family *low intensity actions* (LIA) of CC, the second (C_2) *moderate intensity actions* (MIA) and the third (C_3) *high intensity actions* (HIA). On the *pro-active* sub-component, students in C_1 are classified as *no active* (NA), those in C_2 as *slightly active* (SA) and those in C_3 as *moderately active* (MA). From this classification arises that the second component is strongly biased towards negative categories.

The result depicted in this first analysis is that the level of *Cultural Capital* is measured on the basis of actions made by students or by their families which are not "calibrated" with respect to the intensity of CC owned by the population of students surveyed.

Furthermore, results could suggest that the rule chosen in order to classify a student response as *positive* (i.e., the actions described in the item had to be made *frequently*) seems to be too restrictive with respect to the overall level of CC observed in the sample. This consideration holds for both sub-components: just one item out of twelve has a rate of positive answers greater than 50%. The next part of the paper is devoted to an explorative analysis of the characteristics of the actions (which signal different intensity of CC) selected as indicators of the two sub-components in order to assess how much they are calibrated with respect to the level of CC owned by the surveyed population.

4 Assessing the Difficulty Level of the Survey Questionnaire Using a Bidimensional IRT

In this section we use some tools provided by the Item Response Theory (IRT) in order to get a relative measure of the difficulty level of the questionnaire. The aim is to better understand the LCA results on the light of the characteristics of the items used to scale the two unobservable sub-components of the CC. An item in the questionnaire is considered relatively difficult with respect to another if it requires a higher level of *family made* or *pro-active* CC in order to be positively answered. Basically, IRT models assumes that the chance to score positively to an item depends on two parameters related to that item (in psychometric literature such parameters are called *difficulty* and *discrimination*) and on a subject parameter (*ability parameter*). Higher levels of the *ability* (the latent variable) imply an increase in the probability to observe a positive response to each item [2, 6, 7]. In order to jointly measure both sub-components of students' CC – CC_{PA} and CC_{FM} – two ability parameters are introduced in the model. This is done by considering the latent variable CC as a bidimensional random variable with a known distribution.

Specifically, the probability that unit j answers positively to an item i is modeled as function of a *difficulty* parameter (β_i), a discrimination parameter (λ_i) and two person parameters ($\boldsymbol{\theta}_j$)

$$logit(\pi_{ij}) = \beta_i + \sum_{r=1}^{2} \lambda_{ir}\theta_{jr}; \tag{2}$$

the latter have been specified bivariate normal $\boldsymbol{\theta}_j[\theta_{FMj}, \theta_{PAj}] \sim \mathcal{N}(0, \Sigma)$. The main advantage of modelling the latent variable CC using a bidimensional model rather than fitting a model for each set of indicator is that the connection between the two sub-components are specifically taken into account and the estimates of the parameters relay on the overall observations. Moreover, the approach allows to have estimates of the difficulty parameters which are comparable across the two sub-components [9]. In the framework of the quantification of the CC the lower is β_i, the higher is the intensity of the CC measured by the aspect i and the higher is the minimum level of CC required to students in order to provide a positive answer. Thus, the higher is β_i the easier is the item (i.e. the lower is the intensity of CC measured by a question). The vector $\boldsymbol{\Lambda}_i$ $[\lambda_{i1}, \lambda_{i2}]$ is composed by two binary indicators which specify on which dimension item i loads. We made the assumption that items have the same power to discriminate between subjects with different levels of ability by fixing loadings equal to one on each sub-component. Each θ_{jr} measures the intensity of sub-component r of the latent variable, namely CC_{PA} or CC_{FM}, in subject j. The higher is the level of θ_{jr} in student j, the greater is the probability that he/she answers positively to items which tap on dimension r.

Looking at the sub-component C_{FM} (Table 4) the easiest item is I_4 (*to travel frequently for holidays with family*) with an *odds* to observe a positive answer equal about to 3. The *odds* associated to items I_2 ($\beta_2 = -0.43$, *odds* = 0.65) and I_3 ($\beta_3 = -0.64$, *odds* = 0.53) highlight that both are relatively easier than the remaining three items (I_1, I_6, I_5). The most difficult item is I_5 ($\beta_5 = -2.39$) with an *odds* to answer positively equal to 0.09.

In the second sub-component CC_{PA}, the two easiest items $I_8(-0.57)$ and I_{12} (-1.57) have odds equals to 0.56 and 0.21; the most difficult items are I_9 and I_{11} which have item parameters equal to -4.29 and -5.02 and *odds* close to 0.

Table 4 Results of the bidimensional item response model

Item parameter estimates					
Item	Coef. CC_{FM} (odds)	p-value	Item	Coef. CC_{PA} (odds)	p-value
I_1	−1.41 (0.24)	0.00	I_7	−2.33 (0.10)	0.00
I_2	−0.43 (0.65)	0.00	I_8	−0.57 (0.56)	0.00
I_3	−0.64 (0.53)	0.00	I_9	−4.29 (0.01)	0.00
I_4	1.12 (3.06)	0.00	I_{10}	−2.51 (0.08)	0.00
I_5	−2.39 (0.09)	0.00	I_{11}	−5.02 (0.01)	0.00
I_6	−1.30 (0.27)	0.00	I_{12}	−1.57 (0.21)	0.00
Random effects estimates					
var(θ_{FM}): 0.82 (SE= .14), var(θ_{PA}): 1.33 (SE= .27), $cor(\theta_{FM}, \theta_{PA})$ 0.72					

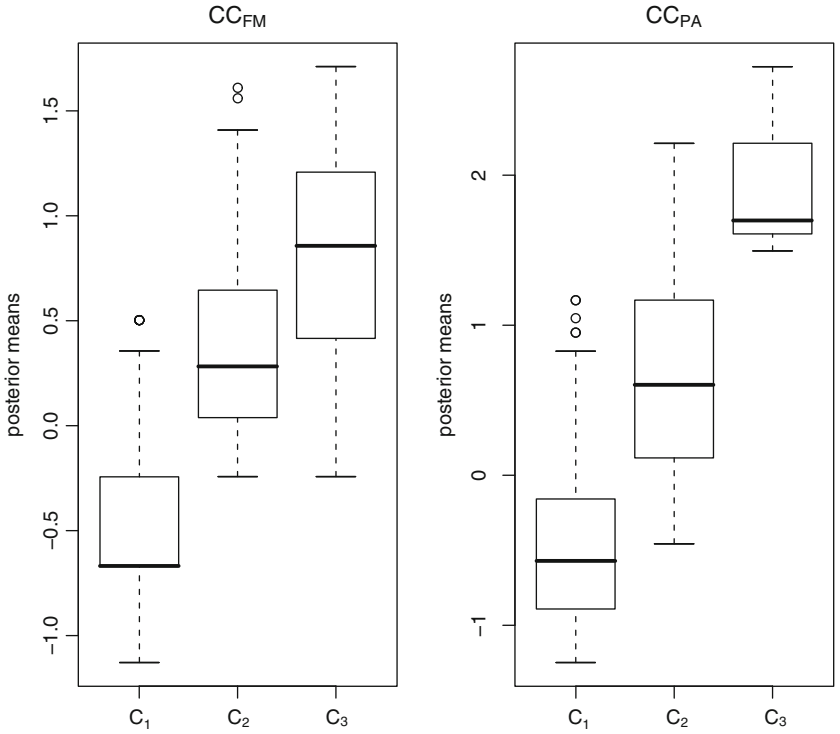


Fig. 1 Box plot of posterior means of students θ_{FM} and θ_{PA} by CC_{FM} and CC_{PA} class membership

The main results singled out by the model is that the structure of the test appears to be "too difficult" with respect to the average level of the CC owned by the surveyed students. Specifically, excluding item I_4, all item parameters have a negative sign and the highest *odd* to get a positive answer is 0.65. On the second subcomponent the test appears to be even more difficult to cope with: four items upon six have *odds* equal or lower than 0.10 (i.e. item I_7, I_9, I_{11}).

The *posterior means – empirical Bayes predictions* [10] – of the person parameters for both sub-components (CC_{FM} and CC_{PA}) show that differences in the intensity of *cultural capital* are clearer highlighted by using the LCA methods which considers both latent variables categorical. The distribution of the posterior means of students person parameters on the two sub-components conditional upon the class-membership is depicted in Fig. 1. The bunching of the sample in three clusters obtained with LCA seems to be adequate and this result validates the classification obtained adopting LCA.

5 Some Final Remarks

The attention of this research has been focused on the analysis of the items composing the sections of questionnaire addressed to measure two sub-components of

the latent variable *Cultural Capital*, namely CC_{FM} and CC_{PA}, and on their relative effectiveness in highlighting differences in the amount of *Cultural Capital* owned by students. For each sub-component the LCA was used in order to classify students in three mutually exclusive classes characterized by different intensity of the amount possessed of the latent variables. The *bidimensional model* (IRT), adopted in order to validate the results of the LCA, provides a relative evaluation of the difficulty of the questions relaying on responses to the overall set of 12 indicators. It shows also a classification of students and items which are consistent with the results obtained using LCA: the most *difficult* items are those which are scored positively just by students belonging to LC C_3, whereas the probability to answer positively to items (relatively) *easy* does not show significant differences among the three categories.

References

1. Agresti, A.: Categorical Data Analysis. Wiley-Interscience, Hoboken, NJ (2002)
2. Baker, F.B., Kim. S.H.: Item Response Theory: Parameter Estimation Techniques. Dekker, New York, NY (2004)
3. Barholomew, D.J., Steele, F., Galbraith, J.I., Moustaki, I.: The Analysis and Interpretation of Multivariate Analysis for Social Scientists. Chapman & All, Boca Raton, FL (2002)
4. Bourdieu, P.: The forms of capital. In: Richardson, J.G. (ed.) Handbook of Theory and Research for the Sociology of Education. Greenwood Press, New York, NY (1986)
5. Bourdieu, P.: Raisons pratiques. Sur la theorie de l'action. Edition du Seuil, Paris (1994)
6. De Boeck, P., Wilson, M. (eds.): Item Response Models: A Generalized Linear and Non Linear Approach. Statistics for Social and Behavioral Sciences. Springer, New York, NY (2004)
7. Fisher, G.H., Molenaar, I.W.: Rasch Models, Foundations, Recent Developments, and Applications. Springer, New York, NY (1995)
8. Linzer, D.A., Lewis, J.: poLCA: Polytomous Variable Latent Class Analysis. R package version 1.2. http://userwww.service.emory.edu/dlinzer/poLCA/ (2008)
9. Rijmen, F., Briggs, D.: Explanatory Item Response Models: A Generalized Linear and Non Linear Approach, Chapter Multiple Person Dimensions and Latent Item Predictors, pp. 111–166. Springer, New York, NY (2004)
10. Skrondal, A., Rabe-Hesketh, S.: Generalized Latent Variables Modeling. Chapman & Hall, Boca Raton, FL (2004)

Assessment of Latent Class Detection in PLS Path Modeling: a Simulation Study to Evaluate the Group Quality Index performance

Laura Trinchera

Abstract Structural Equation Models assume homogeneity across the entire sample. In other words, all the units are supposed to be well represented by a unique model. Not taking into account heterogeneity among units may lead to biased results in terms of model parameters. That is why, nowadays, more attention is focused on techniques able to detect unobserved heterogeneity in Structural Equation Models. However, once unit partition obtained according to the chosen clustering methods, it is important to state if taking into account local models provides better results than using a single model for the whole sample. Here, a new index to assess detected unit partition will be presented: the Group Quality Index. A simulation study involving two different simulation schemes (one simulating the so called null hypothesis of homogeneity among units, and the other taking into account the heterogenous sample case) will be presented.

1 Introduction

Heterogeneity among units is an important issue in statistical analysis. Treating the sample as homogeneous, when it is not, may seriously affect the results [6]. In Structural Equation Models (SEM) [2, 7] all the units are most often supposed to be well described by a unique model. Nevertheless, this hypothesis may often turn to be false. Recently, several techniques able to provide clustering in PLS Path Modeling (PLS-PM) [8, 10] have been presented [5, 6, 9]. However, no matter which method is used to cluster units, once the latent groups are identified, it is important to assess the differences between the detected classes of units and to evaluate the quality of the obtained partition. The first point essentially entails comparing the obtained local models to one another as well as with the global model. In PLS-PM framework only non parametric procedures and resampling methods, such as a bootstrap based

L. Trinchera (✉)
Department of Signal Processing & Electronic Systems, SUPELEC, Gif-sur-Yvette, France
e-mail: laura.trinchera@supelec.fr

B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-13312-1_29,

technique, are available. As regards the second point, i.e. assess the quality of the obtained partition, no specific index or methods have been developed until now. Here we meet this need by presenting a new index to evaluate the quality of the obtained partition: the Group Quality Index (*GQI*).

The remainder of the paper it is organized as follows: first we introduce the *GQI* (cf. 2), then a simulation study to asses the *GQI* properties is presented (cf. 3), to conclude a discussion on the obtained results and of the directions of further research is provided (cf. 4).

2 A New Index to Assess Group Separation in PLS-PM: The Group Quality Index

Assessing the quality of a PLS-PM is a difficult task. It is well known, that PLS-PM is a completely distribution free approach [10]. Thus, standard fit index and inferential process are not yet valid. Moreover, PLS-PM does not seem to optimize a well established global scalar function. Hence, no comparable global goodness of fit criteria are available. Furthermore, it is a variance-based model strongly oriented to prediction. Thus, model validation focuses on the model predictive capability. Following this idea, Amato et al. [1] recently proposed the Goodness of Fit (*GoF*) index. This remains the only available measure to evaluate the global model fitting in a PLS-PM model. Such index has been developed in order to take into account the model performance in both the measurement and the structural model, that is why two different parts compose the index:

$$GoF = \sqrt{\frac{\sum_{q:P_q>1}\sum_{p=1}^{P_q} Cor^2\left(x_{pq}, \hat{\xi}_q\right)}{\sum_{q:P_q>1} P_q} \times \frac{\sum_{j=1}^{J} R^2\left(\hat{\xi}_j, \hat{\xi}_{q:\xi_q\rightarrow\xi_j}\right)}{J}} \qquad (1)$$

where P_q is the number of manifest variables in the q-th block, x_{pq} is the generic manifest variable in the q-th block, $\hat{\xi}_q$ is the generic latent variable score, J is the number of endogenous latent variables in the model and $\hat{\xi}_j$ is the generic endogenous latent variable score.

By looking at Eq. in (1) it is possible to notice that both terms of the product under the square root can be seen as portions of explained variances. As it is well known the R^2 index in a simple regression is an indicator of how well the model fits the data. In fact, the smaller the variability of the residual values around the regression line relative to the overall variability is, the better the prediction obtained by the model is. The residuals play a central role in stating the quality of a model. Following this idea it is possible to rewrite the *GoF* index using residuals as:

$$GoF = \sqrt{\frac{1}{\sum_{q:P_q>1} P_q} \sum_{q:P_q>1} \sum_{p=1}^{P_q} \left(1 - \frac{\sum_{i=1}^{N} e_{ipq}^2}{\sum_{i=1}^{N} \left(x_{ipq} - \overline{x}_{pq}\right)^2}\right)} \times \sqrt{\frac{1}{J} \sum_{j=1}^{J} \left(1 - \frac{\sum_{i=1}^{N} f_{ij}^2}{\sum_{i=1}^{N} \left(\hat{\xi}_{ij} - \overline{\hat{\xi}}_j\right)^2}\right)} \tag{2}$$

where e_{ipq} is the measurement model residual for the i-th unit, corresponding to the p-th manifest variable in the q-th block, i.e. the communality residual, and f_{ij} is the structural model residual for the i-th unit, corresponding to the j-th endogenous block. These two kinds of residuals are the same as used in REBUS-PLS algorithm. For further information about how computing these residuals please refers to Trinchera [9] and Esposito Vinzi et al. [5]. In particular, the communality residuals are the residuals of the simple regressions of each manifest variable on the corresponding latent variable, while the structural residuals are the residuals of the OLS simple and multiple regressions of the endogenous latent variables on their exogenous latent variables.

If more than one class is taken into account, i.e. if the N units are split into K classes each one of size n_k, the GoF index as expressed in Eq. (2) can be reformulated leading to the GQI. Therefore, in the case of K classes the GQI can be expressed as:

$$GQI = \sqrt{\sum_{k=1}^{K} \frac{n_k}{N} \left[\frac{1}{\sum_{q:P_q>1} P_q} \sum_{q:P_q>1} \sum_{p=1}^{P_q} \left(1 - \frac{\sum_{i=1}^{n_k} e_{ipqk}^2}{\sum_{i=1}^{n_k} \left(x_{ipq} - \overline{x}_{pqk}\right)^2}\right)\right]} \times \sqrt{\sum_{k=1}^{K} \frac{n_k}{N} \left[\frac{1}{J} \sum_{j=1}^{J} \left(1 - \frac{\sum_{i=1}^{n_k} f_{ijk}^2}{\sum_{i=1}^{n_k} \left(\hat{\xi}_{ijk} - \overline{\hat{\xi}}_{jk}\right)^2}\right)\right]} \tag{3}$$

This index is equal to the GoF in the case of a unique class, i.e. when $K = 1$ and $n_1 = N$. In other words, the GQI computed for the whole sample as a unique class is equal to the GoF index computed for the global model.

If local models performing better than the global model are detected the GQI index will be higher than the GoF value computed for the global model. As a matter of fact, local models performing better than the global model mean working with residuals that are smaller than the ones computed for the global model. And this directly entails obtaining a higher GQI index than the one obtained for the global model. Of course, the GQI can be considered as an average of the class specific GoF index. Nevertheless, expressing the GQI as in Eq. (3), allows us to directly compare

the same index among different partitions of the units (and with the aggregate solution of the global model too).

To assess the quality of the detected partition it is possible to perform a permutation test procedure [3] involving T random replications of the unit partition (keeping constant the group proportions as detected by the chosen clustering method). In this way an empirical distribution of the *GQI* index will be obtained. The *GQI* of the partition obtained by the chosen clustering method will be compared to the empirical distribution in order to assess if the detected partition performs better than a random assignment of the units, and better than the global model.

In the next section a simulation study to investigate the properties of the *GQI* is presented. The use of *GQI* to asses unit partition in a real case application is shown in [4].

3 Simulation Study

3.1 Design of the Numerical Example and Data Simulation

This simulation study aims at testing the *GQI* capability in assessing unit partition in two different situations, i.e. when the simulated data are affected by unobserved heterogeneity and the simulated local models really differ as regards model parameters, and when the simulated data are strictly homogenous, i.e. when the simulated local models do not differ. Here, a simple marketing type model will be used. The postulated model is composed of one latent endogenous variable, *Customer Satisfaction*, and two latent exogenous variables, *Price Fairness* and *Quality* (cf. Fig. 1). Each latent exogenous variable (*Price Fairness* and *Quality*) has five manifest variables (reflective mode), and the latent endogenous variable (*Customer Satisfaction*) is measured by three indicators (reflective mode). Here, we want to assess if in case of heterogenous data, the partition showing the highest *GQI* is the one whit the highest prediction power, i.e. the simulated one. This study intentionally uses a clear cut example of a marketing related path model for data simulation

Fig. 1 Experimental model

purposes. The data generation procedure is based on the LISREL-type approach. In other words, once the model parameters are established, the data are generated according to the implied covariance matrix, using a specific SAS-IML© macro developed by the author. For both the simulation schemes two latent classes, each of 200 units, are supposed to exist. Thus, the data on the aggregate level for each of the numerical examples includes 400 units. Moreover, for each of the postulated simulation scheme 100 sets of simulated data are computed. In total, the analysis involves 200 marketing related numerical examples on different sets of simulated data.

3.1.1 Simulation Scheme for the Heterogeneous Data-Sets

Unobserved heterogeneity involving both the structural and the measurement models directly means working with local models that are different as regards both the path coefficient values and the measurement model parameter values (i.e. the loading and outer weight values). In a simple model, as the one postulated above, heterogeneity in the model implies detecting price sensitive consumers, or those requiring price fairness, and consumers who have the strongest preference for another particular product attribute, e.g. quality. For more details on simulation scheme for heterogenous data-sets please refers to Table 1. 100 data-sets keeping the postulated features have been simulated. For each of these 100 data-set the *GQI* index is computed for both the global model (i.e. by computing the residuals of each unit

Table 1 Simulated values for model parameters

	Heterogenous data-sets		Homogenous data-sets
Model parameters	Class 1	Class 2	Both class 1 and class 2
No. of units	200	200	200
Path Coefficients:			
Price→*Sat*	0.9	0.1	0.8
Quality→*Sat*	0.1	0.9	0.8
Loadings *Price*:			
P_1	0.9	0.9	0.9
P_2	0.9	0.9	0.9
P_3	0.1	0.9	0.9
P_4	0.9	0.9	0.9
P_5	0.9	0.9	0.9
Loadings *Quality*:			
Q_1	0.9	0.9	0.9
Q_2	0.9	0.9	0.9
Q_3	0.9	0.1	0.9
Q_4	0.9	0.9	0.9
Q_5	0.9	0.9	0.9
Loadings *Satisfaction*:			
S_1	0.9	0.9	0.9
S_2	0.9	0.9	0.9
S_3	0.9	0.9	0.9

from the global model regardless of the unit membership to a class) and the simulated local models (i.e. by computing the residuals of each unit from its own local model). Afterwards, for each simulated data-set, 100 random replications of the unit partition in two classes (keeping constant the group proportions as simulated) are computed in order to perform a permutation test. In this way an empirical distribution of the *GQI* index is obtained. The *GQI* obtained for the simulated partition is compared to the empirical distribution in order to assess if the detected partition (in our case the simulated partition) performs better than a random assignment of the units, and better than the global model.

3.1.2 Simulation Scheme for the Homogeneous Data-Sets

In the case of homogenous data-sets all the units are supposed to be well described by a unique model. Two fictitious latent classes showing the same model parameters both in the measurement and in the structural models have been simulated [see Table 1]. 100 data-sets keeping the postulated features have been simulated. For each of these 100 data-sets the *GQI* index is computed for both the global model and the simulated fictitious local models. Once again, for each simulated data-set, 100 random replications of the unit partition in two classes (keeping constant the group proportions as simulated) are computed in order to perform a permutation test. Of course, we expect that the *GQI* indexes for both the global model solution and the (fictitious) partitioned data solution are similar. Moreover, we expect that the *GQI* value computed for the partitioned data solution is not an extreme value of the obtained empirical distribution.

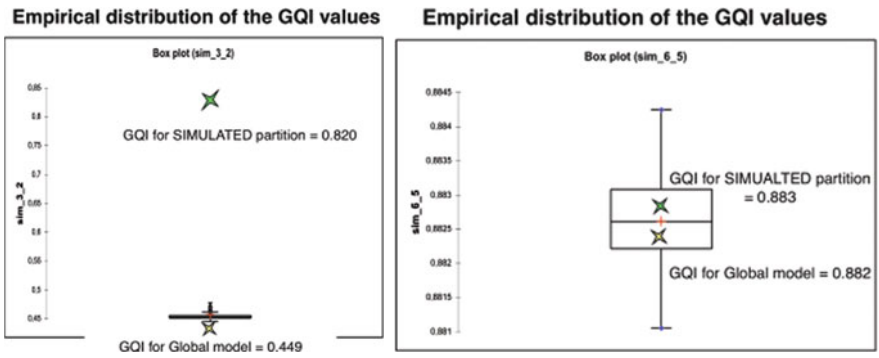
3.2 Simulation Study Results

Following the permutation test approach, each of the 200 data-sets (both homogenous and heterogeneous data) has been randomly divided 100 times into two classes of the same size as the simulated ones. The *GQI* has been computed for each of the random partitions of the units. An empirical distribution of the *GQI* values for a two class partition of the units is therefore obtained for each of the simulated data-sets.

Firstly we present the results obtained for the heterogeneous data-sets. In particular, in Table 2 and in Fig. 2(a) the results obtained as regards one of the 100 simulated heterogenous data-sets are shown. In Fig. 3, instead, the *GQI* distribution for all the 100 heterogeneous data-sets is shown. For each of the simulated heterogenous data-sets, the *GQI* value obtained from the simulated partition of the units, i.e. for real different latent classes, is definitely an extreme value of the distribution (cf. Figs. 2(a) and 3). Moreover, analyzing the box-plot obtained for the empirical distribution of the *GQI* values for a generic heterogenous data-set (cf. Fig. 2(a)), it is possible to notice that the *GQI* computed for the global model (i.e. the *GoF* value computed for the global model) is the smaller value obtained for the *GQI*, except for extreme solutions. This means that a unit partition always surpassed the performance of the global model. In other words, the global model has to be definitely considered as affected by heterogeneity. Moreover, the *GQI* value obtained for the

Table 2 Permutation test results for a generic heterogeneous data-set and a generic homogeneous data-set : simple statistics

Simple statistics	Heterogenous data-set	Homogenous data-set
No. of observations	102	102
Minimum	0.445	0.881
Maximum	0.831	0.884
1^{st} Quartile	0.451	0.882
Median	0.453	0.883
3^{rd} Quartile	0.456	0.883
Mean	0.429	0.883
Lower bound on mean (95%)	0.450	0.883
Upper bound on mean (95%)	0.465	0.882
GQI for SIMULATED partition	0.820	0.882
GQI for the GLOBAL model	0.449	0.883



(a) results for an heterogenous data-set (b) results for an homogenous data-set

Fig. 2 Empirical distribution of the GQI values obtained by permutation test

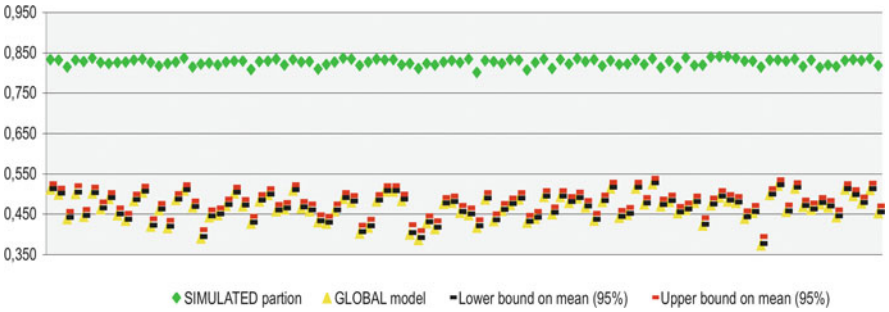


Fig. 3 Permutation test results for all the heterogeneous data-sets

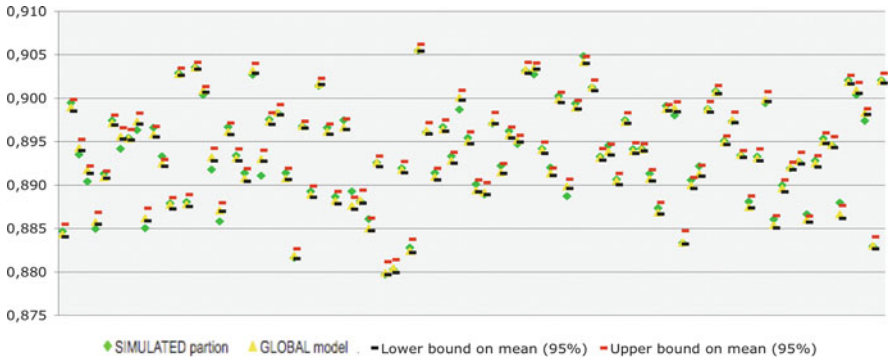


Fig. 4 Permutation test results for all the homogeneous data-sets

simulated partition is the highest obtained value. In Table 2 the simple statistics concerning the empirical distribution of a generic heterogenous data-set are presented. Here we can notice that the *GQI* index computed for the simulated partition is an extreme value as regards the empirical confidence interval ($\alpha = 0.05$) obtaining by permutation test. To conclude, analyzing the Fig. 3, it is possible to notice that similar results are obtained for all the heterogenous simulated data-sets. This allows us to assess that in the case of heterogenous data the simulated partition of the units is better than a random assignment of the units, and is definitely better (in terms of prediction power) than the global model solution.

Results obtained for the homogeneous data-sets are presented in Table 2 and in Figs. 2(b) and 4. Once again results obtained for a generic homogenous data-set are presented in Table 2 and in Fig. 2(b), while the empirical distributions for all the homogeneous data-sets are shown in Fig. 4. Differently form the heterogeneous case, in homogenous data-sets the *GQI* value obtained for the fictitious latent classes is close to the global model ones, as it was obviously expected. As a matter of fact the two latent classes show the same model parameters than the global model. Thus residuals from the local models are similar to residuals computed from the global model. Moreover, random partitions of units in two classes do not improve the predictive power of the models. Following the permutation test approach in the case of homogeneous data-sets no unit partition has to be considered better than the global model solution, i.e. none of the *GQI* values can be considered as an extreme value (cf. Fig. 2(b)). Similar results are obtained for all the homogeneous data-sets. In fact, the empirical confidence interval ($\alpha = 0.05$) for each of the 100 homogeneous data-sets always contains both the global model solution and the simulated one.

4 Discussion and Conclusions

Here, a new index to assess detected unit partition has been presented: the Group Quality Index (*GQI*). This index is a reformulation of the *GoF* index in a multi-group optic. It allows to assess the quality of the obtained unit partition when

performing a clustering method in PLS-PM. This simulation study shows that in the case of homogeneous datasets, the *GQI* computed for a unit partition equals the *GQI* computed for the non partitioned data-set. Instead, in the case of heterogeneous datasets, the *GQI* computed for the *best* unit partition is an extreme value of the *GQI* empirical distribution. Thus, we can conclude that the *GQI* index can be considered as a good indicator to assess if taking into account local models provides better performance (in terms of predictivity power) then using a single model for the whole sample. As future developments are concerned a more complex and more complete simulation study need to be performed so as to consider differences in groups size. Moreover, statistical significance of differences between local parameters needs to be further investigated.

Acknowledgments The Author thanks Vincenzo Esposito Vinzi and Michel Tenenhaus for the invaluable advices.

References

1. Amato, S., Esposito Vinzi, V., Tenenhaus, M.: A global goodness-of-fit index for PLS structural equation modeling. Technical Report, HEC School of Managment, France (2005)
2. Bollen, K.A.: Structural Equations with Latent Variables. Wiley, New York, NY (1989)
3. Edgington, E.: Randomization Test. Marcel Dekker Inc., New York, NY (1987)
4. Esposito Vinzi, V., Trinchera L., Amato, S.: PLS path modeling: from foundations to recent developments and open issues for model assessment and improvement. In: Esposito Vinzi, V., Chin, W., Henseler, J., Wang, H. (eds.) Handbook "Partial Least Squares: Concepts, Methods and Applications", Computational Statistics Handbook Series, vol. II. Springer, Europe 47–82 (2010)
5. Esposito Vinzi, V., Trinchera, L., Squillacciotti, S., Tenenhaus, M.: REBUS-PLS: a response-based procedure for detecting unit segments in PLS-PM. Appl. Stoch. Model. Bus. Ind. **24**, 439–458 (2008)
6. Hahn, C., Johnson, M., Herrmann, A., Huber, F.: Capturing customer heterogeneity using a finite mixture PLS approach. Schmalenbach Bus. Rev. **54**, 243–269 (2002)
7. Kaplan, D.: Structural Equation Modeling: Foundations and Extensions. Sage, Thousands Oaks, CA (2000)
8. Tenenhaus, M., Esposito Vinzi, V., Chatelin, Y.M., Lauro, C.: PLS path modeling. Comput. Stat. Data Anal. **48**, 159–205 (2005)
9. Trinchera, L.: Unobserved heterogeneity in structural equation models: a new approach in latent class detection in PLS path modeling. PhD Thesis, DMS, University of Naples (2007)
10. Wold, H.: Modelling in complex situations with soft information. In: Wold, H. (ed.) Third World Congress of Econometric Society, Toronto, Canada (1975)

Part VI
Latent Variables and Related Methods

Non-Linear Relationships in SEM with Latent Variables: Some Theoretical Remarks and a Case Study

Giuseppe Boari, Gabriele Cantaluppi, and Stefano Bertelli

Abstract The object of the work is to take into account non-linear relationships in path analysis models with latent variables. Some theoretical remarks are made to introduce the context where the presence of non-linearity is to be considered with reference to both the inner and the outer model.

Diagnostic tools to test the existence of a non-linear relationship are also presented, mainly with reference to the so-called Kano model. In particular, a procedure based upon the regression of the response variable, with respect to properly defined dummy variables, is considered.

An application to data coming from a survey on the customers of a financial organization is finally presented.

1 Introduction

The structural equation models (SEM) with latent variables typically consider the following relations

$$\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta} \tag{1}$$

$$\mathbf{x} = \boldsymbol{\Lambda}_{\mathbf{x}}\boldsymbol{\xi} + \boldsymbol{\delta} \qquad \mathbf{y} = \boldsymbol{\Lambda}_{\mathbf{y}}\boldsymbol{\eta} + \boldsymbol{\varepsilon} \tag{2}$$

where the inner model (1) states the structural linear relationship among exogenous latent variables, $\boldsymbol{\xi}$, and the endogenous ones, $\boldsymbol{\eta}$, explained by the matrices of coefficients $\mathbf{B}$, lower triangular, and $\boldsymbol{\Gamma}$; the outer measurement model (2) defines the linear relationship, so-called reflective, among the latent variables, $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$, and the corresponding manifest variables $\mathbf{x}$ and $\mathbf{y}$. When some proxy variables of the formative type are present, see [2], some relations in the measurement model are inverted or of the MIMIC (Multiple Indicators Multiple Causes) type. All previous variables are defined to be the differences from their average values.

G. Boari (✉)
Dipartimento di Scienze statistiche, Università Cattolica del Sacro Cuore, Milano, Italy,
e-mail: giuseppe.boari@unicatt.it

B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-13312-1_30,

However, in several applications, the relations (1) and (2) do not appear to be of the linear type; this is mainly attributable, with no loss of generality, to the following factors:

1. in the structural model
   a. presence of quadratic relations
   b. interaction effects
2. in the measurement model
   a. scaling problems
   b. Kano model relationships.

2 Presence of Non-Linearity in the Inner Model

We remember that the structural model is usually assumed to be of the recursive nature; then a multiple regression approach may be adopted during the Partial Least Squares (PLS) estimating stage of the structural coefficients. Therefore the preceding 1.a and 1.b cases of non-linearity can be easily dealt with, in the inner model, by considering, where necessary, additional variables consisting, in the former case, of the square of the regressors, while in the latter case, of the product of those for which the interaction effect is assumed. These new latent variables are obtained from the latent scores estimated during the iterative stage of the PLS algorithm and with respect to the linear path model initially formulated. For example, with reference to Ping (cf. [5], [6] and [7]), in order to consider also second order relationships, we can formulate the model (1) by considering, for example, the following one

$$\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \mathbf{B}_2\boldsymbol{\eta}^2 + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\Gamma}_2\boldsymbol{\xi}^2 + \boldsymbol{\zeta} \tag{3}$$

where $\boldsymbol{\eta}^2$ and $\boldsymbol{\xi}^2$ are vectors whose elements are the squares of the corresponding elements in $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$; the matrices $\mathbf{B}_2$ and $\boldsymbol{\Gamma}_2$, lower triangular, present non zero coefficients in correspondence to the existence of quadratic recursive relations.

Note that the preceding model (3) is considered only during the phase of estimation of the regression coefficients, while, during the first PLS phase, that is to say the iterative one (aimed at the construction of the latent scores), the estimation procedure is to be treated according to (1), which does not take into account non-linear relationships, since the base model, null hypothesis "of linearity", is being tested.

3 Presence of Non-Linearity in the Outer Model

The non-linear relations, possibly occurring in the reflective measurement model (2), may arise in consequence of several factors; we will take into account only the

two mentioned above, that is the problems deriving from scaling procedures and those concerning the relationships considered by the Kano model.

3.1 Scaling Problems

The non-linear nature of the relations among the generic latent variable and the corresponding connected proxy variables may ensue from the well-known problem of the use of conventional measurement scales, usually employed to gather the data (Likert scales, with a fixed number, k, of steps are typically used). In this case, the approach proposed by Thurstone (cf. [8]), also presented in [9], seems to be the more promising, attaining the following results: linearity among the transformed proxy variables and the corresponding latent variable; distributional normality of the random variables.

Let X be the specific measured variable: we recall that the Thurstone approach assumes that the observations are generated by a monotone transformation of the realizations of a normal random variable, W, which describes the so-called objective "continuum" implicitely used by the interviewed subjects, in providing their evaluations.

A simplified scaling procedure may consist in assuming, for this normal random variable, a mean value corresponding to the sample median $\hat{x}_{0.50}$ and a standard deviation given by the following relation

$$\hat{\sigma} = \max\left(\hat{x}_{0.84} - \hat{x}_{0.50}, \hat{x}_{0.50} - \hat{x}_{0.16}\right),$$

where $\hat{x}_p$ is the quantile of order p of the sample distribution of X. The transformations of the observed scores x_j are then obtained with the inverse of the cumulative distribution function $\Phi_W(\cdot)$ of that normal distribution and the empirical one, $\hat{F}(\cdot)$, according to the following relationship

$$x_j^* = \Phi_W^{-1}(\hat{F}(x_j)) \qquad (j = 1, ..., k-1)$$

and conventionally defining

$$x_k^* = \max\left(x_{k-1}^*, \hat{x}_{0.50} + 3\hat{\sigma}\right).$$

Observe that the empirical distribution $\hat{F}(\cdot)$ is defined by means of the sample distribution of X and $\max(\cdot)$ is now used to avoid that the last scale level will assume a value less than x_{k-1}^*.

3.2 Kano Model Relationships

In other occasions, the non-linear relation among the proxy variables and the latent one may be explained by making use of the Kano models, particularly suitable for the evaluations regarding some tangible aspects chosen to measure the level of quality or satisfaction assigned by the customers to a specific aspect of a product or a service.

We will consider, for instance, the various items of a questionnaire defining the overall evaluation about the personnel of a financial branch: courtesy, technical competence, responsiveness capacity and personal look. We recall that every item is associated to a single element of the sets of manifest variables **x** or **y** .

Relationships (2) may be classified in the following three types (see [4]):

- basic (must-be or expected)
- linear
- attractive (or exciting).

The first and third types are characterized by the behaviour represented in the two drawings of Fig. 1. The identification of the type of relation to be used in a particular situation should be identified, for instance, by observing the scatter-plot diagrams of the available observations for each item and the corresponding levels of the overall satisfaction, properly reconstructed when direct observations are not available.

In particular the x axis represents the perceived level of fulfilment of the attribute, while the y one represents the corresponding satisfaction level. With regard, e.g., to relationships of the "basic" type, a complete fulfilment of the attribute does not improve satisfaction; on the converse, if the attribute is totally unfulfilled, this causes a complete dissatisfaction.

In fact, considering the i^{th} proxy variable X_{ij} $(i = 1, ..., p_j)$ linked to the latent ξ_j, we may distinguish, with reference to the effective availability of the latent scores ξ_j, among the three following situations:

- the scores are estimated by the PLS procedure;
- the scores are obtained by simply summing the observed values;
- the scores are available from a further observed variable (partial overall).

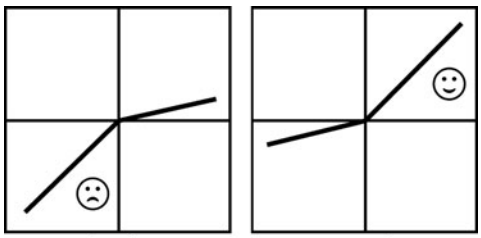


Fig. 1 Types of Kano non-linear relationships *left*: must-be or expected *right*: exciting

Observe that the relationship among the latent variable describing the perceived level of satisfaction and its manifest indicators is, as already mentioned, of the reflective type, while the Kano model assumes that satisfaction depends upon the perceived level of fulfilment of the attribute. This is consistent with our suggestion of testing the null hypothesis "of linearity", since evidence of non-linearity in Kano relationship supports the hypothesis of non-linearity also in relationship (2).

In the literature, two different procedures are suggested, with the aim of identifying the presence of non-linearity, which are based, respectively, on the analysis of proper 2×2 contingency tables, or, on the other hand, on the regression analysis of models referring to appropriate dummy variables.

We recall, however, that the statistics produced by the reliability analysis, which has to be performed, see [10], preliminary to the analysis, for testing the existence and the validity of the latent construct corresponding to ξ_j (under the assumption of additivity of the measures), may give information useful in showing the possible presence of non linear relations; in particular, in addition to the Cronbach alpha and the Dillon-Goldstein Rho indices, the "item to total correlation" and "alpha if deleted" statistics may be used to this purpose.

3.2.1 Contingency 2 × 2 tables

The procedure, presented in [1], may be, in our opinion, effectively employed only when the conventional scale used to gather the data explicitly provides the indifference position (objective zero). In this case, the presence of non linearity may be showed by the analysis of 2×2 contingency tables built as follows.

With reference to the generic manifest variable X_{ij} and to the available scores of the corresponding latent variable ξ_j, let define the events

$$\begin{array}{ll} D = \text{dissatisfaction, if } \xi_j \leq \text{ indifference level} & \bar{D} \text{ otherwise} \\ F = \text{failure,} \qquad\qquad\quad \text{if } X_{ij} \leq \text{ indifference level} & \bar{F} \text{ otherwise} \end{array}$$

and compute the occurrences deriving by the two-way classification showed in the following table

$$\begin{array}{cc|cc} & & F & \bar{F} \\ D & a & b \\ D & c & d \end{array} \tag{4}$$

containing the relative frequencies a, b, c, d.

When both b and c are approximately closed to zero ($b = c = 0$) we can suspect the presence of linearity. Furthermore, we may distinguish between "basic" and "attractive" relationship by observing, before all, that the main interest is the study of X_{ij} as a function of ξ_j and that the group of manifest variables linked to ξ_j should present a mutual positive correlation, being elements of the same measurement scale: the analysis is to be managed, by consequence, with reference to the conditional events $F|D$, $\bar{F}|D$, $F|\bar{D}$ and $\bar{F}|\bar{D}$. In particular, once defined the statistic

$$S = \frac{a}{a+b} - \frac{c}{c+d}\,, \tag{5}$$

it may be observed that $S \to 1$ suggests a "basic" relationship, while $S \to 0$ denotes an "attractive" relationship.

We have to remark that, in order to correctly obtain in practice the entries of the contingency table (4), a proper definition of the indifference value has to be given, also considering that the estimated latent scores do not possess an explicit indefference level. We suggest to use the medians of the involved variables.

3.2.2 Regression with Dummy Variables

In [3] a very simple procedure for detecting the presence of non linearity is suggested; it is based on the regression analysis of the relation among the dependent variable "sub-overall" (the ξ_j scores, both estimated or computed) and the dummy variables H_i and L_i, $i = 1, ..., p_j$, defined as

$$H_i = 1 \text{ if } X_{ij} \geq \text{ third quantile}, \qquad H_i = 0 \text{ otherwise} \tag{6}$$

$$L_i = 1 \text{ if } X_{ij} \leq \text{ first quantile}, \qquad L_i = 0 \text{ otherwise} \tag{7}$$

by estimating the parameters of the following regression model:

$$\xi_j^* = \text{const} + \sum_{i=1}^{p_j} (a_i H_i + b_i L_i)\,. \tag{8}$$

When the coefficients a_i and b_i, relating to the manifest X_{ij}, are significant and have opposite sign, then a linear relation is present between X_{ij} and ξ_j. If only a_i is significant, it is the case of an "attractive" attribute; while if only b_i is significant the attribute X_{ij} may be defined to be of the "basic" type.

3.3 An Application and Concluding Remarks

The procedures previously described were then applied to a real case, in order to go deep into the analysis of a model of Customer Satisfaction and compare their performances. In particular, the applicative example we are introducing concerns the evaluations given by 200 customers on the facilities of a financial service during a wider Customer Satisfaction Survey.

The corresponding reference model is represented by a structural equation model with latent variables and it is used to estimate the overall customer satisfaction level and its relationship with other measurable aspects of the service. Fig. 2 shows that part of the model that refers to the evaluation of the physical appearance of the offices and their practical organization (ξ_1), measured by means of the following three manifest variables:

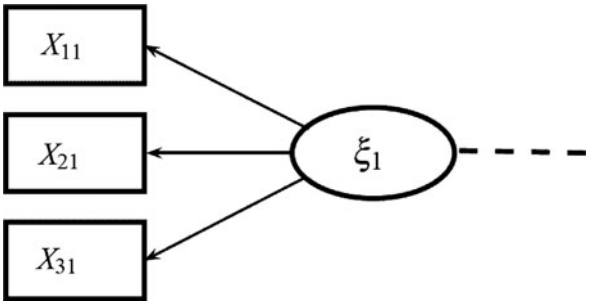


Fig. 2 Path diagram section of the relationships considered

$$X_{11} = \text{opening time}$$
$$X_{21} = \text{neatness of the offices}$$
$$X_{31} = \text{comfort and privacy.}$$

The scores of the latent variable ξ_1 were estimated with a Partial Least Squares (PLS) procedure; then, in order to evaluate more in detail the kind of relationship between ξ_1 and the proxies X_{11}, X_{21} and X_{31}, the dummy variables defined in (6) and (7) were computed and model (8) coefficients estimated.

The Top Managers of the financial organization, that performed the survey, had the suspect that linearity was not definitely appropriate to capture the effective relations among the involved variables; for this reason we proposed to implement this innovative approach to detect the possible presence of non-linearity with regard to the analysis of the particular aspect "facilities" of the site where the financial service is provided.

Table 1 shows the results of the application of the contingency 2×2 tables procedure. Observe that the test for non-linearity may be also performed in a sligthly different way than the form (5) proposed by the authors. One can directly test, for example, the equality of the following percentages

$$\frac{b}{a+b} \qquad \text{versus} \qquad \frac{c}{c+d},$$

whose rejection gives evidence of non-linearity.

Table 2 summarizes the numerical results of the regression with dummy variables procedure.

Table 1 Testing non-linearity by means of contingency 2×2 tables

Attribute	$b/(a+b)$	$c/(c+d)$	z	*p-value*
opening time	0.20	0.26	1.011	0.312
neatness of the offices	0.52	0.13	6.476	**0.000**
comfort and privacy	0.30	0.19	1.873	0.061

Note that in the reference structural model (1) and (2) the variables are considered to be differences from the corresponding average; hence the "const" in the model (8) has zero value.

It can be observed that both procedures suggest that the "neatness of the offices" attribute (see the corresponding boldface $p - value$ figure) presents non-linearity. Furthermore, on the basis of the regression procedure, we may clearly classify this attribute as "basic": it leaves customer not particularly satisfied when fulfilled but dissatisfied if unfulfilled.

The remaining relationships appear to be of the linear type.

Table 2 Parameter estimates of model (8), used to check the presence of non-linear relationships

	L_i		H_i	
Attribute	b_i	$p\text{-}value$	a_i	$p\text{-}value$
opening time	−0.864	0.000	0.250	0.045
neatness of the offices	−0.596	0.000	−0.044	0.703
comfort and privacy	−0.750	0.000	0.613	0.000

To conclude, we can assess that the regression procedure gives a more complete interpretation of the actual relationship existing among variables. It also overcomes the problem of the indifference level definition.

References

1. Conklin M., Pogawa, K., Lipovetsky, S.: Customer satisfaction analysis: identification of key drivers. Eur. J. Oper. Res. **154**, 819–827 (2004)
2. Diamantopoulos, A., Riefler, P., Roth, K.P.: Advancing formative measurement models. J. Bus. Res. **61**, 1203–1218 (2008)
3. Füller, J., Matzler, K., Faullant, R.: Asymmetric effects in customer satisfaction. Ann. Tourism Res. **33**(4), 1159–1163 (2006)
4. Kano, N., Seraku, N., Takahashi, F., Tsuji, S.: Attractive quality and must-be quality. In: Hromi, J.D. (ed.) The Best on Quality, vol. 7, pp. 165–186. International Academy for Quality, ASQ Quality Press, Milwaukee, WI (1996)
5. Ping, A.R. Jr.: A parsimonious estimating technique for interaction and quadratic latent variables. J. Mark. Res. **32**, 336–347 (1995)
6. Ping, A.R. Jr.: Estimating latent variable interactions and quadratics: the state of this art. J. Manage. **22**(1), 163–183 (1996)
7. Ping, A.R. Jr.: Latent variable regression: a technique for estimating interaction and quadratic coefficients. Multivariate. Behav. Res. **31**(1), 95–120 (1996)
8. Thurstone, L.L.: The Measurement of Values. University of Chicago Press, Chicago, IL (1959)
9. Zanella, A.: A statistical model for the analysis of customer satisfaction: some theoretical aspects and a simulation result. Total Qual. Manag. **9**(7), 599–609 (1998)
10. Zanella, A., Cantaluppi, G.: Some remarks on a test for assessing whether Cronbach's coefficient α exceeds a given theoretical value. Stat. Appl. **18**, 251–275 (2006)

Multidimensional Scaling Versus Multiple Correspondence Analysis When Analyzing Categorization Data

Marine Cadoret, Sébastien Lê, and Jérôme Pagès

Abstract Categorization is a cognitive process in which subjects are asked to group a set of object according to their similarities. This task was used for the first time in psychology and is becoming now more and more popular in sensory analysis. Categorization data are usually analyzed by multidimensional scaling (MDS). In this article we propose an original approach based on multiple correspondence analysis (MCA); this new methodology which provides new insights on the data will be compared to one specified procedure of MDS.

1 Introduction

Categorization, also known as sorting task, is a data collection which consists in asking J subjects to partition a set of I objects function of their similarities. Once collected, data are gathered in a co-occurrences matrix which is most of the times analyzes by multidimensional scaling (MDS).

An alternative consists in gathering the data in an individuals$\times$variables data table, where the individuals are the objects and the variables are the subjects considered as qualitative variables: each category corresponds to a group of objects (these data can also be viewed as a concatenation of J incidence matrices where each matrix contains as many rows as there are objects and as many columns as the subject j provides groups). Therefore this data table can be analyzed by multiple correspondence analysis (MCA). This MCA is the core of a new approach to analyze categorization data [2]. In comparison to [2] that refers only to MCA to analyze categorization data, the aim of this article is to compare the analyses of these data by MDS on the one hand and MCA on the other hand.

M. Cadoret (✉)
Laboratoire de mathématiques appliquées, Agrocampus Ouest, 35042 Rennes Cedex, France
e-mail: marine.cadoret@agrocampus-ouest.fr

B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-13312-1_31,

2 Methods

2.1 Multidimensional Scaling

Let C denotes the co-occurrences matrix of dimensions $I \times I$ in which the general term $c(i, l)$ corresponds to the number of subjects that have put the objects i and l in a same group. From C, several matrices of distance can be calculated. A first one, denoted D, in which the general term $d_{MDS1}(i, l)$ corresponds to the number of subjects that haven't put i and l together, is calculated:

$$d_{MDS1}(i, l) = J - c(i, l)$$

This matrix is the one the most used in sensory analysis: Lawless [10], Lawless et al. [11], Faye et al.[6].

A second possible matrix of distances [3] is:

$$d_{MDS2}(i, l) = \sqrt{2(J - c(i, l))}$$

In MDS, once the distance is chosen, two types of MDS are proposed: metric MDS on the one hand and non-metric MDS on the other hand. Both MDS work from a dissimilarities or distances matrix but the metric MDS is based on the distances whereas the non-metric MDS is based on the ranks (in this case, the use of one of the two distances is equivalent).

When analyzing categorization data, the analyst uses usually non-metric MDS; but this choice is done without giving a real justification. In this context the distances themselves have a real meaning with respect to the data and deserve to be considered as such which brings us to consider only the case of metric MDS. In the sequel when we use the acronym MDS it will refer to metric MDS.

In metric MDS, when a distance d is chosen, a matrix of scalar products between objects, denoted W, is computed using the Torgerson's formula. The general term of this matrix denoted $w(i, l)$ is obtained the following way:

$$w(i, l) = -\frac{1}{2}(d^2(i, l) - d^2(i, .) - d^2(., l) + d^2(., .)),$$

where $d^2(i, .) = \frac{1}{I}\sum_{j=1}^{I} d^2(i, j)$ and $d^2(., .) = \frac{1}{I^2}\sum_{j=1}^{I}\sum_{i=1}^{I} d^2(i, j)$.

This matrix is then diagonalized in order to obtain the coordinates of the objects in a new coordinate space. This procedure is called principal coordinates analysis [1, 8].

When the distance used is Euclidean, all the eigenvalues are positive or null; when the distance is not Euclidean, some eigenvalues can be negative. In this last case an Euclidean approximation is realized for example in considering only the eigenvectors associated with positive eigenvalues (this way of doing is optimal [4, 7]).

2.2 Multiple Correspondence Analysis

Categorization data are gathered in a table of dimensions $I \times J$ in which each row i corresponds to an object, each column j corresponds to a subject, and the cell (i, j) corresponds to the label of the group to which the object i belongs, for the subject j. Each column of the table can be assimilated to a qualitative variable with K_j categories, where K_j denotes the number of groups used by the subject j.

The data table is of type individuals $\times$ qualitative variables and thus concerns multiple correspondence analysis. In this analysis, the data are taken into account via the so-called disjunctive table, denoted X, which comprises here I rows and $K = \sum K_j$ columns of general term x_{ik} which is equal to 1 if object i belongs to group k and 0 if not.

MCA provides a representation of the objects that is obtained similarly to the one provided by PCA (by maximizing the inertia of the projected scatter plot on a new coordinate basis). This set of objects lies in a K-dimensional space (more precisely, considering the constraints in MCA, it lies in a $(K - J)$-dimensional space). In this space, the distance between two objects i and l is thus defined:

$$d^2_{MCA}(i, l) = \frac{1}{J} \sum_k \frac{I}{I_k} (x_{ik} - x_{lk})^2$$

where I_k denotes the number of objects in the group k. For this (Euclidean) distance, (1) two objects i and l are superimposed if they were put together by all the subjects, (2) two objects are all the more distant as they were placed in two different groups by a great number of subjects.

More precisely, a group k (associated with subject j) contributes to this distance in a way inversely proportional to its size: the assignment to a group of small size moves an object away from all the others.

2.3 Elements of Comparison Between the Two Methods

Firstly we compare MDS to MCA by looking at the distances (induced by each method) between two elements on the one hand and between one element and the center of gravity on the other hand. After that, we look at some helps for interpreting provided for each method.

Distances between two objects. In MDS, for the first transformation, the chosen distance d_{MDS1} between 2 objects i and l can also be expressed through the disjunctive data table:

$$\begin{aligned} d_{MDS1}(i, l) &= J - c(i, l) \\ &= \frac{1}{2} \sum_k (x_{ik} - x_{lk})^2 \end{aligned}$$

Since $(x_{ik} - x_{lk})^2$ is equal to the absolute value of the difference, d_{MDS1} appears as a distance of L^1-type which is not Euclidean. This distance corresponds to the square of the distance of MCA up to the coefficient $1/I_k$.

If we use the second transformation, the distance of MDS becomes:

$$\begin{aligned} d_{MDS2}(i,l) &= \sqrt{2(J - c(i,l))} \\ &= \sqrt{\sum_k (x_{ik} - x_{lk})^2} \end{aligned}$$

As this distance corresponds to the distance of MCA up to the coefficient $1/I_k$, it is Euclidean. This leads to prefer this second transformation to the first one. In the sequel it is this second transformation which is considered.

For both, whatever the transformation used, contrary to MCA, the size of the groups provided by the subjects is not apparently integrated in the distance between two objects.

Remarks: As only MCA seems to take into account the size of the groups, it will be interesting to consider the case where the I_k are constant: it corresponds to the case where the number of groups and the number of objects per group are the same for all the subjects. This may happen in practice as a constraint in some experiments [9]; from a theoretical point of view this case is worth of interest since the direct effect of the size of the groups on the distances is no longer effective. For the second transformation, when the I_k are constant, the distance of MDS corresponds to the the distance of MCA (up to a coefficient $\frac{I}{JI_k}$ to be meticulous, but this constant doesn't change the shape of the cloud).

Distances to the centre of gravity. To specify the role of each object in the analysis, we compute its distance to the centre of gravity (denoted G_I) of the whole objects. In MDS in the case of an Euclidean distance (which is the case of d_{MDS2}), the square distance of an object to the centre of gravity can be expressed through the Torgerson's formula :

$$\begin{aligned} d^2(i, G_I) &= w(i,i) \\ &= d^2(i,.) - \frac{1}{2} d^2(.,.) \\ &= \frac{1}{I} \sum_{l=1}^{I} d^2(i,l) - \frac{1}{2} d^2(.,.) \end{aligned}$$

In the particular case of d_{MDS2}, it can be rewritten:

$$d^2_{MDS2}(i, G_I) = J - \frac{1}{I^2} \sum_k I_k^2 - \frac{2}{I} \sum_k x_{ik} I_k,$$

where the first two terms are constant and only the third one depends on i. In this distance, the size of the groups is integrated : an object i is all the more distant from the centre of gravity as it is often isolated.

In MCA, this distance is obtained the following way:

$$d^2_{MCA}(i, G_I) = \frac{I}{J} \sum_k \frac{x_{ik}}{I_k} - 1,$$

where one can find the impact of I_k. As for MDS, in MCA an object i is all the more distant from the centre of gravity as it is often put in a group of small size.

Remarks: When the I_k are constant, with respect to the distance induced by MCA all the objects i are equidistant from the centre of gravity and this distance is equal to:

$$d^2_{MCA}(i, G_I) = \frac{I}{I_k} - 1$$

In MDS for the second transformation, all the objects are also equidistant from the centre of gravity:

$$d^2_{MDS2}(i, G_I) = J(1 - \frac{I_k}{I})$$

But contrary to MCA, this distance depends on the number of subjects.

Helps for the interpretation. MDS and MCA are based on a singular value decomposition which provides eigenvalues. These eigenvalues provide an indication of the quality of representation associated with each dimension. In MCA, it exists a further interpretation: the eigenvalue associated with a dimension corresponds to the mean of the correlation ratios between the different variables (subjects) and this dimension; the eigenvalues are therefore interpretable. Thus, an eigenvalue λ_s of 1 corresponds to a situation where the correlation ratio between this dimension s and each subject is of 1: therefore, this dimension is a common structure to all the subjects. In categorization, an eigenvalue of 1 corresponds to the case where an object (or a group of objects) was systematically isolated by all the subjects. The disjunctive table once reordered reveals a diagonal block structure on the data. In MCA, an axis associated with an eigenvalue of 1 opposes this object (or this group of objects) to all the others. There is no equivalent property in MDS.

Because MCA works from individual data (and MDS only from aggregate data), different results and representations can be added to MCA. These further representations are very useful for the user. They constitute a complete factorial approach for sorting task data (FAST) described in [2]; some elements are just mentioned here. First this approach provides a representation of the objects and of the categories. In addition it provides a representation of the subjects which can be interpreted jointly. This representation of the subjects is obtained by using the equivalence between MCA and multiple factor analysis (MFA) [5]: to do so each subject is considered

as a group of one variable. Elements of validation by means of confidence ellipses around the objects are also available and are obtained by re-sampling the subjects. In the case of qualified categorization (when people are asked to describe the groups by some words), the label of each group can be the words used to describe the groups. In this case, the optimal representation of the labels/categories provided by MCA becomes a representation of the words.

3 Application

3.1 Data

In order to compare the two approaches (MDS and MCA) we use a data set in which the I_k are not constant (because in this case the results are similar) and more particularly in which one object has been systematically isolated by all the subjects. In this example (cf. Table 1), 3 subjects have realized a categorization on 10 objects and they all have isolated the object J.

Table 1 Categorizations of 3 subjects on 10 objects (*left*) and associated co-occurrences table (*right*)

	S1	S2	S3
A	1	1	1
B	1	1	1
C	1	2	3
D	3	1	2
E	2	3	1
F	2	2	2
G	3	2	3
H	3	3	2
I	2	3	3
J	4	4	4

	A	B	C	D	E	F	G	H	I	J
A	3	3	1	1	1	0	0	0	0	0
B	3	3	1	1	1	0	0	0	0	0
C	1	1	3	0	0	1	2	0	1	0
D	1	1	0	3	0	1	1	2	0	0
E	1	1	0	0	3	1	0	1	2	0
F	0	0	1	1	1	3	1	1	1	0
G	0	0	2	1	0	1	3	1	1	0
H	0	0	0	2	1	1	1	3	1	0
I	0	0	1	0	2	1	1	1	3	0
J	0	0	0	0	0	0	0	0	0	3

3.2 Case of an Object Isolated by All the Subjects

As expected, the first eigenvalue of MCA is equal to 1 (cf. Table 2) and the first axis opposes this object to the others (cf. Fig. 1). The second dimension of MCA opposes the objects A and B (always in a same group, cf. Table 1) to the others.

Concerning MDS, the first dimension opposes objects A and B to the others and therefore corresponds to the second dimension of the MCA (cf. Table 3). The dimensions 2 and 3 which have the same eigenvalue are defined up to a rotation (they correspond to the third and fourth of MCA).

In MDS, the singular case of J appears only on the fourth dimension (cf. Table 3). It is explained by its contribution to the total inertia (cf. Table 4) which is relatively much more important in MCA than in MDS. Thus, in MDS, the association between

Table 2 Eigenvalues and percentage of variance on the dimensions of MCA and MDS

		Dim.1	Dim.2	Dim.3	Dim.4
Eigenvalues	MCA	1	0.667	0.478	0.478
	MDS	6	4.303	4.303	3.6
Percentage of variance	MCA	33.33	22.22	15.94	15.94
	MDS	27.78	19.92	19.92	16.67

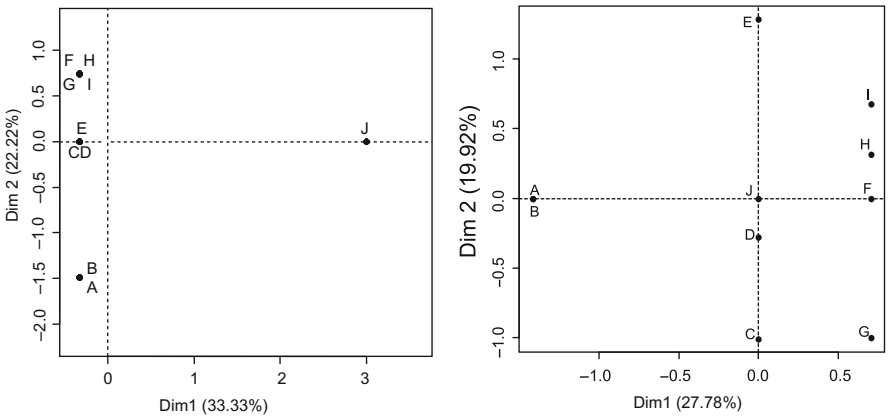


Fig. 1 Representation of the objects in the first plane of MCA (*left*) and MDS (*right*)

A and *B* contributes more to the total inertia than the object *J* itself and appears on the first dimension. Whereas in MCA, the object *J* is so distant to the center of gravity (cf. Table 4) that it is singled out on the first dimension.

This example shows the difference in the hierarchy of the dimensions in MCA and MDS. And especially that an object which is systematically isolated by all the subjects (and so very particular) will appear always on the first dimension of the MCA but not necessary in MDS. However, this particularity is a major feature of the data set which as such must be appear first.

Table 3 Correlation coefficients between factors of MCA and MDS

		MCA			
		Dim.1	Dim.2	Dim.3	Dim.4
MDS	Dim.1	0	1	0	0
	Dim.2	0	0	0.998	−0.063
	Dim.3	0	0	−0.063	−0.998
	Dim.4	1	0	0	0

Table 4 Distances of the objects to the centre of gravity for MCA and MDS

	A	B	C	D	E	F	G	H	I	J
MCA	2.333	2.333	2.333	2.333	2.333	2.333	2.333	2.333	2.333	9
MDS	2.04	2.04	2.04	2.04	2.04	2.04	2.04	2.04	2.04	3.24

4 Conclusion

Practitioners often analyze categorization data using MDS. Another approach using MCA has been proposed. In this paper we compare these two approaches and we show that MCA has two main advantages of high importance:

- The principle of integrating the size of the groups when computing the distances between them is crucial: in particular in the borderline case where one object (or a set of objects) is systematically isolated from the others by the subjects this will be emphasized by MCA (and not by MDS).
- The possibility to supplement the representation of the objects with some other graphical and numerical outputs (just mentioned here but developed in [2]).

References

1. Benzécri, J.-P.: Sur l'analyse factorielle des proximités. Publication de l'ISUP. **13**(4), 235–281 (1964)
2. Cadoret, M., Lê, S., Pagès, J.: A factorial approach for sorting task data (FAST). Food Qual. Prefer. **20**, 410–417 (2009)
3. Cox, T.F., Cox, M.A.A.: Multidimensional Scaling. Chapman & Hall, London (1994)
4. D'AUBIGNY, G.: Vers un renouveau du positionnement multidimensionnel, 4e journées MODULAD organisées par le CISIA (1998)
5. Escofier, B. Pagès, J.: Analyses Factorielles Simples et Multiples. 3eme édn. Dunod, Paris (1998)
6. Faye, P., Brémaud, D., Durand Daubin, M., Courcoux, P., Giboreau, A., Nicod, H.: Perceptive free sorting and verbalization tasks with naive subjects: an alternative to descriptive mappings. Food Qual. Prefer. **15**, 781–791 (2004)
7. Fichet, B.: Distances and Euclidean distances for presence-absence characters and their application to factor analysis. In: De Leeuw, J., Heiser, W.J., Meulman, J.J., Critchley, F. (eds.) Multidimensional Data Analysis, pp. 23–46. DSWO Press, Leiden (1986)
8. Gower, J.C.: Some distance properties of latent root and vector methods in multivariate analysis. Biometrika **53**, 325–338 (1966)
9. Healy, A. Miller, G.A.: The verb as the main determinant of the sentence meaning. Psychon. Sci. **20** 372 (1970)
10. Lawless, H.T.: Exploration of fragrance categories and ambiguous odors using multidimensional scaling and cluster analysis. Chem. Senses **14**, 349–360 (1989)
11. Lawless, H.T., Sheng, T., Knoops, S.: Multidimensional scaling of sorting data applied to cheese perception. Food Qual. Prefer. **6**, 91–98 (1995)

Multidimensional Scaling as Visualization Tool of Web Sequence Rules

Antonio D'Ambrosio and Marcello Pecoraro

Abstract Web Mining can be defined as the application of Data mining processes to Web data. In the field of Web Mining, we distinguish among *Web Content Mining*, *Web Structure Mining* and *Web Usage Mining*. Web Content Mining is the Web Mining process which analyze various aspects related to the contents of a web site such as text, banners, graphics etc. Web Structure Mining is the branch of Web Mining that analyze the structure of the Net (or a sub-part) in terms of connection among the web pages and their linkage design. Finally, Web Usage Mining goal is to understand the usage custom behaviors of web sites users. Within the context of Web Usage Mining, *pattern discovery* and *pattern analysis* allow to *profile* users and their preferences. The sequence rules are association rules ordered in time.
Given a data set coming from a web site which is characterized by a sequence of visits, the proposal is to understand the differences among browsing sections through a Multidimensional Scaling solution, and then obtain a graphical tool which allows to visualize in a new way the sequence rules. The resulting application is half way between Web Usage Mining and Web Structure Mining.

1 Introduction

Web Mining is the Data Mining process applied to Data coming from a single website, a group of websites or from a server [5, 6]. Usually, Web Mining is divided into three main branches: Web Content Mining, Web Usage Mining and Web Structure Mining [8]. The input data of Web Usage Mining process comes usually from log-files or tracking applications. The data are usually connection and visit information (time of connection, visited page, downloaded document etc.). Instead, in Web Content, normally, the data are the text contained in the pages (or specific words like in text mining). In a Web Structure framework we consider the linkage scheme between the page of a website, or between pages published various website. Our idea is to use the traditional input of a Web Usage Mining process to understand the

A. D'Ambrosio (✉)
Department of Mathematics and Statistics, University of Naples Federico II, Naples, Italy
e-mail: antdambr@unina.it

B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-13312-1_32,

similarity among pages of a website, a typical target of a Web Structure Analysis, and then to visualize the relative sequence rules. In this sense, our application is half way between structure and usage. In other words we propose an alternative site map based on "indirect opinions" made by website users. We complete this map with the indication of the stronger relationship between pages, instead of linkage schema usually adopted in these cases. This approach recalls the known Process Mining idea, in which the analyzed website is seen as a unique entity, a unique process where the link, the usage and the content are strictly related and considered together.

2 The Idea

This work provides a visualization method based on the use of Multidimensional Scaling for applications where data are not collected in a direct-way, but in the case that data are collected by an automatic system (tracking) used for the regular server activity. As it is known, Multidimensional Scaling has four main purposes [3]: it is an explorative technique, it can be used for testing structural hypotheses, it is a technique for exploring psychological structures, it can be used as a model of similarity judgments. We propose to understand the differences among browsing sections through sessions analysis via MDS as exploratory tool. The idea is to obtain a similarity visualization among web pages based on user's visit habits. Recall that MDS attempts to model similarity or dissimilarity as distances among points in a geometric space. On such extent web sections are "similar" when a given navigation path is treaded by several users. For this reason we talk about *implicit behaviors* because similarity among web navigation sessions is not obtained in explicit way [4]. Particular attention will be dedicated to data pre-processing task and to the choice of the most appropriated distance measure for the specific problem.

3 The Data

The dataset used for the application comes from UCI machine learning repository. It refers to about a million of navigation sessions collected in a single day on msnbc.com, an American general purpose portal and from the news related portion of msn.com. All the web-pages of this website are originally grouped into seventeen sections (Frontpage, News, Tech, Local etc.). On the rows there are all the single navigation sessions. The columns represent the clicking path. On the first column, there is the first section visited, on the second column the second section visited and so on (Table 1).

Raw data have been processed by putting into columns all of the seventeen section of the website, then for each navigation session the time each section was visited has been counted. The cells contain the absolute frequency of a given visited section in a certain session, namely the total number of clicks made by the users in a given

Table 1 The structure of the data

Session	First section visited	Second section visited	Third section visited	Fourth section visited	Fifth section visited	Sixth section visited	...
1	Frontpage	Sports	News	News	Weather		
2	Frontpage	Opinion	Local	Tech	Opinion	Opinion	Living
3	Weather	Travel	Tech				
4	News	News	News	Local	On-air	Frontpage	
5	BBS	Travel	Business	Travel	Living	Living	Living
6	Frontpage	Sports	Local	Sports	News	Opinion	
...	...	...	...	...	...	...	...

Table 2 The processed data

Session	Frontpage	News	Tech	Local	Opinion	On-air	...
1	1	2	0	0	0	0	...
2	1	0	1	1	3	0	...
...	...	...	...	...	...	...	...

section of the website (Table 2). As the processed dataset is in fact a frequency table, the chi-square distance is used as distance measure for the multidimensional scaling solution because it is the most appropriate in dealing with frequencies.

4 The MDS Solution

Figure 1 shows the Multidimensional scaling solution, the goodness of fit indexes are shown in the Table 3. Note how both STRESS and fit measures are quite good. Recall that STRESS measures could be close to zero as well as fit measures could be close to one.

Looking at the two-dimensional MDS solution, the *front page* is a particular section because it is very often the access to the web site and it also is the bridge for the access to the other sections. Note how sections as *Sports* and *Wheater* differ from *News*, *Local* and *On air*. These are specific sections that are normally visited with specific interests. Note also how the section *Sports* differs from the section *News* even if they can be considered as belonging to the macro-category News.

The first dimension can be interpreted as the *popularity of the sections* as well as the second one can be seen as the *informative content of the sections*. Indeed in the upper side of the figure, there are *Weather*, *Sports* and *MSN sports*, the most specialized contents of the web-page. Instead, in the lower side, there are the most "general" informative web-pages like *News*, *Local* and *On-air*. This kind of analysis seems to work good because it shows how users of a web site remark the differences among sections with their *implicit* behavior. With this approach, the definition of a similarity among web site sections is based on user's behavior. In this way MDS solution gets a graphical representation of the web-site structure seen with the eyes of a particular group of people.

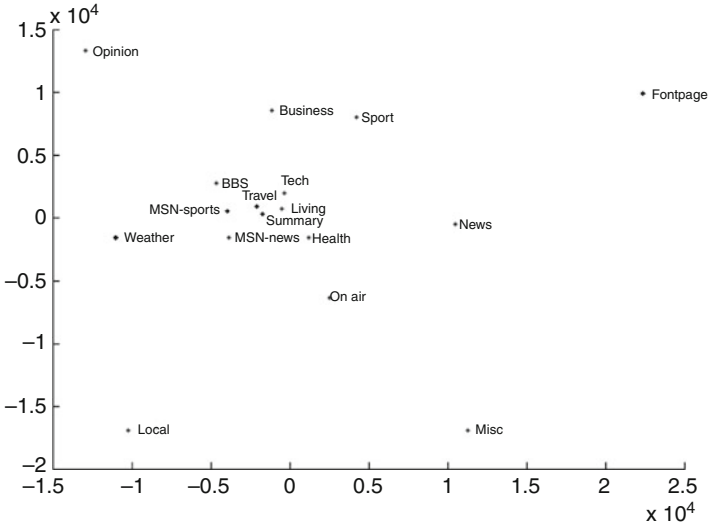


Fig. 1 MDS solution

Table 3 Goodness of fit measures

STRESS and Fit Measures	
Normalized raw STRESS	0.029
STRESS-I	0.171
STRESS-II	0.313
D.A.F.	0.971
Tucker's Coefficient of Congruence	0.985

5 Direct and Indirect Sequence Rules

The graphical MDS solution can be usefully used to visualize the (direct and indirect) sequence rules, namely the association between structures. Association rules are statements to find interesting rules between two or more objects in a large database [1, 7]. A rule is interesting if it satisfies minimum support and minimum confidence threshold (strong rule), so a rule $G \Rightarrow L$ is strong if its support and its confidence respect a minimum threshold.

Let G be an item (in our case, a section) called *antecedent* and let L be an item (a section) called $consequent$. Recall that:

$$Support_{G \Rightarrow L} = P(G \cap L);$$
$$Confidence_{G \Rightarrow L} = P(G \cap L)/P(G),$$

where the support is the proportion of observations in the union of the antecedent and consequent (hence, it is the estimate of the probability of simultaneously observing both the antecedent and the consequent), and the confidence is the support of the

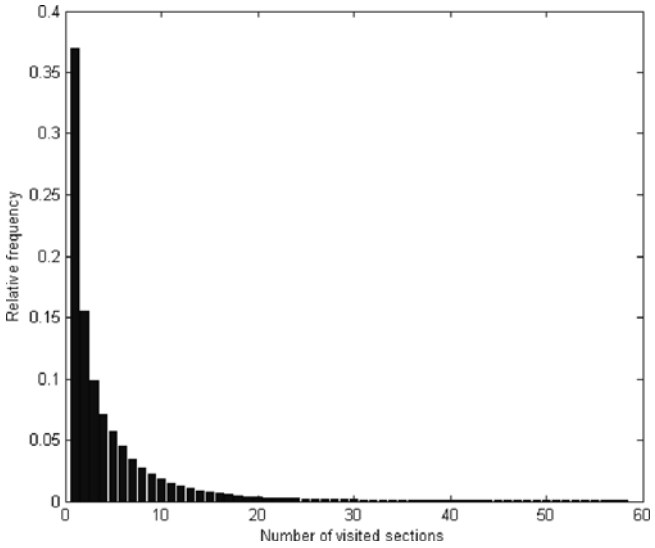


Fig. 2 Bar chart of visited sections

previously mentioned rule divided by the support of the antecedent (hence, it is the estimate of the probability of the consequent given the antecedent).

When the rules are ordered in time, these statements are called sequence rules because G comes first L. Following Blanc and Giudici [2], we distinguish between indirect and direct rules.

A sequence rule is:

- indirect, when between the visited section G and the visited section L, one can visit other sections;
- direct, when one visits first the section G and then, sequentially, the section L.

Looking at Fig. 2, it can be noted that about 37% of visitors gets only one "click" on the web site, as well as more than 50% of visitors leave the web site after no more than two clicks.

For that reason the sequence rules have been computed for visitors with no less than ten clicks in their navigation session.

Notice that visitors can get several clicks on the same section in the same navigation session because each section contains a group of related pages. For instance, the section *Sport* can contain several pages dedicated to different sports.

Figure 3 shows the MDS solution of this reduced version of the dataset and it represents the visualization of the indirect rules. Normalized Raw STRESS and STRESS-I indexes are, respectively, equal to 0.028 and 0.167.

Circles which represent the points in the geometrical space are proportional to the support of the single section *with itself* in the same navigation session.

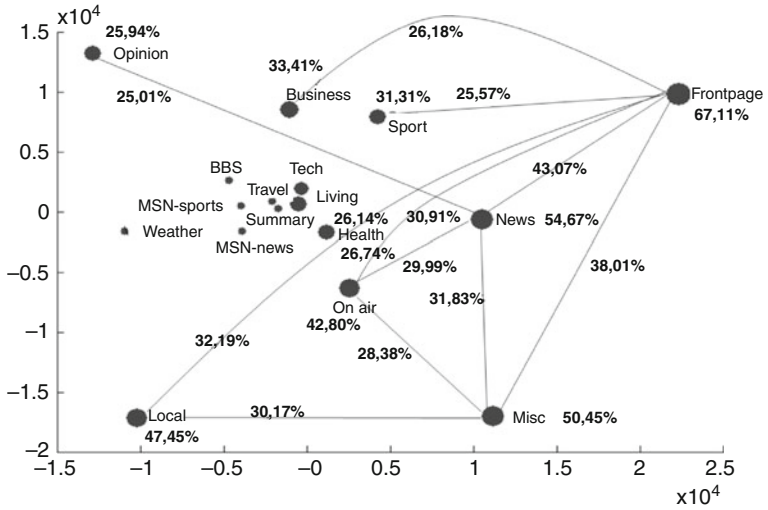


Fig. 3 Indirect rules. Circles are proportional to the support of the single section with itself. Lines connect the support between two sections

Lines connect the support between two sections. The minimum support considered was equal to 0.2.

Note that most of the time the same section is visited (not consecutively) two or more times in the same navigation session. Important indirect rules are also $Frontpage \Rightarrow Misc$, $Frontpage \Rightarrow Business$, $Frontpage \Rightarrow News$, $Frontpage \Rightarrow Local$, $News \Rightarrow Opinion$.

More interesting is the visualization of the direct rules (see Fig. 4). In this case circles are proportional to the support of the single section with itself for two sequential clicks. The lines connect support between two sections, as in the previous figure. Following the suggestion of Blanc and Giudici [2], an imaginary section called *Start* has been added in the figure. In this way, the figure shows at the first click which section has been visited by visitors (just looking at the line between the start section and any other). In this case the minimum support considered was equal to 0.03. The most visited section on the first click is *Frontpage*, *On air* and *News*. The figure emphasizes that most of the sections are visited (at least) two times consecutively, but it depends obviously on the nature of this specific data set: recall that a given section can include several pages.

Looking at both figures, the sequences $Frontpage \Rightarrow News$ and $Frontpage \Rightarrow Misc$ are the ones which are confirmed in both indirect and direct rules. It is straightforward to note that the closeness of the sections in the geometrical space is independent from the sequence rules. The MDS configuration shows how two or more sections are similar given the navigation habit of the users, but it does not mean that there is a clicking sequence among close sections. It is an alternative graphical tool to visualize the structure of a web site compared to the classical site-maps.

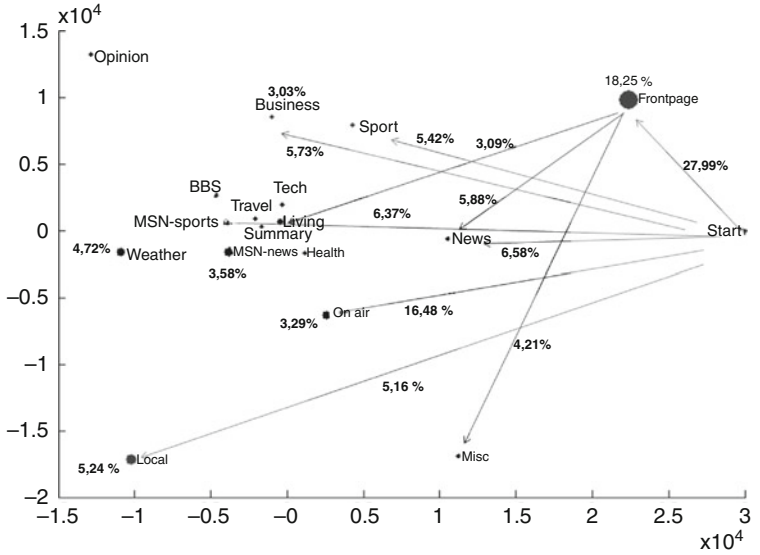


Fig. 4 Direct rules. Circles are proportional to the support of the single section with itself. Lines connect the support between two sections

6 Conclusions

The Multidimensional Scaling analysis allows to visualize the sections of the web site in terms of their similarity. This similarity is governed by the frequency wherewith visitors of the web site visit the sections. In this sense we talk about implicit behaviors of visitors: the MDS solution gets a graphical representation of the websites structure as seen with the eyes of their users. By adding the sequence rules to the MDS solution, it is possible to visualize the connection between the visited sections (governed by given association rules) through a graphical representation which permits simultaneously to keep the similarity structure (or the preference structure) of the web sections. Indeed this kind of analysis returns an interesting graphical tool which allows to represent in a geometrical space the visited sections according their clicking frequency. This visualization allows to better represent, in a subsequent step, the sequence rules because it is a non-random configuration of the web sections (or the web pages) in the geometrical space, and it is an alternative representation compared to the classical site-maps. With such an approach, it is possible investigate about the association among structures but not about the navigation paths: future works will be addressed to this last topic.

Acknowledgments Authors are grateful to the anonymous referees for their useful and interesting comments.

References

1. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In: Proceedings of 20th International Conference on Very Large Data Bases, Santiago, Chile, pp. 487–499 (1994)
2. Blanc, E., Giudici, P.: Sequence Rules for Web Clickstream Analysis. Advances Data Mining. Lecture Notes in Computer Science 2394/-1, pp. 1–14 (2002)
3. Borg, I., Groenen, P.J.F.: Modern Multidimensional Scaling, p. 614. Spriger, New York, NY (2005)
4. D'Ambrosio, A., Pecoraro, M.: Web structure mining through implicit behaviors via multi-dimensional scaling. In Proceedings of the 1st Joint Meeting of the Société Francophone de Classification and the Classification and Data Analysis Group of the Italian Statistical Society (SFC-CLADAG 2008), Book of Short Papers, pp. 261–264 (2008)
5. Etzioni, O.: The world wide web: quagmire or gold mine. Commun. ACM **39**(11), 65–68 (1996)
6. Giudici, P.: Data Mining, p. 424. McGraw-Hill, Milano (2001)
7. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning, p. 536, Springer, New York, NY (2001)
8. Pecoraro, M., Siciliano, R.: Statistical Methods for User Profiling in Web Usage Mining. In: Song, M., Wu, Y.B., (eds.) Handbook of Research on Text and Web Mining Technologies, pp. 359–368. Chapter XXII, Idea Group Inc., Hershey, PA, (2008)

Partial Compliance, Effect of Treatment on the Treated and Instrumental Variables

Antonio Forcina

Abstract Under the assumption that treatment assignment has no direct effect on the response, a non parametric probabilistic model of the distribution involving the latent confounder under partial compliance leads to a generalized definition of the effect of treatment on the treated and reveals that the instrumental variable estimand equals a suitable average of such causal effects only when certain restrictions hold. An application to a popular data set concerning reduction of cholesterol level is used as an illustration.

1 Introduction

The literature on causal inference under non ignorable imperfect compliance is extensive; most contributions are based on the notion of potential outcomes and the distribution of counterfactual quantities. Hernán and Robins [9], in the case of binary compliance, clarify the relation between structural mean models and describe conditions under which the instrumental variable estimator equals the effect of treatment on the treated. When the response is binary, Vansteelandt and Goetghebeur [15] extend structural mean models to allow for logistic link; Goetghebeur and Molenberghs [8] propose a model for the generalized effect of treatment on the treated closely related to the one to be discussed below. The approach based on the notion of principal strata [5] concentrates on estimating the effect of treatment on compliers (see, [1] for the case when compliance is binary and [10], for the extension to the context of partial compliance). Efron and Feldman [3] formulate a parametric regression model which allows them to estimate che overall causal effect. In the case of a binary response, Ten Have et al. [14] examine the properties of an estimator of the marginal causal odds ratio. In the case when treatment assigned, treatment received and response are all binary, Balke and Pearl [2] have computed sharp bounds for the overall average causal effect of the treatment by formulating a plain probabilistic model of the latent distribution within a causal diagram. Geneletti

A. Forcina (✉)
Dipartimento di Economia, Finanza e Statistica, 06100 Perugia, Italy,
e-mail: forcina@stat.unipg.it

B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-13312-1_33,

and Dawid [7] show that definitions and results concerning the effect of treatment on the treated may be derived within a *Decision theoretic* approach to causal inference.

This paper extends the probabilistic formulation of Balke and Pearl [2] to the case where treatment received and response are not binary and concentrates on the average effect of treatment on the treated. With partial compliance, this effect may be defined as a function $\Delta(t_1; t_0)$ giving the improvement that subjects who self selected the amount of treatment $T = t_0$ would have received by taking $T = t_1$ as compared to $T = 0$. We show that, under a set of conditions which extend those given for instance by Hernán and Robins [9], the instrumental variable estimand is a suitable average of quantities of the form $\Delta(1; t)$, the effects of full compliance among those who self selected $T = t$. Our results reveal that the usual instrumental variable estimand implies a simple linear model of the form $\Delta(1, t)t = \Delta(t; t)$.

After formulating an unrestricted latent class model in Sect. 2, in Sect. 3 we derive the relation between the instrumental variable estimand and the effect of treatment on the treated. Advantages and limitations of a latent class approach to causal inference are discussed in Sect. 4. The data analyzed by Efron and Feldman [3] are used in Sect. 5 to exemplify and discuss the methods introduced in the paper.

2 A Latent Class Model for Partial Compliance

In the following we assume that a binary treatment Z is randomized within a population, with 0 denoting control and 1 active treatment. Unobservable individual heterogeneity, which may affect compliance behavior and response, will be represented by a qualitative discrete latent variable U and $\mathcal{U}$ will denote the set of its possible levels. Let T denote the actual amount of active treatment taken by a given unit, this is assumed to depend on Z and U; for notational convenience, we assume that T is discrete and varies from 0 (no treatment) to 1 (full compliance). Let Y denote a binary, ordered qualitative or quantitative discrete response which is assumed to be independent of Z given U and T; an assumption known as exclusion-restriction. When Y is binary, 0 will denote failure and 1 success; when Y is ordered, we assume that a suitable set of scores measuring success of treatment have been assigned. Notice that, for patients in the control arm, T measures the amount of active treatment eventually taken, if available.

These statements may be translated into the following causal directed acyclic graph (DAG) which is identical to the one used for example by Balke and Pearl [2] (Fig. 1)

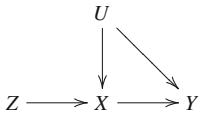


Fig. 1 Causal graph for the basic model of partial compliance

2.1 Notation

In the following, when no ambiguity may arise, conditioning variables will be denoted by their value, for instance, $P(Y = y \mid u, t)$ will mean $P(Y = y \mid U = u, T = t)$. The relevant features of the manifest distribution conditional on Z may be determined by parameters of the form $p_{tz} = P(T = t \mid z)$ and $E(Y \mid t, z)$. For the latent distribution, let the marginal of U be denoted by $\boldsymbol{\pi}$, a vector with elements $P(U = u)$, $u \in \mathcal{U}$, $\boldsymbol{\tau}_{tz}$ denotes the vector whose elements are the compliance probabilities $P(T = t \mid u, z)$, $\forall u \in \mathcal{U}$; similarly $\boldsymbol{\theta}_t$ will be the vector containing the conditional expectations $E(Y \mid u, t)$.

A crucial ingredient of the model is the posterior distribution of the latent U given T, Z which may be interpreted as the distribution of the latent within specific subsets of the population characterized by their behavior relative to treatment and compliance. This probability may be expanded by using marginal independence between U and Z

$$P(U = u \mid t, z) = \frac{P(U = u)P(T = t \mid u, z)}{\sum_u P(U = u)P(T = t \mid u, z)}.$$

Let $\boldsymbol{\pi}_{tz}$ be the vector whose elements are the posterior probabilities $P(U = u \mid t, z)$, $u \in \mathcal{U}$; because the denominator equals $P(T = t \mid z) = p_{tz}$, we may write,

$$\boldsymbol{\pi}_{tz} = \operatorname{diag}(\boldsymbol{\tau}_{tz})\boldsymbol{\pi} / p_{tz}. \tag{1}$$

3 The Effect of Treatment on the Treated

Within the causal DAG presented above, the causal effect of T within a given value of the latent U may be measured by the difference

$$E(Y \mid T = x, u) - E(Y \mid T = 0, u);$$

when u varies in $\mathcal{U}$, this defines the vector $\boldsymbol{\delta}(x) = \boldsymbol{\theta}_x - \boldsymbol{\theta}_0$ of causal effects across the latent that would be obtained by enforcing $T = t$. By averaging the elements of $\boldsymbol{\delta}(x)$ with the posterior probabilities of self selecting a given level $T = t$ of compliance, we obtain the effect of treatment on the treated generalized to the case of partial compoliance

$$\Delta_z(x; t) = \boldsymbol{\delta}(x)'\boldsymbol{\pi}_{tz}, \forall x > 0, \tag{2}$$

this may be interpreted as the average causal effect of an amount $T = x$ of treatment among those who took $T = t$. For the special case of binary compliance see Hernán and Robins [9], p. 367, Heckman (discussion of [1]) and Geneletti and Dawid [7]. When $t \neq x$, $\Delta_z(x; t)$ is not a proper conditional expectation, nonetheless it is a meaningful causal parameter.

3.1 The Instrumental Variable Estimand

Because Z is binary, the probabilistic analog of the so called instrumental variable estimator may be written in the form

$$\delta_{IV} = \frac{Cov(Y, Z)}{Cov(T, Z)} = \frac{E(Y \mid Z = 1) - E(Y \mid Z = 0)}{E(T \mid Z = 1) - E(T \mid Z = 0)}.$$

For what follows it is useful to state the two following results whose proofs are given in the Appendix:

Theorem 1 *The vector of conditional covariances may be decomposed as*

$$Cov(Y, Z \mid \boldsymbol{u}) = \sum_{t>0} diag(\boldsymbol{\tau}_{t1} - \boldsymbol{\tau}_{t0})\boldsymbol{\delta}(t). \tag{3}$$

This equation may be interpreted as the latent class analog of Eq. (9) in Angrist et al. ([1], Sect. 2.2), extended to the case of T non binary. It indicates that, conditional on the latent, the overall effect of Z on Y may be decomposed as the sum of the products of the effect of Z on T times the effect of T on Y.

By using (3) it is possible to express the numerator of the instrumental variable estimand in terms of the partial effects of treatment on the treated

Theorem 2 *The numerator of δ_{IV} may be written as*

$$Cov(Y, Z) = \sum_{t>0} [\Delta_1(t; t)(p_{t1} - p_{t0}) + p_{t0}(\Delta_1(t; t) - \Delta_0(t; t))]. \tag{4}$$

This expression is useful to relate the instrumental variable estimand to an average causal effect as described below.

When T may take one or more values between 0 and 1 and the treatment is not available in the control arm, (4) implies that

$$\delta_{IV} = \frac{\sum_{t>0}(\Delta_1(t; t)/t)(tp_{t1})}{\sum_{t>0} tp_{t1}}; \tag{5}$$

this may be interpreted as follows: under a linear structural mean model Robins [12] saying that $\Delta_1(t; t) = t\psi$, clearly $\delta_{IV} = \psi$. Under the slightly more general model $\Delta_1(t; t) = t\Delta_1(1; t)$, δ_{IV} equals the average effect of full compliance across subjects who self selected different values of $T = t$ with weights proportional to tp_{t1}. This suggests the following extension of the instrumental variable estimand: let $g(t)$ be a known function which is strictly increasing and such that $g(0) = 0$ and $g(1) = 1$; then if we assumed that $\Delta_1(t; t) = g(t)\Delta_1(1; t)$, the appropriate measure of causal effect would be provided by

$$\delta_{IV,g} = \frac{Cov(Y, Z)}{Cov(g(T), Z)}$$

which is still an average of $\Delta_1(1; t)$'s where heavier weights for stronger compliance depend on the g function. Possible instances of $g(t)$ are: t^k, $(\exp(t) -$

$1)/(\exp(1) - 1)$, $log(t + 1)/log(2)$. Though a model for the causal ratio $\Delta_1(t;t)/\Delta_1(1;t)$ cannot be identified from the data alone, when the experiment is double blind and compliance is measured in the control arm, $\Delta_1(t;t)$ may be identified under reasonable assumptions (see [3]) and its shape may allow to explore possible specifications of $g(t)$; an investigation along these lines is contained in Sect. 5.

4 Discussion

This paper describes a plain probabilistic model for the identification of treatment effects under partial compliance. This is not to claim that the latent class formulation proposed here is superior, the idea being that different formalism may be seen as different "languages", "each with their virtues and vices", as Lauritzen [11] put it. Because the latent variable is meant to represent all possible confounders as for instance in Balke and Pearl [2], the causal effects as defined at the beginning of Sect. 3 are not merely probabilistic quantities but a close analog of individual causal effects. By taking this philosophical attitude, the paper indicates that the plain probabilistic formulation reveals certain features, like the bivariate structure of the effect of treatment on the treated extended to partial compliance and its relation to the instrumental variable estimator, which seems to add something new to the understanding of the problem. A latent class model may also be useful as a data generating tool to examine which latent construction is necessary for the causal quantities $\Delta_z(x,t)$ to be, for instance, constant in t and increasing in x or to satisfy the condition of "no effect modification by Z".

Though latent class models imply the existence of perfectly homogeneous latent groups of experimental units; this is mainly a conceptual rather than a substantial restriction as long as we are not interested in estimating the full latent distribution but simply certain average causal effects. Forcina [4], in the much simpler case of binary compliance, in an attempt to identify the full latent distribution, was forced to assume deterministic latent class models. According to Frangakis [6] *intervention* is not a fundamental concept within the latent class formulation, but this is a merely philosophical objection. According to Robins et al. [13] a latent class formulation is isomorphic to a causal DAG; as such, there does not exist any parameter which represents the effect of treatment on the treated and proposed a partially deterministic DAG where such parameter exists. As far as I can understand, this means that a "G-formula" does not exist and thus does not contradicts the results given in Sect. 3.

5 Application

Efron and Feldman [3] (EF for brevity in the following) describe and analyze in detail a placebo-controlled double-blind randomized clinical trial aiming to measure the efficacy of cholestyramine for lowering cholesterol level. Because treatment was not available in the control arm, we may assume that $P(T = t \mid Z = 0) = 0$, thus the instrumental variable estimate is also equal to the average causal effect of

Table 1 Average causal effects among the treated by deciles of compliance

Deciles	D_1	D_2	D_3	D_4	D_5
t	0.04	0.14	0.27	0.41	0.59
$\Delta_1(t;t)$	0.6	10.2	4.0	11.6	18.9
$\sigma(\Delta_1(t;t))$	19.4	20.3	18.3	26.7	25.4
Deciles	D_6	D_7	D_8	D_9	D_{10}
t	0.77	0.90	0.94	0.97	0.99
$\Delta_1(t;t)$	29.9	27.8	42.0	49.5	55.1
$\sigma(\Delta_1(t;t))$	27.4	32.2	31.3	29.2	27.0

treatment on the treated in its extended formulation given in (5); this is equal to 41.4 with a s.e. of 3.54, estimated with the delta method.

Because compliance to placebo was also recorded, EF assumed that an unbiased estimate of $\Delta_1(t;t)$ may be obtained by comparing the average response among treated and untreated with the same quantile level t of compliance. The estimates of $\Delta_1(t;t)$ and their s.e. within deciles of compliance are displayed in Table 1. EF used the corresponding quantile estimates to fit various parametric models; by assuming that $\boldsymbol{\pi}_{t1} = \boldsymbol{\pi}_{t0}$ and that $\boldsymbol{\theta}_0'\boldsymbol{\pi}_{t1}$ is linear in t, they were able to estimate the value $\bar{t}$ for which $\Delta_z(\bar{t};\bar{t})$ equals the causal effect on the whole population: $(\boldsymbol{\theta}_{\bar{t}} - \boldsymbol{\theta}_0)'\boldsymbol{\pi}$. This estimate is then extended linearly to full treatment giving a value of 34.5 ± 4.8; this is smaller than the IV estimator which gives higher weights to the more compliant patients.

Here we use these estimates simply to try different $g(t)$ functions in the model $\Delta_1(t;t) = g(t)\Delta_1(1;t)$ and then compute corresponding generalized instrumental variable estimators. The model is clearly non identifiable because both $g(t)$ and $\Delta_1(1;t)$ depend on t. Table 2 below gives the R^2 measure of fit by weighted least squares, of several alternative forms of $g(t)$ and the corresponding generalized instrumental variable estimate and standard errors under the special assumption that $\Delta_1(1;t)$ was constant. Though the cubic model fits best, we believe that it is too much affected by the last two deciles and provides probably an overestimate of the causal effect. On the basis of these estimates, the quadratic model provides, perhaps, the most reliable estimate.

Jin and Rubin [10] analyzed the same data by assuming the existence of several Principal Strata of subjects characterized by their level of potential compliance relative to treatment and placebo and derived estimates of posterior medians within certain typical Strata in a parametric bayesian model. Their estimate for full compliers (when the correlation between the two compliance behaviors is assumed to be

Table 2 Comparison of different generalized instrumental variable estimators

Model for $\Delta_1(x;x)$	R^2	Estimate	s.e.
ϕx	0.83	41.4	3.33
ϕx^2	0.89	51.6	3.35
ϕx^3	0.91	59.2	3.43
$\phi \frac{\exp(x)-1}{\exp(1)-1}$	0.88	45.9	3.32

0) equals 50 with a s.e. of about 3.2, a result which is in close agreement with the generalized instrumental variable estimate based on a quadratic model.

Acknowledgments The author would like to thank D.R. Cox, P. Dawid, T. Richarson and T. Vanderwheele for helpful discussions.

References

1. Angrist, J., Imbens, G.W., Rubin, D.B.: Identification of causal effects using instrumental variables. J. Am. Stat. Assoc. **91**, 444–472 (1996)
2. Balke, A., Pearl, J.: Bounds on treatment effects from studies with imperfect compliance. J. Am. Stat. Assoc. **92**, 1171–1178 (1997)
3. Efron, B., Feldman, D.: Compliance as an explanatory variable in clinical trials. J. Am. Stat. Assoc. **86**, 9–26 (1991)
4. Forcina, A.: Causal effects in the presence of non compliance: a latent variable interpretation. Metron **64**, 275–301 (2006)
5. Frangakis, C.E., Rubin, D.B.: Principal stratification in causal inference. Biometrics **58**, 21–29 (2002)
6. Frangakis, C.: Discussion of Forcina (2006), pp. 292–299 (2006)
7. Geneletti, S., Dawid, P.: Defining and identifying the effect of treatment on the treated. Research Report, Department of Epidemiology and Public Health, Imperial College, London (2007)
8. Goetghebeur, E., Molenbergs, G.: Estimating efficacy in placebo-controlled clinical trials with ordered non-compliance. J. Am. Stat. Assoc. **91**, 928–934 (1996)
9. Hernán, M.A., Robins, J.M.: Instruments for causal inference. An epidemiologist's dream? Epidemiology **17**, 360–372 (2006)
10. Jin, H., Rubin, D.B.: Principal stratification for causal inference with extended partial compliance: application to Efron-Feldman data. J. Am. Stat. Assoc. **103**, 101–111 (2008)
11. Lauritzen, S.L.: Discussion on causality. Scand. J. Stat. **31**, 189–192 (2004)
12. Robins, J.M.: Correcting for non-compliance in randomized trials using structural mean models. Commun. Stat. Theory Methods **12**, 2379–2412 (1994)
13. Robins, J.M., Vanderwheele, T.J., Richardson, T.: Discussion of Forcina (2006), pp. 288–301 (2006)
14. Ten Have, T.R., Joffe, M., Cary, M.: Causal logistic models for non compliance under randomized treatment with univariate binary response! Stat. Med. **22**, 1255–1283 (2003)
15. Vansteelandt, S., Goetghebeur, E.: Causal inference with generalized structural mean models. J. R. Stat. Soc. **65**, 817–835 (2003)

Appendix

Proof (Propositions 1 and 2) By using the fact that $Z \perp U$ and $Y \perp Z \mid T, U$ and the identity $\boldsymbol{\tau}_{0z} = 1 - \sum_{t>0} \boldsymbol{\tau}_{tz}$,

$$\begin{aligned} E(Y \mid Z) &= \sum_t \sum_u E(Y \mid u, t, Z) P(t \mid u, Z) P(u) \\ &= \sum_t \boldsymbol{\theta}_t' \mathrm{diag}(\boldsymbol{\tau}_{tz}) \boldsymbol{\pi} = \sum_t \boldsymbol{\theta}_t' \boldsymbol{\pi}_{tz} p_{tz} = \theta_0' \boldsymbol{\pi} + \sum_{t>0} (\boldsymbol{\theta}_t - \boldsymbol{\theta}_0)' \mathrm{diag}(\boldsymbol{\tau}_{tx}) \boldsymbol{\pi}. \end{aligned}$$

The numerator of the instrumental variable estimator $E(Y \mid Z = 1) - E(Y \mid Z = 0)$ may then be expanded as follows,

$$= \sum_{t>0} (\boldsymbol{\theta}_t - \boldsymbol{\theta}_0)' \mathrm{diag}(\boldsymbol{\tau}_{t1} - \boldsymbol{\tau}_{t0}) \boldsymbol{\pi} = \sum_{t>0} (\boldsymbol{\theta}_t - \boldsymbol{\theta}_0)' (\boldsymbol{\pi}_{t1} p_{t1} - \boldsymbol{\pi}_{t0} p_{t0})$$
$$= \sum_{t>0} [\Delta_1(t;t)(p_{t1} - p_{t0}) + p_{t0}(\Delta_1(t;t) - \Delta_0(t;t)).$$

Method of Quantification for Qualitative Variables and their Use in the Structural Equations Models

C. Lauro, D. Nappo, M.G. Grassia, and R. Miele

Abstract The article is about the problem of the treatment of qualitative variables in the Structural Equation Models with attention to the case of Partial Least Squares Path Modeling. In literature there are some proposals based on the application of known statistical tecniques to quantify the qualitative variables. Starting from these works we propose an external quantification for only qualitative variables by the Alternating Least Squares, obtaining the optimal quantification (vectors of optimal scaling), a future objective to develop an algorithm that computes simultaneously the vectors of optimal scaling and the optimal regression coefficients, between the variables. We will present an application of our method to a real dataset.

1 Introduction

In social and marketing research the study of qualitative indicators to explain some theories is very important, and their use to clarify the casual relationship is always more frequent. The presence of qualitative indicators in the estimation of casual models, as a Structural Equations Model (SEM [4]), is very frequent, as we remember that in this kind of model generally the ordinal variables that we know to be qualitative variables are used. The escamotage is to assume the continuity for these variables, even if the scale of measure could be different between the variables. With this assumption we can use all statistical techniques created for the study of quantitative variables. This practice has a large use in the literature, especially for the Partial Least Squares-Path Modeling (PLS-PM, [10]), in which distributional hypothesies on data are not necessary. The problems for the estimation are with a major frequence in presence of dicotomic variables (0/1, absence/presence) or with variables expressed on small scale of values (3 or 4 values): in these cases it is easy

C. Lauro (✉)
Department of Mathematics and Statistics, University "Federico II" Naples, 80125 Naples, Italy
e-mail: clauro@unina.it

B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-13312-1_34,

to have a problem of significativity of the estimation because the assumption of continuity of variables is not correct.

We consider a scenario in which we have to estimate a model PLS-PM with qualitative variables (nominal, ordinal) and we propose a different metodology that allows us to obtain an external quantification in order to use the classical algorithm of PLS-PM, to obtain the estimation of the relationship. We have used this approach to estimate a model, in which the latent blocks are composed by nominal and ordinal manifest variables.

2 Different Ways to Quantify the Qualitative Variables

Generally in statistics to analyze a qualitative variable is used the binary coding or the association of an integer number to the modalities (but without the numeric significant) in order to use the quantitative techniques. Sometimes in the PLS-PM with this coding we observe that the estimation of the relationship are not significant, because the coefficients have the interval confidence around the zero. In the context of covariance SEM approach to the estimation of the parameters of a model, the problem is solved by the estimation of the tetracoric/policoric and poliserial correlation.[1] For the PLS-PM a useful validation procedure does not exist when we have to introduce in the model the qualitative indicators to obtain more information. However there are some proposals made by Jakobowicz and Derquenne [3] and P.G. Lovaglio (2002), that try to individuate some solutions to the problem. E. Jakobowicz and C. Derquenne [3] propose an algorithm, called Partial Maximum Likelihood (PML), based on the Generalized Linear Models (GLM), in which they take into account of the different nature of variables (nominal, numerical or ordinal), whose final aim is to obtain a quantification of qualitative variables and the estimation of model parameters. The analysis then continues by performing the classical PLS-PM algorithm. They modify the first step of the PLS-PM algorithm, according to the nature of manifest variables (nominal or ordinal). The authors introduce the concept of reference variable as the initial estimation of the latent variable: it is a manifest variable of any latent block associated to the j-th block that is supposed to better explain the latent concept. The vector of the initial weights will be equal to (Lohmöller [7] has demonstrated that for any initial vector of weights the algorithm of PLS-PM converges):

$$w_{jh}^{0} = cov(x_{jh}, x_{i1}) \tag{1}$$

where x_{i1} is the reference variable chosen between the blocks associated to j block. The authors propose a series of statistical methodologies well known in

[1] The tetracoric correlation coefficient, introduced by Pearson, is the estimated correlation coefficient of two continuous variables distributed as a normal, underlying two dicotomic ordinal variables. The correlation between a continuous and dicotomic (politomic) variable is called *biserial* (*poliserial*) correlation coefficient.

literature, whose differences between themselves are related to the nature of the variable x_{jh} and the reference variable x_{i1}. In particular if the reference variable is numeric and the manifest variable is nominal an ANOVA model will be used. In the opposite case the chosen model will be a logistic one. If they are both categorical a Logistic model with one effect will be used, while if the reference variable is nominal with r categories and the manifest variable is numeric, a polytomic Logistic model will be chosen; if the reference variable is nominal and the manifest variable nominal a Logistic model with one effect will be applied.The inner estimation is the same as in the classical PLS-PM algorithm, while for the outer estimation it is important to consider the nature of the manifest variables. Another proposal in the literature is of P.G. Lovaglio [8], that proposes an algorithm for the estimation of a latent variable measured by causes and indicators. The nature of the observed variables can be nominal, ordinal or numerical: the algorithm computes a regression model in which there are a set of manifest variables X that are explicative, and a set Y of manifest variables that are dependent and that define a latent variable. The algorithm estimates, alternating two steps, the best quantification for the variables X and Y (in the case in which the sets are composed by qualitative variables) and the best estimation of the parameters of the model. The algorithm proposed belongs to the family of Alternating Least Square Optimal Scaling (ALSOS; [2, 9]), and in particular it is based on the join between two approaches: one is the Non Linear Regression of the set Y on X to obtain the optimal quantification for both variables, and the second is the Principal Component Analysis to obtain the estimation of the latent variable as the first component of $\hat{Y}'\hat{Y}$. These two methods, that forms the two steps are alternating until the convergence and the results are the estimation of the regression coefficients and the optimal quantification for the qualitative variables. It is obtained the convergence, because at each iteration the residual is smaller than the previous ieration; the aim is the maximization of the redundance index and of the multiple R^2. The proposal of Jakobowicz and Derquenne is very interesting, but it is costituited by methods based on distributional hypothesis, as the ANOVA or the Logistic model, that is beyond of the characteristics of PLS-PM. Besides in the proposal of Jakobowicz and Derquenne the weights of the manifest variables are not univocally determined, because their estimation depends by their measurement level. The proposal of Lovaglio, based on an iterative algorithm, has characteristics similar to the PLS-PM and its applicability to each kind of variable, it makes it more useful and adaptable. The fundamental characteristic of this algorithm is the simultaneous estimation of the vector of scaling and of the parameters of the model, in this case the regression coefficients. So it has the same characteristics of the PLS-PM with the added of an unique function to optimize respect the parameters of the model. The purpose of this proposal is the estimation of a latent variable and not of a SEM model, obtainig as results the quantification and the estimation of the parameters of the regression model with mixed variables. In the context of PLS-PM and following the previous works (Jakobowicz and Derquenne, Lovaglio) we propose a different method for the external quantification, by (the ALSOS algorithm).

3 Alternating Least Squares Algorithm

Given a data matrix $n \times m$ of metric variables, Principal Component Analysis (PCA) is a common technique to reduce the dimensionality of the data set, projecteing variables into a subspace $\Re_p$ where $p \ll m$. The ECkart-Young theorem states that this classical form of linear PCA can be formulated by means of a loss function. Its minimization leads to a $n \times p$ matrix of component scores and an $m \times p$ matrix of component loadings. The actual computerprograms for PCA impose some restrictions about the completeness of the data matrix and interval measurement of variables. In the social science is usually not justified the assumption of interval scales, and often the data matrix are incomplete. In this case it is possible to use the Nonlinear Principal Component Analysis (NPCA), where the term nonlinear pertains to nonlinear transformation of the observed variables. According to the Gifi [2] terminology the NPCA can be defined as homogeneity analysis with restrictions on the quantification matrix Y_I. The ALS algorithm generalizes the approach of PCA to general types of variables. The Non linear PCA in the ALSOS system is derived as homogenety analysis with some costraints; the loss function is

$$\sigma(X; Y_1, ..., Y_I) = J^{-1} tr(X - G_i Y_i)'(X - G_i Y_i) \tag{2}$$

with the constraint of rank-one

$$Y_i = q_i \beta_i' \quad \text{with} \quad i \in I \tag{3}$$

The constraints are imposed on the multiple category quantifications, with q_i a l_i column vector of single category quantifications for variable i, and β_i a p-column vector of weights (component loadings). In this way each quantification matrix Y_i is restricted to be of rank one, that is the quantifications in p dimensional space are proportional to each other. With the introduction of the rank one restrictions it is possible to have multidimensional solutions for object scores with a single quantification for the categories of the variables, and it is possible to introduce the measurement level of the variables in the analysis. At this point it is necessary to take into account of the restrictions imposed by the measurement level of the variables. This means that we have to project the estimated vector $\hat{q}_i$ on the cone C_i: in the case of ordinal variables the cone C_i is the cone of monotone transformation [5] given by $C_i = q_i | q_i(1) \leq q_i(2) \leq .....q_i(l_i)$. So the projection is obtained across a monotone regression in the metric D_i (weights). In the case of numerical data the corresponding cone is a ray given by $C_i = \{q_i | q_i = \gamma_i + \delta_i s_i\}$, where s_i is a given vector, for example, the original variable quantifications. So the projection problem is a regression problem; for nominal variables the cone is the $\Re_i^l$ space and the projection is done by simply setting $q_i = \hat{q}_i$, so $\hat{Y}_i = \hat{q}_i \hat{\beta}_i'$ and the algorithm proceeds with the estimation of the object scores. This solution is reffered in the literature as the PRINCALS (Principal Components analysis by Alternating Least Squares; [1]) solution (principal component analysis by means of alternating least squares). The PRINCALS model allows the data analyst to treat each variable differently; some

way be treated as multiple nominal and some others as single nominal, ordinal or numerical. Moreover, with some additional effort one can also incorporate in the analysis nominal variables of mixed measurement level, that is variables with some categories measured on an ordinal scale (e.g. Likert scale) and some on a nominal scale (e.g. categories in survey questionnaires corresponding to the answer "do not know").

3.1 Alternating Least Squares Algorithm: The Model for AVSI

The proposed methodology has been applied to the AVSI[2] database obtained by a statistical research [6] made in three Countries of Africa (Rwanda, Uganda and Kenya) with the aim to evaluate if and in which measure there were changes, for the children belonged to the AVSI project, in the hygienic-sanitary condition, in the education and in the environment, after 1 year from the start of the program. The research has regarded a sample of 1,254 children, being this number proportional to the number of children in each Country. Face to face interviews have been performed to fill standardized questionnaires composed by 204 variables, most of which are qualitative. By analyzing the variables and the possible relationship between them it was possible to determine a SEM model that was estimated with the PLS-PM algorithm. It must be remarked that the questionnaire did not create for this kind of analysis, so many problems came up during the application of this technique. The model, called "Status of the child", has as aim to evaluate the factors that impact on the life condition of the children, and it is composed by eight latent variables: three latent endogenous blocks summarize the Status of the child, Family characteristics, House (the characteristics of the house where children live), Avsi intervention, based on three sub latent variables describing the supports offered to the family, for the school and nutrition. An outcome block of the model is associated to guardian satisfaction depending on the general Status reached by the Child in the year of the survey. The problems that came up in the application are relative to the treatment of qualitative variables, codifying as dichotomy (0/1), many of which have been eliminated from the model, because they weren't statistically significant for the model. The model developed after the quantification shows an improvement both in terms of loss of information (a smaller number of manifest variables are eliminated from the model with respect to the other in which same blocks remain with only two manifest variables), in terms of fitting the model to the data, and in terms of significant results. The blocks that presented a major number of non significant manifest variables (m.v.) are "Guardian satisfaction" that passes from four to two manifest variables and "Housing condition" that passes from four to

[2] AVSI Foundation is an international, not-for-profit, non-governmental organization (NGO) founded in Italy in 1972. AVSI has programs in over 40 countries in Africa, Latin America, Eastern Europe, the Middle East, and Asia. AVSI has implemented several programs in education, healthcare, construction, emergency response, water and sanitation, food and nutrition, and psychosocial support for children, adults and even the elderly persons in the community.

two manifest variables; the block "Support for the school" loses two manifest variables in both models, while "Nutritional support" loses only one indicator for both models. The block that remains unchanged is "Support for family",while "Status of child", in the model without quantification, has four manifest variables versus five in the other model. Family environment, instead, has four manifest variables in the model with quantification versus three in the other. So the reduction in the number of eliminated variables is an advantage for the determination of the latent concept, because through the quantification the scale of variables is extended, resulting in better significance for the determination of the causes of Status of children belonging to the AVSI program. The improvements are also in terms of fitting of the model to the data and in terms of the significance of the casual relationship between the latent variables (see Tables 1 and 2).

The results of the inner model without the quantification show that the relationship between "Status of child" and "Guardian Satisfaction" is not significative because the confidence interval is around the zero and the value of the T Statistic is small, then it has been eliminated from the model. In the quantified model, instead, this relationship is significative, even if it keeps having a value of T Statistic high. Regard to the other casual relationship, we can note that the block AVSI intervention in the model quantified has a relative contribute lower than in the other model (respectively 3.48 and 20.82): in the model quantified we have an improvement in the blocks with more manifest variables that impact on the Status of child (the block Housing condition has a contribute of 40.28, while Family environment of 56.23). For the block "AVSI intervention" we can see that with respect to the model not quantified the situation is not changed much: the latent blocks "Family support" and "Support for the school" have the major impact (this is also because these kinds of help are more frequent than the other) on the "AVSI intervention" respect to the "Nutritional and health support" . The relationship between the latent concepts, in the model quantified, are all significative, as also the relation between Status of child and Guardian satisfaction, even if the impact of the first latent is weak (path coefficient is 0.0966). Table 3 reports the values of the normalized Gof index, average communality and average redundancy for both models. The important difference is in the comparison between the two normalized Gof index: the model quantified shows a better fitting to the data (the normalized Gof is equal to 0.40) even if it's inferior to the mean value, with respect to the other model that presented a lower Gof index (0.38). Respect to the average communality the model not quantified has a slightly higher mean value superior with respect to the other model (0.3986 versus 0.3961); also for the average redundancy the quantified model has a better performance (0.113 versus 0.097). From the results we can conclude that by performing an a priori quantification of qualitative indicators it is possible to obtain some improvement in the results of the model, both about the outer model (minor lost of information) and about the inner model (we have better estimation of the causal relationship between latent concepts). About the application we must remember that the questionnaire was not built to be treated with Structural Equation Models and the database had a big number of missing values that affected the stability of the model.

Table 1 Inner model for the model not quantified

Block	Factor	Corr.	Contr. R2	P. Coeff	L. Conf	U. Conf	St. dev	Student't
AVSI intervention	R2	0.999			0.997	1.000		
	Intercept			0.0000				
	Support for school	0.815	36.15	0.443	0.419	0.465	0.011	40.69
	Nutritional support	0.78	27.86	0.36	0.33	0.39	0.013	26.80
	Support for family	0.805	35.99	0.45	0.413	0.48	0.014	31.113
Status of child	R2	0.08			0.05	0.115		
	Intercept			0.0000				
	AVSI intervention	0.15	20.83	0.11	0.03	0.15	0.03	4.14
	Housing condition	0.165	26.53	0.13	0.086	0.17	0.02	6.36
	Family environment	0.22	52.65	0.19	0.15	0.25	0.024	7.78
Guardian satisfcation	R2	0.006			0.0000	0.001		
	Intercept			0.0000				
	Status of child	0.02	100.0000	0.02	-0.043	0.09	0.03	0.77

Table 2 Inner model for the model quantified

Block	Factor	Corr.	Contr. R2	P. Coeff	L. Conf.	U. Conf.	St. dev	Student't
AVSI intervention	R2	0.99			0.99	1.000		
	Intercept			0.00				
	Support for the school	0.82	35.77	0.44	0,42	0.46	0.01	41.63
	Nutritional and health support	0.79	28.09	0.36	0.33	0.39	0.01	26.53
	Support for the family	0.81	36.14	0.45	0.41	0.48	0.014	32.51
Status of child	R2	0.18			0.14	0.21		
	Intercept			0.00				
	AVSI intervention	0.11	3.48	0,06	0.001	0.10	0.02	2.42
	Housing condition	0.30	40.29	0.24	0.19	0.29	0.05	5.06
	Family environment	0.34	56.23	0.30	0.26	0.34	0.02	14.16
Guardian satisfcation	R2	0.001			0.0035	0.0243		
	Intercept			0.0000				
	Status of child	0.1	100.0000	0.1	0.056	0.156	0.026	3.60

Table 3 Comparison between the two models

	Model with quantification	Model not quantified
Gof normalized	0.4005	0.3791
Average communality	0.3961	0.3986
Average redundancy	0.1130	0.0973

4 Conclusion

Despite many proposals in literature for the quantification of qualitative variables, the problem does not have a unique solution and researcher are looking for new ways to solve it. The large use of the PLS-PM, we have said, it is due to the possibility to analyze a matrix with the number of variables bigger than of observation, the absence of distributional hyphotesis, the possibility to estimate the relationship in presence of multi-colinearity. Even if the PLS-PM is useful in all these cases, sometimes it fails in presence of a set of when we have the qualitative manifest variables. Future work is oriented to the development of an algorithm that embraces one or more steps of quantification for only qualitative variables:in this way we can obtain the best quantification, taking into account, from an hand, of the variable nature, and, on the other hand, of the purpose and of the method to use.

The final aim is to have an algorithm capable to estimate a model with all kinds of variables, nominal, ordinal and numerical.

References

1. De Leeuw, J.: HOMALS and PRINCALS some generalizations of principal components analysis. In: Diday, E. et al. (eds.) Data Analysis and Informatics, pp. 231–241. North-Holland Publishing Company, Amsterdam (1980)
2. Gifi, A.: Nonlinear multivariate analysis. Department of data Theory. University of Leiden, The Netherlands (1981)
3. Jakobowicz, E., Derquenne, C.: A modified PLS Path Modeling algorithm handling reflective categorical variables and a new model building startegy. Comput. Stat. Data Anal. **51**, 3666–3678 (2007)
4. Joreskog, K.G.: A general method for estimating linear structural equation system. Goldberger and Duncan, pp. 85–112 (1973)
5. Kruskal, J.B., Shepard, R.N.: A nonmetric variety of linear factor analysis. Psycometrika **39**, 123–157 (1974)
6. Lauro, C., Nappo, D.: Multivariate Analysis of OVC survey data and Decision Support tools – OVC Project AVSI second survey (2008)
7. Lohomoller, J.B.: Latent variables Path Modeling with partial least squares. Physica-Verlag, Heildelberg (1989)
8. Lovaglio, P.G.: La stima di Variabili Latenti da variabili osservate miste. Statistica **LXII**, 203–213 (2002)
9. Young, F.W.: Quantitative analysis of qualitative data. Psychometrika **46**, 357–388 (1981)
10. Wold, H.: Soft modeling: the basic design and some extensions. In: Joreskog, K.G. Wold, H. (eds.) Systems Under Indirect Observations. Part II, pp. 1–54. North Holland, Amsterdam (1982)

Monitoring Panel Performance Within and Between Sensory Experiments by Multi-Way Analysis

Rosaria Romano, Jannie S. Vestergaard, Mohsen Kompany-Zareh, and Wender L.P. Bredie

Abstract In sensory analysis a panel of trained assessors evaluates a set of samples according to specific sensory descriptors. The training improves objectivity and reliability of assessments. However, there can be individual differences between assessors left after the training that should be taken into account in the analysis. Monitoring panel performance is then crucial for optimal sensory evaluations. The present work proposes to analyze the panel performance within single sensory evaluations and between consecutive evaluations. The basic idea is to use multi-way models to handle the three-way nature of the sensory data.

1 Introduction and Data Description

The present work investigates panel performance in sensory experiments from a project considering organic milk at the University of Copenhagen. One of the objectives of the project was to establish knowledge about production of high quality organic milk with a composition and flavor different from conventionally produced milk. Specifically, the basic aim was to describe how specific types of pasture and legumes affect the sensory attributes of milk. Two different sensory experiments were conducted consecutively in 2007: the first in spring and the second in autumn. Milk samples from seven different farms representing two different breeds (Holstein-Friesland and Jersey) were analyzed by *sensory descriptive analysis* [11] (12 attributes in the spring experiment and 16 attributes in the autumn experiment). Information on milk production was also provided: (a) pasture; (b) proportion of legumes. The two experiments described in Table 1 presented small differences: in spring only 6 samples were evaluated; the panel in both experiments included 9 assessors but some of them differed from one experiment to another.

Investigating panel performance in sensory experiments is crucial. Even if the assessors are well trained they are human subjects variable over time and among themselves. Thus monitoring their performance is necessary for the efficient use

R. Romano (✉)
Seconda Universitá degli Studi di Napoli, Napoli, Italy
e-mail: romano.rosaria@gmail.com

B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-13312-1_35,

Table 1 Data description

Spring experiment:
Samples:
7 varieties of milk with respect to:
– 2 cow races: Holstein-Fries (HF), Jersey (JE);
– 7 farms: WB, EMC, UGJ, JP, HM, OA, KI.
Panel:
9 assessors, 3 replicates.
Sensory descriptors:
Odor (green), appearance (yellow), flavor (creamy, boiled-milk, sweet, bitter, metallic, sourness, stald-feed) after taste (astringent0, fatness, astringent20).
Measurement scale:
Continuous scale anchored at 0 and 15.
Production variables:
Pasture; proportion of legumes.
Autumn experiment:
Samples:
6 varieties of milk with respect to:
– 2 cow races: Holstein-Fries (HF), Jersey (JE);
– 6 farms: EMC, UGJ, JP, HM, OA, KI (no WB).
Panel:
9 assessors, 3 replicates.
Sensory descriptors:
Odor (green, feed, stald), appearance (yellow, grey), flavor (creamy, boiled-milk, sweet, bitter, metallic, sourness, feed) after taste (astringent0, fatness, feed, astringent20).
Measurement scale:
continuous scale anchored at 0 and 15.
Production variables:
Pasture; proportion of legumes.

of sensory data. As measuring instruments, they must meet the requirements of all measurement methods. Specifically, assessors should use the scale correctly (location and range); score the same product consistently in different replicates (reliability/repeatability); score in agreement, on average, with the panel (validity/consistency); score significant different to different products (sensitivity).

Different approaches to analyze the assessors' performance have been proposed in literature: *univariate* methods mostly based on the ANOVA model [7, 13]; *multivariate* methods aiming to find a *consensus profile* [1, 9, 15]; *multi-way* methods allowing for a simultaneous analysis of samples, attributes and assessors [8].

In the present work, *multi-way models* [16] are used in order to evaluate the panel performance within and between the experiments. The aim is to analysing data from each single experiment and the relation between the different experiments over the time keeping the natural multi-way structure of the data. First, focus is given on each experiment separately: (a) the PARAllel FACtor (PARAFAC) model [3, 10] is used to investigate individual differences between assessors; (b) the N-way Partial Least Squares (N-PLS) model [2] is used to test the predictive ability of the panel. Then, the model from one experiment is tested using data from the other experiment to investigate the performance of the panel as a whole over the time.

2 Methods

2.1 Modeling Assessors' Performance by PARAFAC

PARAFAC is a generalization of PCA to higher order arrays. Let $\underline{\mathbf{X}}$ be the three way array holding the scores x_{ijk} given by K assessors, on I products, according to J attributes. The model can be written as:

$$x_{ijk} = \sum_{f=1}^{F} a_{if} b_{jf} c_{kf} + e_{ijk} \tag{1}$$

where a_{if}, b_{jf}, c_{kf} are the elements of the loading matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$, and F is the number of *components*. The solution to the model can be found by the Alternating Least Squares (ALS), minimizing the sum of squares of the residual e_{ijk}.

Using PARAFAC, variation in the products space and in the assessors space can be modeled at the same time. PARAFAC permits to investigate individual differences in sensitivity, reproducibility, and consistency [6]. Focusing on the assessors dimension, the higher the loadings the higher the sensitivity. There is disagreement if one or more assessors have loadings opposite to the rest of the panel. The variability of each assessor is shown up in the residual analysis.

2.2 Modeling Panel Predictive Ability by N-PLS

N-PLS model is a straightforward extension of the bi-linear PLS regression [12] in case of higher order arrays. Focus here is on the tri-linear PLS, where the explanatory variables are collected in a three way array $\underline{\mathbf{X}}$ (I × J × K) and the dependent variables in a two-way array $\mathbf{Y}$ (I × M). The algorithm aims at decomposing the cube X into a set of triads satisfying a certain criterion. A triad consists of one score vector ($\mathbf{t}$), one weight-vector ($\mathbf{w^j}$) on the second order, and one weight-vector ($\mathbf{w^k}$) on the third order. In case of one dependent variable y, the algorithm finds the vectors ($\mathbf{w^j}$) and ($\mathbf{w^k}$) that satisfies:

$$\max_{w^j w^k} \left[cov(t, y) | min \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} (x_{ijk} - t_i w_j^j w_k^k)^2 \right] \tag{2}$$

From $\underline{\mathbf{X}}$ the weight vectors ($\mathbf{w^j}$) and ($\mathbf{w^k}$) are determined and these, in turn, define the score vector ($\mathbf{t}$) as the least-squares model of $\underline{\mathbf{X}}$. The scores are successively determined to have maximal covariance with the dependent variable. Finally, the scores are related to the dependent variable by regression.

In presence of several dependent variables it is possible to use the algorithm in (2) to model each dependent variable separately. Alternatively, it is possible to model all the variables simultaneously as in the PLS2 algorithm.

Table 2 PARAFAC results

Components	Core consistency	Explained variance	Sum-sq residuals
Spring experiment			
1	100%	21.3%	19153
2	100%	27.9%	17523
3	unstable results	...	...
Autumn experiment			
1	100%	17.1%	14580
2	100%	23.1%	13534
3	unstable results	...	...

N-PLS is used for the prediction of production data [$y1$: pasture, $y2$: proportion of the legumes] from sensory data. The aim is to test the predictive ability of the panel in the two experiments.

2.3 Modeling Panel Performance Between Experiments

The basic aim here is to compare panel performance in the two experiments. As discussed in Sect. 1, panel compositions are not exactly the same. However, it is possible to consider the panel as a whole. Under this assumption, a PARAFAC model with two components is performed on the autumn experiment data and used to predict the spring data. Residuals from this model are then compared with residuals from PARAFAC with two components on the spring data. If residuals from application of PARAFAC on spring data are very low as compared to the residuals from the prediction, then considerable differences between the two evaluations exist.

3 Results

PARAFAC model is performed on the data centered across samples [5]. This preprocessing removes differences between assessors in level. No scaling is applied to any mode. No scaling on the attributes, because they are on the same unit and range. Scaling of the assessors has diverse implications and different scaling methods may be used [14]. However, this is not the focus of the work, so the choice is to not scaling this mode. Analysis of core consistency [4], explained variance and residual analysis presented in Table 2 suggest using a model with two components in both spring and autumn experiments. Models with additional components seem to be quite unstable.

Results from PARAFAC model with two components in the two experiments are shown in Fig. 1. For sake of space only loadings from assessors' mode are presented. In spring, the first component shows a good agreement of the panel but different sensitivities: assessors 8, 1 and 5 are the most sensitive. On the second factor it seems there is no consistency as the assessors are divided into two groups: assessors 7 and 8 are the most sensitive, whereas assessors 2, 6 and 9 are not sensi-

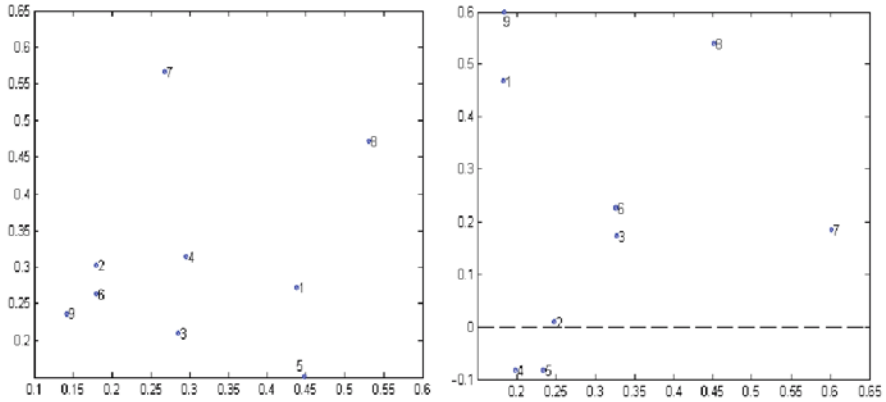


Fig. 1 PARAFAC assessors' loadings in the spring experiment (*left*); PARAFAC assessors' loadings in the autumn experiment (*right*)

tive at all. Similar results in autumn: good panel agreement on the first component with differences in sensitivity and no consistency on the second component. Here there is also disagreement as some assessors have opposite loadings with respect to the panel. Residual analysis provides results (not shown here) on the variability of each assessor with respect to the single attribute and over all the attributes together. A comparative analysis of the results in both experiments, including loadings and residuals as well, shows that the panel in autumn performed better than in spring: in autumn there was a group of good assessors, whereas in spring only a reduced number. The improvement can be due to a better panel performance (training effect) but also to differences in the samples in the two experiments (season effect). However, the first hypothesis seems more realistic. The panel improvement may be due much more to the previous experience as the improvement is related to the agreement of the panel in describing the sensory descriptors rather than on a clearer discrimination of the samples.

N-PLS modeling with a number of factors from 1 to 5 is applied on data from the two experiments. The aim is to make a calibration model for the prediction of quantitative production data [$y1$: pasture, $y2$: proportion of legumes] from the sensory assessment of milk. Figure 2 shows the *root mean squares errors* (RMSE) from calibration (RMSEC) and cross validation (RMSECV), respectively, obtained for the two dependent variables. Segmented cross-validation including all replicates in the same segment is used to avoid optimistic results due the fact that replicates of the samples are used for their own prediction. The calibration suggests a good model with 5 components. However, RMSCV shows that the explained variance is just over-fitting and not real information. Hence, the panel used in the spring experiment provides a not valid model for predicting the production variables.

Results from cross validation in the autumn experiment (not shown here) suggest that in case of $y1$ the explained variance is just over-fitting, while considering $y2$ an improvement in the prediction by introduction of the second and the third factor is obtained. The panel used in the autumn experiment provides then a valid model

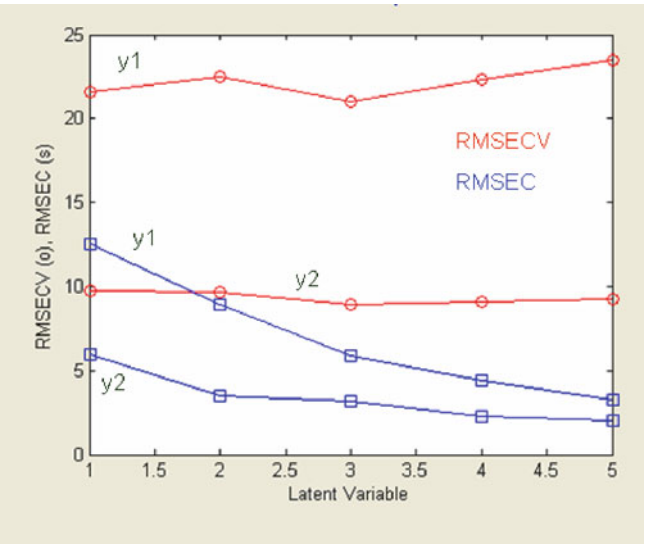


Fig. 2 N-PLS on spring experiment

for predicting the second dependent variable (proportion of the legumes). In Fig. 3 a slight linear trend is observed for $y1$ (first raw) and a clear linear relation for $y2$ (second raw) for any choice of number of components from 1 to 5. Hence, the panel predictive ability in the autumn experiment is better than in spring. However, it must be stressed that more information was provided in the autumn experiment (4 additional attributes).

Residuals from PARAFAC on the autumn experiment for the prediction of the spring experiment and residuals from PARAFAC on the spring data directly are compared. No preprocessing for both PARAFAC models has been selected in order to make the results comparable. There are four more attributes in autumn (compared to spring) and one more sample in spring (compared to autumn), which were

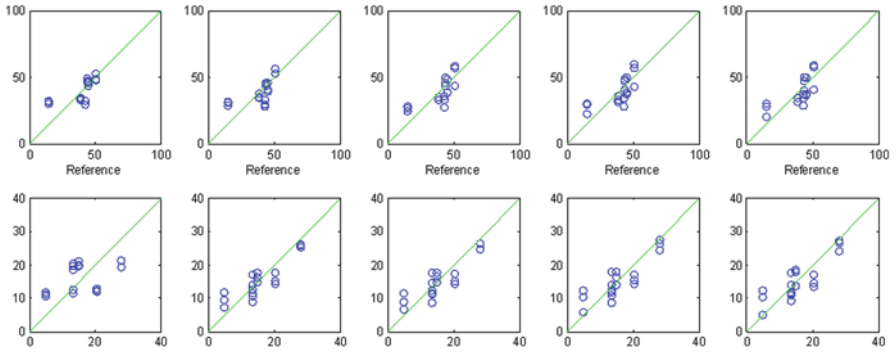


Fig. 3 N-PLS on autumn experiment

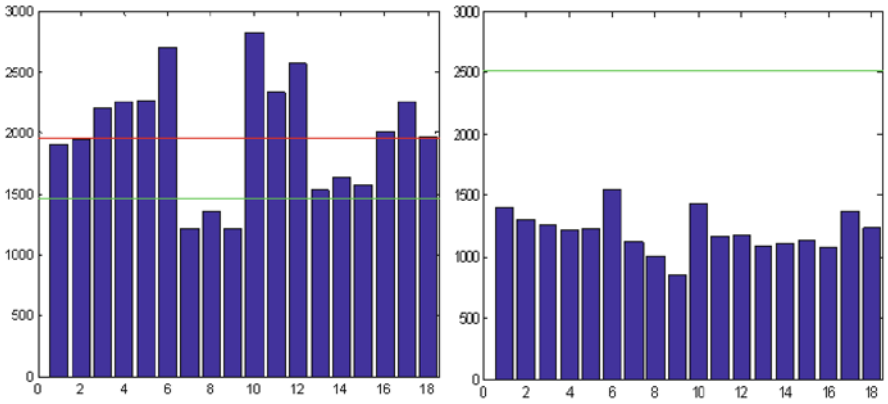


Fig. 4 PARAFAC residuals

eliminated before the prediction. The residuals with respect to the samples mode shown in Fig. 4 are not very different from each other. Of course, the ones from the spring experiment are the lowest but they have similar shape: samples 6, 10, 12, 17 present the highest values in both experiments. This shows the similarity between the autumn and the spring experiment, i.e. the spring experiment has a structure similar to the autumn experiment. There are some differences, but these may be due to the season effect as the two evaluations span from spring to autumn.

4 Conclusions

The aim of the project was to use multi-way models for monitoring panel performance within and between the two experiments. Results from PARAFAC model have shown that the panel in the autumn experiment performed better than the panel in the spring. The improvement can be due to a better panel performance and/or to differences in the samples between the two experiments. Regarding to the panel performance for predicting production data, it was possible to build a valid model on the autumn experiment only, and for one single dependent variable. However, it must be stressed that more information was provided in the autumn experiment. To further investigate this aspect, we have used the PARAFAC model from the autumn data to predict results on spring data. We found out that spring experiment had a structure similar to the autumn experiment. There were some differences, but these may be due to the season effect as the two evaluations span from spring to autumn. Thus, the conclusion is that even if the assessment in spring was much noisy and it was not possible to build a valid model due to a lack of information, it provided a valid sensory evaluation as well.

Acknowledgments This work is supported financially by the *Organic Milk of High Quality* project. The authors would like to thank Professor Rasmus Bro for his intellectual contribution.

References

1. Arnold, G.M., Williams, A.A.: The use of generalised procrustes analysis in sensory analysis. In: Piggott, J.R. (ed.) Statistical Procedures in Food Research. Elsevier, Amsterdam (1986)
2. Bro, R.: Multiway calibration. Multilinear PLS. J. Chemom. **10**, 47–61 (1996)
3. Bro, R.: PARAFAC Tutorial and applications. Chemom. Intell. Lab. Syst. **38**, 149–171 (1997)
4. Bro, R., Kiers H.A.L.: A new efficient method for determining the number of components in PARAFAC models. J. Chemom. **17**, 274–286 (2003)
5. Bro, R., Smilde, A.K.: Centering and scaling in component analysis. J. Chemom. **17**, 16–33 (2003)
6. Bro, R., Quannari, E.M., Kiers, H.A.L., Naes, T., Frost, M.B.: Multi-way models for sensory profiling data. J. Chemom. **21**, 1–10 (2007)
7. Brockhoff, P.B., Skovgaard, I.: Modelling individual differences between assessors in a sensory evaluation. Food Qual. Prefer. **5**, 215–224 (1994)
8. Brockhoff, P.B., Hirst, D., Naes, T.: Analysing individual profiles by three-way factors analysis. In Naes, T., Risvik, E. (eds.) Multivariate Analysis of Data in Sensory Science, pp. 307–342. Elsevier, Amsterdam (1996)
9. Escofier, B., Pagés, J.: Analyses factorielles simples et multiples: objectifs, méthodes et interprétation. Dunod, Paris (1988–1998)
10. Harhman, R.M.: Foundations of the PARAFAC procedure: model and conditions for an 'explanatory' multi-mode factor analysis. UCLA Work. Pap. Phon. **13**, 1–84 (1970)
11. Lawless, H.T., Heymann, H.: Sensory Evaluation of Food. Chapman and Hall, New York, NY (1998)
12. Martens, H., Naes, T.: Multivariate Calibration. Wiley, Chichester (1989)
13. Naes, T.: Handling individual differences between assessors in sensory profiling. Food Qual. Prefer. **2**, 187–199 (1990)
14. Romano, R., Brockhoff, P.B., Hersleth, M., Tomic, O., Naes, T.: Correcting for different use of the scale and the need for further analysis of individual differences in sensory analysis. Food Qual. Prefer. **19**, 197–209 (1990)
15. Schlich, P.: Defining and validating assessor compromises about product distances and attribute correlations. In: Naes, T., Risvik, E. (eds.) Multivariate Analysis of Data in Sensory Science, pp. 259–306. Elsevier, Amsterdam (1996)
16. Smilde, A., Bro, R., Geladi, P.: Multi-Way Analysis. Applications in the Chemical Sciences. Wiley, New York, NY (2004)

A Proposal for Handling Categorical Predictors in PLS Regression Framework

Giorgio Russolillo and Carlo Natale Lauro

Abstract To regress one or more quantitative response variables on a set of predictor variables of different nature, it is necessary to transform non-quantitative predictors in such a way that they can be analyzed together with the other variables measured on an interval scale. Here, a new proposal to cope with this issue in Partial Least Squares (PLS) regression framework is presented. The approach consists in quantifying each non-quantitative predictor according to Hayashi's first quantification method, using the dependent variable (or, in the multivariate case, a linear combination of the response variables) as an external criterion. The PLS weight of each variable which is quantified according to the proposed approach is coherent with the statistical relationship between its original non-quantitative variable and the response variable(s) as expressed in terms of Pearson's correlation ratio. Firstly, the case where one variable depends on a set of both categorical and quantitative variables is discussed; then, a modified PLS algorithm, called PLS-CAP, is proposed to obtain the quantifications of the categorical predictors in the multi-response case. An application on real data is presented in order to enhance the properties of the quantification approach based on the PLS-CAP with respect to the classical approach based on the dummy code of the categorical variables.

1 Introduction

The PLS regression (PLSR) approach [3] has become a common tool in many areas of social and economic sciences. However, PLSR is thought to handle quantitative variables, whereas in these fields researchers are often interested in the investigation of the dependence structure of a set of response variables on a block of predictor variables that are measured at different scale levels (nominal, ordinal or interval). Hence, it arises the need of handling categories so as to make them numerical.

A simple approach to cope with the quantification problem, which can be easily used in whatever regression context, is to replace each non-quantitative predictor

G. Russolillo (✉)
Università degli Studi di Napoli "Federico II", Napoli, Italy
e-mail: giorgio.russolillo@unina.it

B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-13312-1_36,

with the corresponding dummy matrix. This approach, however, does not consider the concept of categorical variable as a unicum, because categories are analyzed as they were distinct variables. Hence, the model assigns a value to the impact of each category, while the researcher is interested in the impact of each explanatory variable on response(s). In order to overcome these problems, a better strategy seems to be the quantification of each category by a numeric value, in such a way that each qualitative variable is transformed in a corresponding quantitative variable to be used into the PLS regression model.

In the OLS framework, MORALS [5] and ACE [1] algorithms are the most largely used in literature to optimize the transformation functions according to the multiple or canonical correlation criterion.

In the following a quantification criterion transform categories into values, in such a way that each qualitative predictor is transformed in a unique novel quantitative variable, is presented in PLS framework.

2 The Univariate Response Case

PLSR predicts a single (PLS1 algorithm) or several (PLS2 algorithm) dependent variables both as a linear combination of a set of predictor variables, and as a linear combination of a set of latent variables $\mathbf{t}_1 \dots \mathbf{t}_a \dots \mathbf{t}_A$. At the same time, it is also a powerful visualization tool, because latent variables compose a lower dimensional subspace in which information on predictor variables, useful to explain the responses, is resumed. It is a very flexible regression tool, able to handle large data sets regardless of the shape of the data matrices, the presence of a (limited) number of missing data and multicollinearity. In this section PLS1 algorithm background will be provided and a criterion to quantify categorical predictors will be presented.

2.1 PLS1 Algorithm Backgrounds

Let $\mathbf{x}_1 \dots \mathbf{x}_j \dots \mathbf{x}_P$ be a set of predictor variables and $\mathbf{y}$ be a response variable measured on N observations. The first PLS component $\mathbf{t}_1$ is built as a linear combination of the $\mathbf{X}$-variables whose weights are the P elements of the vector $\mathbf{w}_1$ ($\mathbf{t}_1 = \mathbf{X}\mathbf{w}_1$). The vector $\mathbf{w}_1$ is computed as $w_1 \propto cov(\mathbf{X}, \mathbf{y})/var(\mathbf{y})$, i.e. as the normalized OLS regression coefficients of each $\mathbf{x}_j$ on $\mathbf{y}$. If variables are centered and normalized to unitary variance, $\mathbf{w}_{1_j} \propto cov(\mathbf{x}_j, \mathbf{y})$.

The second PLS component is computed working on the residuals $\mathbf{y}_1$ of a OLS simple regression of $\mathbf{y}$ on $\mathbf{t}_1$ and the residuals $\mathbf{X}_{1_j}$ of OLS simple regressions of each $\mathbf{x}_j$ on $\mathbf{t}_1$; in formulas:

$$\mathbf{y}_1 = \mathbf{y} - c_1\mathbf{t}_1, \text{ where } c_1 = \mathbf{y}'\mathbf{t}_1/\mathbf{t}'_1\mathbf{t}_1$$
$$\mathbf{x}_{1_j} = \mathbf{x}_j - p_{1_j}\mathbf{t}_1, \text{ where } p_{1_j} = \mathbf{x}'_j\mathbf{t}_1/\mathbf{t}'_1\mathbf{t}_1$$

The second component is defined as $\mathbf{t}_2 = \mathbf{X}_1\mathbf{w}_2$, where $\mathbf{w}_2 \propto cov(\mathbf{X}_1, \mathbf{y}_1)/var(\mathbf{y}_1)$. However, it can be expressed even in terms of the $\mathbf{X}$-matrix as $\mathbf{t}_2 = \mathbf{X}\mathbf{w}_2^*$.

We use the same procedure for computing the following components $\mathbf{t}_a = \mathbf{X}_{a-1}\mathbf{w}_a = \mathbf{X}\mathbf{w}_a^*$.

The search of new components is stopped when the last component does not significantly improve the model predictive capability, in accordance with a cross-validation procedure.

PLSR coefficients are computed as $\mathbf{b}^{PLS(A)} = \sum_{a=1}^{A} c_a \mathbf{w}_a^*$. As a consequence, the regression coefficient of a generic predictor $\mathbf{x}_j$ in a single component PLSR model depends directly on its weight in the construction of the component, and it is proportional to $\rho_{(\mathbf{x}_j,\mathbf{y})}$.

2.2 A Quantification Criterion for Categorical Predictors in Univariate PLS Regression

Let's consider a PLSR model where $\mathbf{y}$ depends on $M + L = P$ variables where $\mathbf{x}_1^{qq} \ldots \mathbf{x}_m^{qq} \ldots \mathbf{x}_M^{qq}$ are quantitative predictors and $\mathbf{x}_1^{ql} \ldots \mathbf{x}_l^{ql} \ldots \mathbf{x}_L^{ql}$ are categorical ones. Without loss of generality, all quantitative variables in the model are assumed centered and normalized to unitary variance.

From a geometrical point of view, each quantitative predictor is a vector in the N-dimensional space defined by the rows of $\mathbf{X}$. Let $\mathbf{G}_l$ of order N x K_l be the indicator matrix of the categorical predictor $\mathbf{x}_l^{ql}$ having K_l categories. Each categorical variable $\mathbf{x}_l^{ql}$, instead, can be geometrically represented as the subspace spanned by the columns of $\mathbf{G}_l$. Any vector in this subspace is a quantification of $\mathbf{x}_l^{ql}$ that respects the constraint for which observations belonging to the same group assume the same value. Since the idea underlying the model is that each independent variable is a predictor of the response variable, it seems coherent to quantify categorical predictors in such a way that each resulting quantified variable is able to explain at best the response. Starting from this idea, the proposed approach is based on a PLS regression model in which each categorical predictor $\mathbf{x}_l^{ql}$ is transformed in a linear combination of the columns of $\mathbf{G}_l$, denoted $\mathbf{x}_l^{qq}$, that maximizes Pearson's correlation coefficient $\rho_{(\mathbf{x}_l^{qq},\mathbf{y})}$. To optimize this criterion, $\mathbf{x}_l^{qq}$ is computed as the projection of $\mathbf{y}$ on the space spanned by the columns of $\mathbf{G}_l$; the resulting vector is then normalized to unitary variance in order to make it homogeneous with the other variables:

$$\mathbf{x}_l^{qq} \propto \mathbf{G}_l \left(\mathbf{G}'_l\mathbf{G}_l\right)^{-1} \mathbf{G}'_l\mathbf{y}.$$

This quantification procedure corresponds to the application of the Hayashi's first quantification method [2] to each categorical variable: Hayashi proposed this quantification criterion in order to predict a quantitative criterion variable (in our case the $\mathbf{y}$ variable) on the basis of the information concerning the categorical attributes of each observation. Each quantified predictor is positively correlated to $\mathbf{y}$; moreover,

the strength of this correlation is measured by the between group deviance of $\mathbf{y}$ given $\mathbf{x}_l^{ql}$ categories, because $bet(\mathbf{y}|\mathbf{x}_l^{ql}) = codev(\mathbf{x}_l^{qq}, \mathbf{y})$. This equivalence is very useful for the interpretation of the parameters $\mathbf{w}_1$ and $\mathbf{b}^{PLS(1)}$ of the model because, since they are a function of $\rho_{(\mathbf{x}_l^{qq},\mathbf{y})}$, they can be expressed (and interpreted) as a function of the correlation ratio $\eta^2_{(\mathbf{x}_l^{ql},\mathbf{y})}$. In particular, it can be easily shown that

$$b_l^{PLS(1)} \propto \eta_{\mathbf{y}|\mathbf{x}_l^{ql}}$$

Hence, the proposed approach for handling categorical predictors assures that, in the PLS framework, the importance that regression algorithm gives to quantified variables is consistent with the capability of the categories of $\mathbf{x}_l^{ql}$ in predicting $\mathbf{y}$. From this point of view, the quantification criterion allows to transfer the predictive properties of original categorical variables to the PLS regression coefficients of their quantifications .

3 The Multivariate Response Case

When there are more response variables $\mathbf{y}_1 \ldots \mathbf{y}_q \ldots \mathbf{y}_Q$, PLS regression decomposes both $\mathbf{X}$ and $\mathbf{Y}$ as a product of a common set of components $\mathbf{t}_1 \ldots \mathbf{t}_a \ldots \mathbf{t}_A$ and a set of specific loadings (respectively $\mathbf{p}$ and $\mathbf{c}$). The weights of the first component are the elements of the normalized vector $\mathbf{w}_1$ maximizing covariance between $\mathbf{t}_1$ and a linear combination $\mathbf{u}_1$ of the response variables. As in the univariate case, the subsequent components are calculated working on $\mathbf{X}_a$ and $\mathbf{Y}_a$, which are respectively the regression residuals of $\mathbf{X}_{a-1}$ and $\mathbf{Y}_{a-1}$ on $\mathbf{t}_a$, under the constraints that $\mathbf{t'}_a\mathbf{t}_{a-1} = 0$ and $\mathbf{w'}_a\mathbf{w}_a = 1$.

3.1 The PLS2 Algorithm

From the computational point of view, the difference between PLS1 and PL2 regression is that in the multivariate case each component is extracted through an iterative algorithm. In each iteration, $\mathbf{Y}$-scores, $\mathbf{X}$-weights, $\mathbf{X}$-scores and $\mathbf{Y}$-weights are sequentially calculated each one as a function of the previous one. In particular, the iterative procedure for the computation of the first PLS component is the following:

Step 0: Select whatever initial vector $\mathbf{u}_1$ (typically the first column of $\mathbf{Y}$)
Step 1: $\mathbf{w}_1 \propto \mathbf{X'u}_1$
Step 2: $\mathbf{t}_1 = \mathbf{Xw}_1$
Step 3: $\mathbf{c}_1 \propto \mathbf{Y't}_1$
Step 4: $\mathbf{u}_1 = \mathbf{Yc}_1/\mathbf{c}_1^{'}\mathbf{c}_1$

Repeat steps 1–4 until the convergence of $\mathbf{u}_1$.

As matter of fact, PLS1 algorithm is a particular case of the PLS2 one where $\mathbf{c}_a$ is a scalar and $\mathbf{u}_a$ is proportional to $\mathbf{y}$.

3.2 *Quantifying the Categorical Predictors in the Multivariate Case: The PLS-CAP Algorithm*

In the PLS2 algorithm, $\mathbf{w}_1$ is calculated as a function of $\mathbf{u}_1$; in particular, each weight w_{1_j} is proportional to $\rho_{(\mathbf{x_j},\mathbf{u_1})}$.

Coherently with this feature of the PLS2 algorithm, the proposal here is to replace each $\mathbf{x}_l^{ql}$ by the (normalized) orthogonal projection of $\mathbf{u}_1$ on the space spanned by the columns of $\mathbf{G}_l$

$$\mathbf{x}_l^{qq} \propto \mathbf{G}_l \left(\mathbf{G'}_l\mathbf{G}_l\right)\mathbf{G'}_l\mathbf{u}_1.$$

As a consequence, the weight of a quantified predictor in the construction of the first component can be interpreted in terms of $\mathbf{u}_1$ correlation ratio square root given the categories of $\mathbf{x}_l^{ql}$. Similarly, the generic single component regression coefficient $b_{lq}^{PLS(1)} = w_{1_l}c_{1_q}$ can be interpreted both as a function of $\rho_{(\mathbf{x_l^{qq}},\mathbf{u_1})}$ and as a function of $\eta_{(\mathbf{u_1}|\mathbf{x_l^{ql}})}$

As a matter of fact, quantified predictors cannot be obtained by a one-step procedure because of the iterative computation of PLS2 parameters: $\mathbf{x}_l^{qq}$ is a function of $\mathbf{u}_1$, but $\mathbf{u}_1$ is, in its turn, a function of $\mathbf{x}_l^{qq}$. In order to overcome this problem, an adjusted PLS algorithm, called PLS-CAP (PLS for CAtegorical Predictors), is proposed.

PLS-CAP computes the first component through an iterative procedure which yields the quantified predictor too. In this iterative procedure (shown in Fig. 1) there are two additional steps, as compared to the classical PLS algorithm. At the first, the quantified predictors are calculated following the previously shown criterion:

$$\text{For each l in (1:L): } \mathbf{x}_l^{qq} \propto \mathbf{G}_l \left(\mathbf{G'}_l\mathbf{G}_l\right)\mathbf{G'}_l\mathbf{u}_1$$

Then, the predictor matrix is updated by juxtaposing the new quantified variables to the quantitative ones:

$$\mathbf{X} = \left[\mathbf{X}^{qt}|\mathbf{X}^{qq}\right]$$

The iterative procedure of the PLS-CAP algorithm calculates at the same time the first component and the quantification for each category. Similarly to the classical PLS2 algorithm, the PLS-CAP can be applied even in the univariate regression case as it is a generalization of the latter.

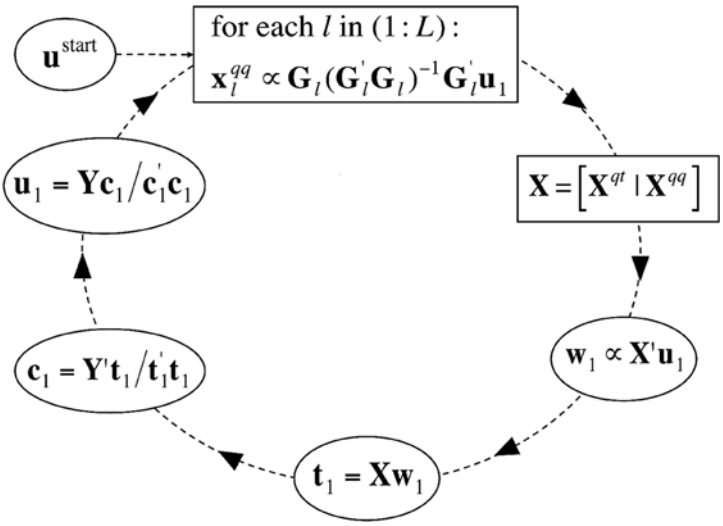


Fig. 1 The PLS-CAP iterative procedure. The steps of the classical PLS algorithm are rounded by an ellypses, while the steps characteristic of PLS-CAP algorithm are rounded by a rectangle

4 An Application to Real Data: The "Cars" Dataset

In the "Cars" dataset a dependent variable (price), six quantitative predictors (displacement, horse power, length, width, weight, speed) and two qualitative predictors (nationality and trimmings), having respectively six and three categories, are measured on eighteen car models. It is a simple dataset, but very useful to pinpoint the differences between the PLS-CAP approach and the classical one, in which each categorical predictor is transformed into an indicator matrix.

The data set was modeled in two ways. In the first, the variable "price" is regressed on six quantitative and nine dummy predictors ("Italy", "Germany", "France", "Japan", "Great Britain", "URSS", "very good trimmings", "good trimmings" and "medium trimmings"). In the second, there are the two (quantified) categorical variables instead of the nine dummy predictors. The two models are named "PLS-Dummy" and "PLS-CAP" respectively. The "PLS-Dummy" regression model assigns a regression coefficient to each quantitative predictor and each category, coded as a dummy variable. On the contrary, the "PLS-CAP" model finds out a regression coefficient for each predictor variable, independently on the nature.

Cross validation procedure led us to keep two components for both the models according to the SIMCA-P 10.0 software rules. Even if PLS-CAP model is built on a lower number of predictor variables, its explicative ability (according to R^2 index) is fully comparable with the explicative ability of the PLS-Dummy model.

Predictive abilities of the two models, according to Q^2 index, were investigated. The Q^2 index is a cross validated R^2 index: it measures the model ability in predicting observations which are not used in the construction of the model. PLS-CAP model showed a higher performance in this sense (Table 1).

Table 1 The comparison between explicative and predictive power of the two models

Model	R^2	Q^2
PLS-dummy	0.93	0.79
PLS-VIP	0.93	0.83

Finally, the ranking of the predictors on the basis of their importance in the prediction was investigated, according to the VIP (Variable Importance in the Prediction) index [4]. The VIP score for the j^{th} variable is calculated as

$$\mathrm{VIP}_{A_j} = \sqrt{\frac{p}{\sum_{a=1}^{A}\sum_{q=1}^{Q} R^2_{(\mathbf{y}_q,\mathbf{t}_a)}} \sum_{a=1}^{A}\left[\sum_{q=1}^{Q} R^2_{(\mathbf{y}_q,\mathbf{t}_a)}\right] w^2_{a_j}}$$

where $R^2_{(\mathbf{y}_q,\mathbf{t}_a)}$ represents the part of variability of $\mathbf{y}_q$ explained by $\mathbf{t}_a$. The VIP index is a useful tool for variable selection in PLSR framework: it provides a measure of the impact of all the dependent variables, and so it can be used both in the univariate and multivariate cases. Moreover, since the average of squared VIP scores equals 1, "greater than one rule" is generally used as a criterion for variable selection.

In Table 2, the predictors ordered by their importance in the prediction (according to VIP index) are shown. In the PLS-CAP model the variable "trimmings" is definitively the most important in the prediction, whereas the PLS-Dummy model does not yield a univocal value for this variable, but just for each of its modalities. Since the categories take first, but also fourth and tenth place in the ranking, it is not trivial to understand the real importance of trimmings in the prediction of the price. Moreover, these results can deceive, because the pseudo-variable "good trimmings"

Table 2 The predictor variables ranked by VIP index in PLS-Dummy and PLS-CAP models

(a) PLS-Dummy model		(b) PLS-CAP model	
Variable	VIP	Variable	VIP
Trimmings (VG)	1.62	Trimmings	1.47
Horse power	1.43	Horse power	1.07
Weight	1.35	Weight	1.01
Trimmings (G)	1.27	Length	0.94
Length	1.20	Displacement	0.91
Displacement	1.17	Width	0.89
Width	1.08	Speed	0.83
Speed	1.07	Nationality	0.71
Nationality (U)	0.85		
Trimmings (M)	0.59		
Nationality (D)	0.57		
Nationality (F)	0.40		
Nationality (I)	0.37		
Nationality (GB)	0.19		
Nationality (J)	0.17		

seems to be a bad predictor of the price, at the opposite of "very good trimmings" and "medium trimmings"; what really happens is that trimmings as a whole is a very good predictor of price, because cars with good trimmings have a medium price, while cars with very good trimmings have a higher price and cars with medium trimmings have a lower price.

5 Conclusions

PLS-CAP algorithm makes PLSR able to work even with predictor variables measured at a different scale level. Relations between quantified predictors and response variable(s) can be read in terms of both correlation ratio and correlation coefficients. The first interpretation assures the coherence of the quantified predictor coefficients with respect to the explicative power of the original categorical variable. The latter assures comparability with the coefficients of the other predictors.

Since PLS-CAP keeps the concept of a categorical variable as a unique entity, the PLSR model gains in terms of interpretability without losing in terms of predictivity.

Acknowledgments The present paper is financially supported by National Interest Research Project (PRIN) 2006 (co-financed by Italian Ministry of University and Research) *Multivariate statistical models for the ex-ante and the ex-post analysis of regulatory impact*. National Coordinator: Prof. Carlo Natale Lauro (University of Naples "Federico II")

References

1. Breiman, L., Friedman, J.: Estimating optimal transformations for multiple regression and correlation. J. Am. Stat. Assoc. **80**, 580–598 (1985)
2. Hayashi, C.: On the prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematicostatistical point of view. Ann. Inst. Statist. Math. **3**, 69–98 (1952)
3. Tenenhaus, M.: La Regression PLS. Technip, Paris (1998)
4. Wold, S., Johansson, E. Cocchi, M.: PLS: partial least squares projections to latent structures. In: Kubinyi, H. (ed.), 3D QSAR in Drug Design: Theory, Methods and Applications, pp. 523–550. ESCOM Science Publishers, Leiden (1993)
5. Young F.W., Jan de Leeuw J., Takane Y.: Regression with qualitative and quantitative variables: an alternating least squares method with optimal scaling features. Psychometrika **41**, 505–529 (1976)

Part VII
Symbolic, Multivalued and Conceptual Data Analysis

On the Use of Archetypes and Interval Coding in Sensory Analysis

Maria Rosaria D'Esposito, Francesco Palumbo, and Giancarlo Ragozini

Abstract Archetypal analysis is a statistical method aiming at synthesizing a set of multivariate observations through few points not necessarily observed. On the other hand, coding data as interval values allows to include variability and variation in the data itself. This work proposes the use of archetypal analysis for interval-coded sensory data to synthesize profiling data taking into account assessor panel variability.

1 Introduction

In sensory analysis, properties or attributes of products are judged by a panel of assessors through the use of their senses. These attributes are mainly exploited as statistical variables in order to profile the products of several brands, to analyze relationships among them, and to explain individual differences on the basis of other individual features [9]. The sensory data collection process usually yields a three-way data table where a panel of assessors evaluate n different brands or products with respect to a set of p attributes. When the focus is on the products, the within panelist variability may be a relevant issue. Even if the judges are usually trained in the use of the measurement scale and in the evaluation techniques, individual judgments may be biased by subjective impressions. In order to overcome this problem, panel average scores for each attribute are tipically considered prior to further analysis on product descriptions [9]. Collapsing the assessor dimension allows us to simplify the data analysis process and to use classical parametric and non parametric multivariate methods. However, it ignores individual differences, and reduces the capabilities for panel monitoring [5].

Alternatively, a wide range of statistical methods that consider the whole set of scores have been proposed [1, 5, 10]. These methods provide information about relationships among assessors, among attributes and among products. They are,

M.R. D'Esposito (✉)
Dipartimento di Scienze Economiche Statistiche, Universitá di Salerno, 84084 Fisciano, Salerno, Italy
e-mail: mdesposito@unisa.it

B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-13312-1_37,

however, quite complex, and hence a variety of plotting tools have been invented to simplify both interpretation and presentation of results [2, 5, 11, 15]. All these methods exploit very simple graphics such as line plots, correlation plots and Manahattan plots or rely on some dimension reduction techniques. As they mainly focus attention on each assessor, on each product or on each attribute in turn, it is difficult to gain an overall view of data characteristics and relationships.

In this paper, we focus on product description. To overcome some of the abovementioned problems – loss of information due to averaging scores and plots that are not fully suited to a simple and thorough representation of the data – we propose a method that combines *archetypal analysis* and *interval-valued data* coding. The archetypes are multivariate objects that make it possible to synthesize data through few representative products. They yield well-separate sensory profiles with which all the other products can be compared and provide information that can be visualized through a set of both simple and more sophisticated graphical representations. On the other hand, the interval-coding allows us to keep part of the panelist variability among assessors without averaging.

The paper is organized as follows: Sections 2 and 3 present the essentials of archetypal analysis and interval coding for sensory data, respectively. Section 4 provides our proposal for extending archetypal analysis to interval-coded sensory data. Section 5 has an illustrative example.

2 Archetypal Analysis

In this section, we recall the essentials of archetypal analysis [4], focusing on its useful characteristics in sensory data profile analysis. Archetypal analysis is a statistical method aiming at synthesizing a set of multivariate observations through few points which are not necessarily observed. These points, the archetypes, can be considered a sort of "pure" types as all the data points must be a combination of them. In addition, to ensure that these *pure* points are as close as possible to the observed data, archetypes must be also a convex combination of the data points.

Let $\mathbf{X}$ be a $n \times p$ data matrix having x_{ij} $(i = 1 \ldots n;\ j = 1 \ldots p)$ as general element. Formally, the archetypes $\mathbf{a}'_k$, $k = 1, \ldots, m$, are those points in the p-dimensional Euclidean space such that

$$\mathbf{a}'_j = \beta'_j \mathbf{X} \tag{1}$$

with

$$\beta_{ji} \geq 0 \qquad \forall j, i \quad \beta'_j \mathbf{1} = 1 \quad \forall j, \tag{2}$$

where $\mathbf{X}$ is the observed data matrix, $\mathbf{a}'_j$ the j th row of the $\mathbf{A}$ matrix, and the convex combination coefficient β_{ji}'s are the n elements of the β'_j vectors, i.e. the weights of the n observations in determining the j th archetype.

At the same time, all the points $\mathbf{x}'_i$ should also be expressed as a mixture of archetpes:

$$\mathbf{x}'_i = \boldsymbol{\alpha}'_i \mathbf{A} \tag{3}$$

with

$$\boldsymbol{\alpha}_{ij} \geq 0 \quad \forall i, j \quad \boldsymbol{\alpha}'_i \mathbf{1} = 1 \qquad \forall i, \tag{4}$$

where $\mathbf{x}'_i, i = 1, \ldots, n$, are the observed data, $\mathbf{A} = (a_{ji})$ is the matrix containing the archetype coordinates, and $\boldsymbol{\alpha}'_i$ is the vector of the convex combination coefficients of the m archetypes for the i-th data point, with generic elements $\boldsymbol{\alpha}_{ij}$, $j = 1, \ldots, m$.

Equations (1), (2), (3) and (4) imply that the archetypes coincide with the vertices of the data convex hull [12]. However, the number of these is usually too large to provide a useful synthesis of the data and, thus, a smaller number m of them has to be chosen. This goal is achieved by modifying Eq. (3) and choosing the m archetypes $(\mathbf{a}'_1, \ldots, \mathbf{a}'_m)$ as those points that:

$$(\mathbf{a}'_1, \ldots, \mathbf{a}'_m; \alpha'_1, \ldots, \alpha'_m) : \sum_{i=1}^{n} \left\| \mathbf{x}'_i - \alpha'_i \mathbf{A} \right\|_2^2 = \text{min!} \tag{5}$$

holding Eq. (1), and the constraints (2) and (4).

The solution to this minimization problem depends on m, and in order to make this choice Cutler and Breiman [4] suggest looking at the quantity:

$$RSS(m) = \sum_{i=1}^{n} \left\| \mathbf{x}'_i - \tilde{\mathbf{x}}'_i(m) \right\|_2^2 \tag{6}$$

where $\tilde{\mathbf{x}}'_i(m) = \boldsymbol{\alpha}'_i(m) \cdot \mathbf{A}(m)$ are the best approximations of the observations $\mathbf{x}'_i$ through the m archetypes. The residual sum of squares $RSS(m)$ is then the sum of the squared Euclidean distances of the observed data from their best approximation, and therefore it measures to what extent the m archetypes synthesize the data.

Archetypal analysis has found application in several fields: physics and astronomy, medicine, performance analysis and benchmarking [2, 8, 12, 14]. In our framework, archetypal analysis will make it possible to define some non-observed products that synthesize all the information and that are characterized by well-separate sensory profiles.

3 Interval Coding in Sensory Analysis

As previously stated, in sensory experiments a group of assessors usually express their judgment on a set of products according to some perception variables. In this framework, we deal with two sources of variability: variability among assessors

and variability among products. Of course, the analysis aims to study the variability among products. However, the scores of a panel of assessors are registered in order to check for possible biased evaluations due to the subjective influences of the assessor. For this reason, variability among assessors can be considered a sort of undesired side effect. In sensory analysis, this issue is usually addressed by taking averaging into consideration, and thus completely loosing all the information regarding panelist variability. In our opinion, therefore, *interval data* coding represents a better solution than simple score averaging as it allows the two variability sources to be kept separate and yet included in the analysis at the same time.

With respect to the classical *single-valued* data, indeed, interval data can capture different sources of *incertitude* in the data such as measurement errors and repeated measures.

The interval coding of the assessor judgment scores can be achieved by considering the range, or the interquartile range, of scores on each attribute. In this way panel variability is included in the analysis. In order to also consider differences among replicates, appropriate interval coding could be envisioned.

The generic $n \times p$ interval data matrix $[\mathbf{X}]$ has row $[\mathbf{x}]'_{\mathbf{i}}$, with general term $[x]_{ij} = [\underline{x}_{ij}, \overline{x}_{ij}]$, $i = 1, \ldots, n$ and $j = 1, \ldots, p$, with $\underline{x}_{ij}$ and $\overline{x}_{ij}$ as the minimum and maximum observed values. From a geometric point of view, the $[\mathbf{x}]'_{\mathbf{i}}$'s are parallelotopes in a p-dimensional space.

The general term $[x]_{ij}$ can be also represented by the *midpoint* x^c_{ij} and *range* (or *radius*) x^r_{ij} notation:

$$[x]_{ij} = [\underline{x}_{ij}, \overline{x}_{ij}] = [x^c_{ij} - x^r_{ij}, x^c_{ij} + x^r_{ij}]. \tag{7}$$

The interval matrix $[\mathbf{X}]$ is split into $\mathbf{X}^c$ and $\mathbf{X}^r$ which are called center and range matrices, respectively.

4 Archetypes for Interval Coded Data

In order to define a generalization of archetypal analysis for interval coded data, recall that archetypes can be obtained as those points providing the best approximation of observed data in terms of closeness. To achieve this best approximation, it is necessary to define a distance function for use in the minimization problem (5).

As regards interval data, from a geometric point of view, a general unidimensional $[x] = \{x^c, x^r\}$ interval represents a *compact* and *closed* subset of the space $\mathbb{R}$. Let us define the unidimensional space of all compact subsets with $\mathbb{IR}$. The distance between two subsets $([x]_i, [x]_{i'})$ in $\mathbb{IR}$ is defined according to the Hausdorff distance as:

$$\begin{aligned} H([x]_i, [x]_{i'}) &= \max\{|\, \overline{x}_i - \overline{x}_{i'} \,|, |\, \underline{x}_i - \underline{x}_{i'} \,|\} = \\ &= |\, x^c_i - x^c_{i'} \,| - |\, x^r_i - x^r_{i'} \,| \end{aligned} \tag{8}$$

Defining the distance in the space $\mathbb{IR}^p$ is a well known but still unresolved issue in the statistical analysis of *interval-valued* data, where $\mathbb{IR}^p$ indicates the p-dimensional space of all compact and closed subsets in $\mathbb{R}^p$. In fact, it is not possible to generalize the distance in (8) to $\mathbb{IR}^p$. In order to obtain a good approximation of the Hausdorff metric in $\mathbb{R}^p$, out of all the proposed approaches we will adopt the one proposed by de Souza and de Carvalho [6].

Similarly to the single value case, some interval archetypes [**A**] which should synthesize the locations and the shapes of all the other data are defined. These archetypes are parallelotopes such that the others can be expressed as a convex combination of them, and they are a convex combination of all the others. We will show that archetypal analysis for interval data can be solved for the two sets of archetypes: $\mathbf{A}^c = (a^c_{ji})$ and $\mathbf{A}^r = (a^r_{ji})$, respectively, in the spaces of centers and ranges. To ensure a unique solution, D'Esposito et al. [7] have proposed imposing the constrain that the mixture coefficients $\boldsymbol{\alpha}'_i$ are the same for the two spaces, representing the algebraic linkage between the two spaces. The parallelotope-archetypes [**A**] are such that they minimize the quantity:

$$HRS(m) = \sum_{i=1}^{n}\sum_{k=1}^{p}\left(\left|x^c_{ik} - \sum_{j=1}^{m}\alpha_{ij}a^c_{jk}\right| + \left|x^r_{ik} - \sum_{j=1}^{m}\alpha_{ij}a^r_{jk}\right|\right), \tag{9}$$

where α_{ij} indicates the weight of j-th archetype on the i th statistical unit. The $HRS(m)$ function in (9) represents the distance among the observed parallelotopes and their representation in terms of archetypes in the Hausdorff metrics generalization in $\mathbb{IR}^p$. As for the single value case, the minimization problem in (9) is subject to the constraints $\alpha_{ij} \geq 0\ \forall i, j, \quad \alpha'_i\mathbf{1} = 1\ \forall i$, given that:

$$\mathbf{a}'^c_j = \beta'^c_j\mathbf{X^c} \qquad \text{and} \qquad \mathbf{a}'^r_j = \beta'^r_j\mathbf{X^r} \tag{10}$$

with

$$\beta^c_{ji} \geq 0 \qquad \beta^r_{ji} \geq 0 \qquad \forall j, i \quad \beta'^c_j\mathbf{1} = 1 \qquad \beta'^r_j\mathbf{1} = 1 \quad \forall j, \tag{11}$$

All computational details are skipped for sake of space. However, it is important to remark that solutions cannot be achieved by the classical alternate least squares. The minimization is achieved by solving a mathematical programming problem [3, 7]. Note that, the solution for the archetypes obtained in terms of midpoints and ranges can be re-expressed in terms of intervals using the correspondence in (7).

5 An Illustrative Example

This section shows some of the main results obtained by exploiting the capabilities of interval data archetypal analysis on a dataset concerning cheese sensory profiles [13]. The data refer to a sample of 14 cheeses and to a double sensory experiment

which involved a panel of 12 assessors using scores on the scale from one to nine. In this example we have kept 13 of the original attributes and we have limited our attention to the first sensory experiment. The 13 variables considered fall into three groups: odour, flavour and fatness. *Acidic* (Acidic), *Intensity* (Int), *Rancid* (Rancid) and *Sun* (Sun) are considered with respect to both *Odour* and *Flavour*, they can be distinguished according to the label suffix: od and fl. *Sweet* (Sweet), *Salty* (Salty), *Bitter* (Bitter), *Metallic* (Met) form the second group and are only flavour variables. *Fatness* (Fat) differs from all the other variables and represents the third singleton group because it can be ascribed neither to the flavour nor to the odour. In this example, we define the interval data for each attribute using the upper and lower scores from the 12 assessors, ending up with 13 interval data for each of the 14 cheeses. On the basis of the $HRS(m)$ values, as m varies, and on the interpretability of the archetypes, we set $m = 3$. The solution is achieved by solving the mathematical optimization problem in (9). Using 100 different random starting points, we verified the stability of the reported solution. This guarantees that the three computed archetypes are not local minima.

Figure 1 represents all 14 cheeses in terms of the related α weights in the Cartesian space $\mathbb{R}^{(3)}$. As a consequence of the linear constraints $\sum_j \alpha_{ij} = 1$ $(\forall i = 1, \ldots, m)$, it has to be noted that the $\boldsymbol{\alpha}_i'$ weights lie in a $(m-1) = 2$ dimensional space, i.e. in a planar triangle. The three vertices of the triangle correspond to the three archetypes: in particular $\mathbf{a}_1'$ coincides with the cheese C13, $\mathbf{a}_2'$ corresponds to the points C7 and C12, and finally $\mathbf{a}_3'$ with the cheese C14. The remaining

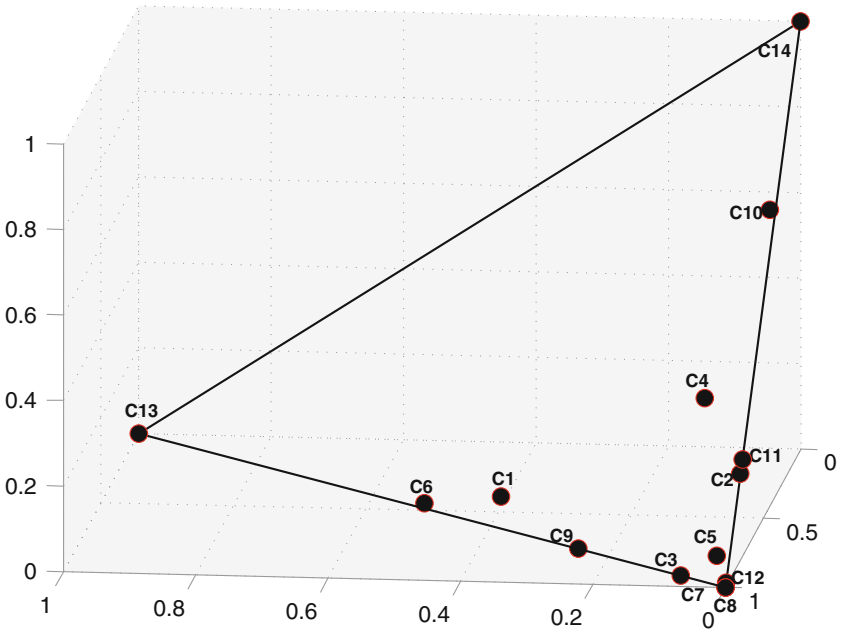


Fig. 1 Scatterplot of the 14 cheeses in $\mathbf{R}^3$: the coordinates of each cheese are the α weights

Table 1 the α weights for the 14 cheeses with three archetypes

Product	α_1	α_2	α_3
C_1	0.389	0.556	0.054
C_2	0.000	0.800	0.200
C_3	0.077	0.923	0.000
C_4	0.076	0.611	0.314
C_5	0.021	0.929	0.050
C_6	0.514	0.476	0.009
C_7	0.000	**1.000**	0.000
C_8	0.000	0.992	0.008
C_9	0.251	0.749	0.000
C10	0.010	0.326	0.665
C11	0.000	0.774	0.226
C12	0.000	**1.000**	0.000
C13	**1.000**	0.000	0.000
C14	0.000	0.000	**1.000**

10 objects can be expressed in terms of linear combination of these three points using the scores in the vectors $\{\beta^c, \beta^r\}$. The weights are also reported in Table 1, where the four rows referring to the archetypes have been highlighted. Figure 1 and Table 1 point out that the majority of the cheeses are clustered around the archetype $\mathbf{a}'_2$, while the other two archetypes are isolated points. Hence, $\mathbf{a}'_2$ sensory profile synthesizes the majority of the cheese profiles, apart from the other two archetypes and the cheese C10 which is closer to $\mathbf{a}'_3$.

The three archetypes expressed as intervals are also graphically displayed as three stars in Fig. 2, where in each star the two polygons refer to the lower and upper bound of the archetype intervals.

Looking at Fig. 2 and Table 2 it is worth noting that the differences among cheeses can be ascribed mainly to acidity, rancidity, saltiness and, to a lesser degree, bitterness. $\mathbf{a}'_1$ and $\mathbf{a}'_3$ are quite similar and they synthesize cheeses with low rancidity. However, $\mathbf{a}'_1$ is also characterized by higher scores on acidity (and also higher ranges), while $\mathbf{a}'_3$ is characterized by higher scores on saltiness and bitterness. On the other hand, the other archetype $\mathbf{a}'_2$ presents much higher scores on rancidity and much lower scores on acidity. Furthermore, the midranges in Table 2 show that rancidity is perceived according to a very subjective scale. Indeed, both the odour

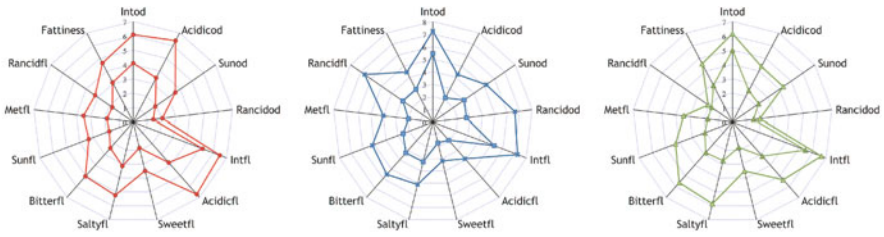


Fig. 2 Archetypes displayed as *stars*: in each star the two polygons refer to the lower and upper bound of the archetype intervals. $\mathbf{a}'_1$, $\mathbf{a}'_2$, $\mathbf{a}'_3$ from *left* to *right*

Table 2 *Midpoints* and *midranges* archetypes components

	Odour (*suffix* "od")				Flawour (*suffix* "fl")								
Var	Int	Acidic	Sun	Rancid	Int	Acidic	Sweet	Salty	Bitter	Sun	Met	Rancid	Fat
Archetypes midpoints													
$\mathbf{a}_1^{c\prime}$	5.086	4.929	2.736	1.740	5.850	5.237	2.662	4.196	3.721	2.558	2.673	2.509	3.859
$\mathbf{a}_2^{c\prime}$	6.345	3.207	4.120	4.664	6.251	2.896	2.432	4.181	4.407	3.861	2.976	4.774	3.696
$\mathbf{a}_3^{c\prime}$	5.584	3.407	3.272	1.709	6.062	4.255	2.648	4.310	4.221	3.166	2.597	1.921	3.721
Archetypes midranges													
$\mathbf{a}_1^{r\prime}$	0.999	1.465	0.856	0.320	0.643	1.477	0.821	1.052	1.318	0.766	0.827	0.730	0.769
$\mathbf{a}_2^{r\prime}$	0.887	1.052	1.081	1.956	1.000	0.992	0.741	0.935	1.149	1.317	1.011	1.863	0.781
$\mathbf{a}_3^{r\prime}$	0.575	0.950	1.050	0.225	0.600	1.100	0.850	1.550	1.400	1.100	0.850	0.150	0.850

and the flavour components are characterized by large intervals between the lower and the upper bound.

This example shows how the analysis carried out by archetypes plus interval-coded data can easily highlight product characteristics.

Acknowledgments The authors are grateful to T. Næs and R. Romano for the cheeses dataset.

References

1. Bro, R., Qannari, E.M., Kiers, H.A.L., Næs, T., Frost, M.B.: Multi-way models for sensory profiling data. J Chemom. **22**, 26–45 (2008)
2. Chan, B.H.P., Mitchell, D.A., Cram, L.E.: Archetypal analysis of galaxy spectra. Mon. Not. R. Astron. Soc. **338**(3), 790–795 (2003)
3. Corsaro, S., Marino, M.: Archetypal Analysis of Interval Data, Reliable Computing, **14**, 105–116 (2010).
4. Cutler, A., Breiman, L.: Archetypal analysis. Technometrics **36**, 338–347 (1994)
5. Dahl, T., Tomic, O., Wold, J.P., Næs, T.: Some new tools for visualizing multi-way sensory data. Food Qual. Prefer. **19**, 103–113 (2008)
6. de Souza, R.M.C.R., de Carvalho, F.A.T.: Clustering of interval data based on city-block distances, Pattern Recognit. Lett. **25**(3), 353–365 (2004)
7. D'Esposito, M.R., Palumbo, F., Ragozini, G.: Archetypal analysis for interval data in marketing research. Ital. J. Appl. Stat. **18**, 343–358 (2006)
8. D'Esposito, M.R., Ragozini, G.: A new R-ordering procedure to rank multivariate performances. Quad. Stat. **10**, 5–21 (2008)
9. Dijksterhuis, G.: Multivariate data analysis in sensory and consumer science: an overview of developments. Trends Food Sci. Technol. **6**, 206–211 (1995)
10. Lea, P., Næs, T., Rodbotten, M.: Analysis of Variance for Sensory Data. Wiley, Chichester, UK (1997)
11. Næs, T.: Detecting individual differences among assessors and differences among replicates in sensory profiling, Food Qual. Prefer. **9**, 107–110 (1998)
12. Porzio, G.C., Ragozini, G. Vistocco, D.: On the use of archetypes as benchmarks. Appl. Stoch. Model. Bus. Ind. **24**, 419–437 (2008)
13. Romano, R., Brockhoff, P.B., Hersleth, M, Tomic, O., Næs, T.: Correcting for different use of the scale and the need for further analysis of individual differences in sensory analysis. Food Qual. Prefer. **19**, 197–209 (2008)

14. Stone, E.: Exploring archetypal dynamics of pattern formation in cellular flames. Physica D. **161**, 163–186 (2002)
15. Tomic, O., Nilsen, A., Martens, M., Næs, T.: Visualization of sensory profiling data for performance monitoring. LWT – Food Sci. Tecnol. **40**, 262–269 (2007)

From Histogram Data to Model Data Analysis

Marina Marino and Simona Signoriello

Abstract The aim of this work is to propose a new approach for dealing with histogram data in symbolic data analysis framework. The idea is to approximate histogram data using B-spline functions in order to synthetize the information within data trough some characteristic function parameters. This parameters will be the new data that could be, subsequently, analyzed with methodologies of multidimensional data analysis.

1 Introduction

The Symbolic Data Analysis techniques that have been developed during the last decade [1], represent new and well adoptable instruments for the complex nature of real phenomena. Different methods [2] are developed for different types of data to analyze, which is the starting point for any statistical analysis.

The simplest symbolic data is the interval data that consider the variation of a phenomenon in term of its bounds or intervals. It means that this kind of data give us little information regarding the interval variability owing to the hypothesis of uniform distribution throughout the interval of phenomenon variation. In order to take a better description of real world into account the use of histogram data could be more appropriate [5, 7, 12]. The techniques developed up to now that analyze the relation among histograms are based on probability density or on cumulative frequencies of the histograms. In this paper, we propose a new way to treat with histograms: all the histograms are transformed in models by means of approximations with functions belonging to the same family in order to synthetize the information within histograms trough some characteristic function parameters. So data are described by a set of parameters and an error term due to approximation. The type of function used to approximate the histograms characterizes the choice of the parameters of the model. Among approximation functions, it is chosen to use B-splines [3].

M. Marino (✉)
Department of Agricultural Engineering and Agronomy, University of Naples Federico II, 80055 Portici, Naples, Italy
e-mail: marina.marino@unina.it

B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-13312-1_38,

This paper is organized as follow: in Sect. 2 we briefly recall the definition of histogram data and a new type of symbolic data, called "Model Data", is proposed; moreover, we recall some basic properties of B-spline. In Sect. 3, we present a way to approximate a histogram using B-spline functions. Finally, in Sect. 4, starting from real data, the histogram transformation process is shown.

2 Histogram Data and Model Data

2.1 Histogram Data

Due to recent developments in data warehousing, a huge amount of continuous data are stored at any occurrence. In these case, aggregation of some kind is necessary even if only to reduce the dataset to a more manageable size for subsequent analysis. There are innumerable ways to aggregate such datasets. In many real experiences, data are collected and/or represented by frequency distributions. If Y is a numerical and continuous variable, many distinct values y_i can be observed. In these cases, the values are usually grouped in a smaller number H of consecutive and disjoint bins I_h (groups, classes, intervals, etc.). The frequency distribution of the variable Y is given considering the number of data values n_h falling in each I_h.

The histogram data, as symbolic data, is described by a partition of an interval into buckets (or sub intervals) weighted by probabilities or relative frequencies. For a generic variable, the i-th histogram data is a model to represent an empirical distribution described as a set of H ordered pairs $Y(i) = (I_h, \pi_h)$ such that:

$$
\begin{aligned}
&I_{hi} \equiv \left[\underline{z}_{hi}, \overline{z}_{hi}\right] \ \underline{z}_{hi} \leq \overline{z}_{hi} \in \Re, \\
&\bigcup\nolimits_{h=1,\ldots,H} I_{hi} = \left[min_{h=1,\ldots,H}\{\underline{z}_{hi}\}, max_{h=1,\ldots,H}\{\overline{z}_{hi}\}\right], \\
&\pi_h \geq 0, \quad \sum\nolimits_{h=1,\ldots,H} \pi_h = 1.
\end{aligned}
$$

2.2 Model Data

In this paper, the term "Model Data" is referred to a set of parameters of the mathematical model used to approximate histogram data.

The data represented by a histogram are transformed into a model that synthesizes the shape of distribution with a certain error, obviously depending on the kind of approximation. We are looking for the best trade-off between model and error. This concerns the choice of the model, or better the choice of the number of parameters to use in the approximation, and the error due to the approximation. Since our data have been suitable processed as a function, they may be summarized through function parameters and some indices of goodness of fit. So for each variable we will get m functions, each of which corresponds to the i th observation. New data have to be proportional in number to the function parameters, in this way any func-

Table 1 Table of the parameters of the new data

	Parameters			
	var 1	var 2	$\cdots$	var s
oss 1	$b_{111}, \ldots, b_{11l}, I_{11}$	$b_{121}, \ldots, b_{12l}, I_{12}$	$\cdots$	$b_{1s1}, \ldots, b_{1sl}, I_{1s}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
oss m	$b_{m11}, \ldots, b_{m1l}, I_{m1}$	$b_{m21}, \ldots, b_{m2l}, I_{m2}$	$\cdots$	$b_{ms1}, \ldots, b_{msl}, I_{ms}$

tion will be replaced by its own parameters and new data will be as many as the $units \times variables \times number\ of\ parameters$. Assuming that all the functions have l parameters $(b_1, \ldots, b_l)$ and an appropriate index (I) of goodness of fit, we can summarize the data as in Table 1.

The problem is now to derive a suitable function that, from a mathematical point of view, is the best approximation of the data. Among approximation functions, we choose to use spline functions [3] because of the simplicity of their construction, their easiness and accuracy of evaluation, and their capacity to approximate complex shapes through rather smooth curve. In particular, we focus our attention on B-splines that are spline functions that has minimal support with respect to a given degree, smoothness, and domain partition.

2.2.1 B-spline

In this section we briefly recall some basic properties of B-spline [3] which are essential in our discussion.

The B-spline functions of degree p compose a basis in the subspace of all the spline functions of degree p. Actually, a spline function of degree p, defined on a knots set $\{t_k\}_{k=0,\ldots,n}$, can be expressed as a linear combination of B-spline functions $B_{i,p}$ on the same knots sequence $\{t_k\}_{k=0,\ldots,n}$:

$$S(t) = \sum_{i=0}^{m} P_i B_{ip}(t), \tag{1}$$

where P_i are $m+1$ control points and the B-spline functions are built in the following way:

$$B_{i,1}(t) = \begin{cases} 1 & t_i \le t \le t_{i+1} \\ 0 & \text{otherwise} \end{cases}$$
$$B_{i,p}(t) = \frac{t - t_i}{t_{i+p-1} - t_i} B_{i,p-1}(t) + \frac{t_{i+p} - t}{t_p - t_{i+1}} B_{i+1,p-1}(t).$$

Therefore, a B-spline curve involve a set of $m+1$ control points, a vector of $n+1$ knots and a degree p, and for them the following expression must holds: $n = m + p + 1$.

Note that, the term B-spline usually refers even, to a spline curve parametrized by spline functions that are expressed as linear combinations of B-splines. In particular, a B-spline function exploits all the properties of a spline functions and take advantage of the following ones: *strong convex hull property* (a B-spline curve is included in the convex domain of its control polygon); *local change property* (changing the position of the control points P_i affects the curve $S(t)$ only in the interval $[t_i, t_{i+p+1})$); *affine invariance property* (if an affine transformation is applied to a B-Spline curve, the result can be achieved by the affine image of its control points).

3 Histogram Approximation by B-spline

We would like to derive a smoother approximation from a histogram to the underlying distribution. We can do this by constructing a spline function $s(t)$ of degree p, pass through the starting and the final points of histogram, whose average value over each bar interval equals the height of that bar. Let $z(i)$ be the left edge of the i-th bar and $h(i)$ its height, we want our spline $s(t)$ to satisfy:

$$\left(\int_{z(i)}^{z(i+1)} s(t)dt\right) / (z(i+1) - z(i)) = h(i).$$

Since our purpose is to be able to compare different histograms, we are going to transform the obtained spline functions in B-spline function to compare their control points.

Fixed $p = 2$, what we get are as many control points as the number of histogram bars. Indeed, since the relation $m = n - p - 1$ holds and having imposed that the curve have to pass through starting and final points, (i.e. the knots sequence is $\{t_0, t_0, t_1, \ldots, t_n - 1, t_n, t_n\}$), we will have $m = n - 1$. In that case, the obtained model perfectly fits the histogram shape and gives an approximation error equal to zero (Fig. 1).

However, our aim is to work with a number of parameters (i.e. control points) less than the number of histogram bars, otherwise we could simply deal with the frequencies of the related histogram classes.

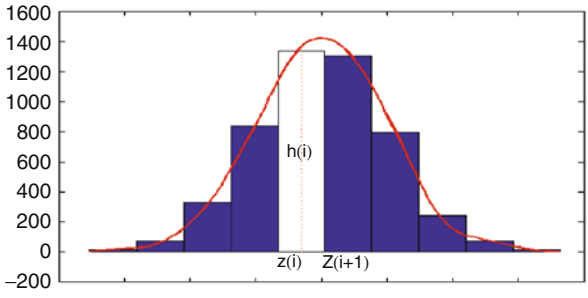


Fig. 1 Smoothing a histogram by quadratic spline

The idea is to get a priori a smaller fixed number of control points equal for all the histograms and thus to obtain different approximation errors. A way to reduce the number of control points is to decrease the number of knots and/or to increase the spline function degree. However, the degree of the B-spline usually do not exceed three, so the number of knots have to be reduced. According to some experimental results, we set the number of knots inside the interval $\left[int\left(\frac{k}{2}\right) - 1, int\left(\frac{k}{2}\right) + 1\right]$, where k is the number of histogram bars, and later work out a suitable index of goodness of fit that allows us to know the approximation quality. Then, the problem is how to find an optimal knots sequence. The location of knots should be established in terms of the best fitting function.

So the starting point is to define a way to compute the approximation error. We can obtain a measure of error as the sum of the squares of the differences between the histogram bar area and the spline function area. In this way we build the objective function to minimize in order to find out the optimal sequence of knots. So we have to solve the following bound-constrained optimization problem:

$$\begin{array}{ll} argmin & \sum_{h=1}^{H} \int_{L_i}^{R_i} [s(t) - h(i)]^2 \, dt \\ s.t. & t_0 \leq t \leq t_n \\ & |t_i - t_{i+1}| > (z_{i+1} - z_i) \, . \end{array} \tag{2}$$

In that way we are going to get a different knots sequence for each histogram and for this reason the B-spline parameters will not be comparable. In order to set the same set of knots for every histogram, a first idea could be to create a knots sequence as the average of the obtained knots. Starting from that sequence we will build final approximating B-spline functions for each histogram. Note that the error introduced replacing the optimal sequence with the sequence of knots average is incorporated in the new approximation error.

It is worth to remark that the control points obtained from the model built on the optimal knots sequence give us information on the histogram form. This form is not referable to a density function and is defined by parameters not statistically interpretable. Subsequently, it can happen that B-spline approximation function, having to pass through the histogram extremities, can assume also negative value.

4 Histogram Transformation Process: An Example

We have considered a dataset representing the sequential "Time Biased Corrected" state climatic division monthly Average Temperatures recorded in the 48 states of US from 1895 to 2007 (Hawaii and Alaska are not present in the dataset).[1]

[1] The original dataset is freely available at the National Climatic Data Center website of US http://www1.ncdc.noaa.gov/pub/data/cirs/drd964x.tmpst.txt

The starting point of transformation process is a single value units×variables matrix where each unit is observed in N occasions. The units are the 48 states, the variable are the months of the year and the occasions are the years.

The steps of the process are summarized as follow:

- Step 1: Histograms building.
  Starting from original data, it is possible to build histograms by pooling occasions. We want to obtain standardized histograms with the same number of bars of same width. Regarding the number of classes, the Sturges formula [10] has been used: $K \simeq 1+log_2 N$ where N is the number of occasions; in our case, being $N = 113$, we set $K = 8$. Moreover, to have a comparison among histograms we transform the histogram in [0, 1] by means:

$$y = \frac{x - min(x)}{max(x) - min(x)} \tag{3}$$

  where x is the occasions vector. In this way we have built histograms with the same bins $\{z_1, \ldots, z_{K+1}\}$. At the end, a new matrix is obtained where in each column there is an histogram variable (see histograms in Fig. 2)
- Step 2: Choose the optimal knots sequence.
  As said before we want to construct B-splines on a number of knots within $\left[int\left(\frac{k}{2}\right) - 1, int\left(\frac{k}{2}\right) + 1\right]$. For our data we choose to work on five knots, two knots are fixed to the extremities, while the others are chosen by the optimal process (2) inside the interval (0, 1). So, we obtain three significant values for each histogram (the other two are equal for all the histograms).
- Step 3: Average knots sequence computation.
  Our purpose is to compare the control points P_i of (1) and so the $B_i(t)$ must be built on the same knots sequence to get the same bases. In order to get the same knots sequence for each histogram variable, a mean vector of the knots is built. So, we will have a mean knots sequence for each variable.
- Step 4: B-spline building.
  B-splines are constructed starting from the average knots sequence by means of (1). So each histogram is approximate by means of a B-spline (Fig. 2).
- Step 5: Calculation of parameters matrix.

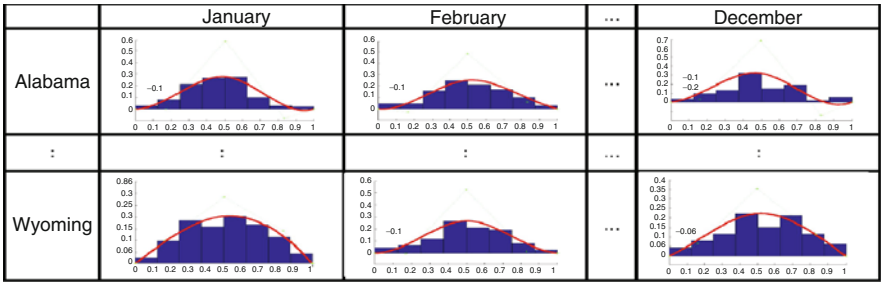


Fig. 2 Matrix of histogram data approximated by B-spline

Table 2 Table of the parameters of real data

		January	February	⋯	December
Alabama	c.p.	$-0.001,\ 0.591,\ -0.081$	$-0.028,\ 0.482,\ 0.056$	⋯	$0.029,\ 0.689,\ -0.139$
	err.	$9.48E-5$	$6.01E-5$	⋯	$5.76E-4$
	loc.	46.80	47.95	⋯	47.25
	size	27.00	20.70	⋯	18.30
⋮	⋮	⋮	⋮	⋮	⋮
Wyoming	c.p.	$-0.124,\ 0.569,\ 0.146,$	$-0.003,\ 0.287,\ 0.211$	⋯	$0.050,\ 0.279,\ 0.101$
	err.	$2.99E-4$	$8,72E-5$	⋯	$5.51E-4$
	loc.	16.65	21.50	⋯	19.95
	size	25.50	21.80	⋯	21.50

We will build the matrix containing the control points that are the parameters of the approximation function which will be used during subsequently analysis. Moreover, we would like to keep information about goodness of fit. A measure of this can be the minimum of the objective function in (2).
Note that, we worked on histogram data that were normalized in the interval [0, 1] and we approximated them by means of B-spline functions whose control points will be calculated to get information about the histogram shape. Doing the transformation on [0, 1] the histogram shape and the spline do not change.
Finally, since a histogram is a symbolic data that is characterized by three fundamental measures: *location*, *size* and *shape*, we need to retrieve data about *location* and *size*. The information about the histogram *location* come from $(max(x)+min(x))/2$, while the information about the size come from the width of all the interval that is $max(x) - min(x)$.

To sum it up, we have a matrix of 12 blocks (one for each month) of order 48×6 where 48 is the number of states considered and 6 is the number of parameters (Table 2). The first three parameters are the B-spline control points and give us information about the *shape*, the fourth parameter is the error term, the fifth and the sixth are the *location* and the *size* of the histogram.

5 Conclusion and Future Work

In this work we have presented a way to deal with histograms by means of a suitable approximation functions. The aim was to synthetize the information within histograms trough B-spline function parameters (and an approximation error term) to take into account histogram shape. The choice of B-spline function is one of the possible choices. In a different context we could consider approximating histograms through density functions of probability as long as they enter in a family of functions and have the same number of parameters and as long as the latter are comparable. The last case brings a certain inflexibility to the model although contributing with a notable simplicity and interpretation. Our approach, indeed, offers major flexibility

in the choice of model and a possible situation comparability not referable to one single theoretical model. Alternative proposals are the use of moment generating functions or "Lambda Generalized" [11] that can represent a compromise between flexibility and statistical meaning.

Moreover, to consider the same set of B-splines knots for all histograms to compare parameters among histogram we chose to compute the average of the optimal set of knots for each histogram. Of course, more sophisticated choice can be used; in particular, a choice based on global optimization approach have to be investigated.

Finally, the transformation carried out on histograms in this paper, allows us to work in a more simple way on this kind of symbolic data. After the histograms have been transformed, the next stage will be their treatment according to classical techniques of Multidimensional Analysis as factorial and/or classification methods. Among factorial techniques, it is relevant the use of methods suitable for the study of variable described by a block of parameters as the Multiple Factor Analysis [4]. Furthermore, to classify symbolic models the core problem is to define an adequate distance between models [9]; an idea can be the generalization of distance defined in [8], as proposed in [6].

References

1. Billard, L., Diday, E.: Symbolic data analysis: conceptual statistics and data mining. Wiley Series in Computational Statistics, Chichester (2006)
2. Bock, H.-H., Diday, E.: Analysis of symbolic data. Exploratory methods for extracting statistical information from complex data. Springer, Heidelberg (2000)
3. De Boor, C.: A Practical Guide to Splines. Springer, New York, NY (1978)
4. Escofier, B., Pagés, J.: Multiple factor analysis (AFMULT package). Comput. Stat. Data Anal. **18**, 121–140 (1994)
5. Irpino, A., Verde R., Lechevallier Y.: Dynamic clustering of histograms using Wasserstein metric. In: Rizzi, A., Vichi, M. (eds.) COMPSTAT 2006, pp. 869–876. Springer, Berlin (2006)
6. Marino, M., Signoriello, S.: Hierarchical clustering of histogram data using a "Model Data" based approach. Stat. Appl. **20**(1), 49–59 (2008)
7. Rodríguez, O., Diday, E., Winsberg, S.: Generalization of the principal components analysis to histogram data. In: 4th European Conference on Principles and Practice of Knowledge Discovery in Data Bases, Lyon, France (2000)
8. Romano, E., Giordano, G., Lauro, N.C.: An inter-models distance for clustering utility functions. Stat. Appl. **17**(2) (2006)
9. Signoriello, S.: Contributions to symbolic data analysis: a model data approach. Ph.D. Thesis. Department of Mathematics and Statistics, University of Naples Federico II (2008)
10. Sturges, H.A.: The choice of a class interval. J. Am. Stat. Assoc. **21**, 65–66 (1926)
11. Karian, Z.A., Dudewicz, E.J.: Fitting Statistical Distributions: The Generalized Lambda Distribution and Generalized Bootstrap Methods. CRC press, New York, NY (2000)
12. Verde, R., Irpino, A.: A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data. In: Batanjeli, V., Bock, H.H., Ferligoj, A., Ziberna, A. (eds.) Data Science and Classification, pp. 185–192. Springer, Berlin (2006)

Use of Genetic Algorithms When Computing Variance of Interval Data

Jaromír Antoch and Raffaele Miele

Abstract In many areas of science and engineering it is of great interest to compute different statistics under the interval uncertainty. Unfortunately, this task often turns out to be very complex. For example, finding the bounds of the interval that includes all possible values produced by the calculation of quantities like variance or covariance for interval valued dataset is a NP-hard task. In this paper a *genetic algorithm* is proposed to tackle with this problem. An application of the algorithm is presented and compared with the result of an exhaustive search using the same data, which has been performed on a grid computing infrastructure.

1 Introduction

Use of interval analysis in engineering is a topic of great interest because many measurement instruments return interval data. The same type of problems arise in statistics whenever we are measuring unprecisely defined objects as, e.g., length of the disease, height or volume of the tree, etc. One of the main reasons why interval analysis became a hot topic in informatics is rounding and/or discrete representation of continuous processes in computers, etc.

It is easy to accept the idea that interval representation can be richer than the scalar one, however, more difficult to be analysed. Typical reason why interval data are reduced to the scalars lies in the fact that there are routinely not available adequate techniques to treat them in their native form. Unfortunately, transformation of the interval data into the scalars is typically accompanied by the loss of information.

Interval analysis can be roughly described as follows. Let us have two intervals $\mathbf{x} = [\underline{x}, \overline{x}]$ and $\mathbf{y} = [\underline{y}, \overline{y}]$, and some operation **op**. Then interval operation **op** on $\mathbf{x}$ and $\mathbf{y}$ can be defined as

$$\mathbf{op}\,(\mathbf{x}, \mathbf{y}) = \{\mathbf{op}(x, y) \,|\, x \in \mathbf{x},\ y \in \mathbf{y}\}\,. \tag{1}$$

J. Antoch (✉)
Department of Probability and Statistics, Charles University of Prague, CZ-186 75 Prague 8, Czech Republic
e-mail: jaromir.antoch@mff.cuni.cz

B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-13312-1_39,

In other words, the result of an interval operation applied to **x** and **y** corresponds to the range of all possible values of **op** if applied to any couple $(x,\ y)$ such that $x \in \mathbf{x}$ and $y \in \mathbf{y}$.

Interesting examples of interval data can be found in [4, 9]. Another approach, which is described in [6], synthesizes a variable with the set of all possible values that could be obtained from it by supplying real-valued arguments from the respective intervals. For a detailed overview of the application of interval statistics see [5, 10]. For other ways how to treat interval data see [2].

Interval analysis considered in this paper consists in finding extremes of studied operator (operation) when applied to the interval data. Generally, it corresponds to solving following box-constrained optimization problem, i.e.

$$\min, \max \{ f(x_1, x_2, \cdots, x_n) \mid x_i \in [\, a_i, b_i \,],\ i = 1, \ldots, n \}, \tag{2}$$

where the objective function f is expression of the statistic of interest and x_i's represent possible realizations of the (interval) observations that, as a whole, define a hypercube $\mathcal{K} = \prod_{i=1}^{n} [\, a_i, b_i \,]$. For some functions there exist analytical solution, e.g. for the sample mean. In this case the bounds of the interval mean are the mean of the inferior endpoints and the mean of the superior endpoints. For details see [5]. It can be shown that for monotone functions it is often possible to get the interval in an explicit form. However, more frequent are the situations in which this is not possible. Moreover, finding the endpoints of the solution set becomes a complicated problem in which specialized algorithms are to be applied when searching for solutions together with the exploitation of specific properties of the objective function in an effort to reduce the amount of computation to be performed.

Figure 1 shows examples of different types of interval data. As pointed out in [5], it appears that for different types of intervals different algorithms are efficient. All the datasets that do not belong into the first nine categories are qualified as *general*. For such data sets it is typical that finding an exact solutions in an affordable amount of time is practically impossible, because the only way how to proceed is exhaustive search over the multidimensional cube defined by all intervals.

2 Specific Problem

Let us consider n intervals $I_i = [a_i, b_i],\ a_i \le b_i,\ i = 1, \ldots, n$, and denote by $\mathcal{K}$ their cartesian product, i.e. $\mathcal{K} = I_1 \otimes \ldots \otimes I_n = [a_1, b_1] \otimes \ldots \otimes [a_n, b_n] \subset \mathbb{R}^n$, where $\mathbb{R}^n$ denotes n-dimensional Euclidean space. Our task is to find among all the vectors falling into $\mathcal{K}$ that one which has maximal variance, i.e. to find

$$\mathbf{x}^{max} = \arg\max_{\mathbf{x} \in \mathcal{K}} \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \overline{x}_n \right)^2, \tag{3}$$

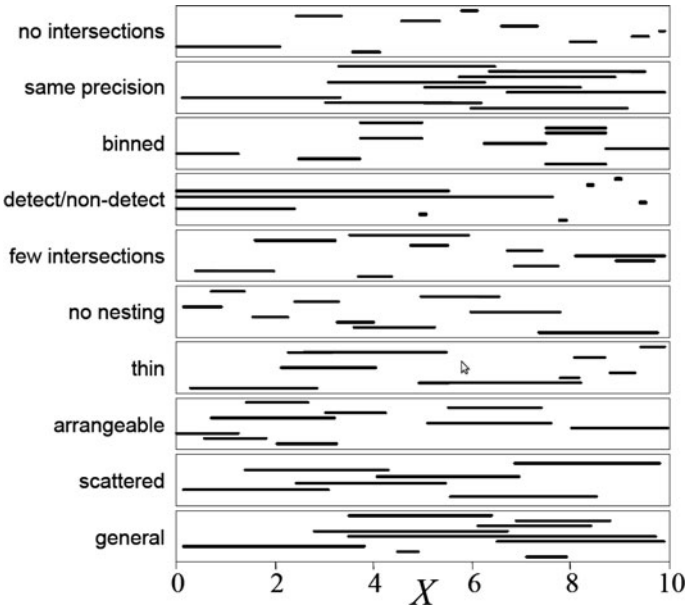


Fig. 1 Different types of interval dataset. Source: [5]

where $\overline{x}_n = n^{-1} \sum_{i=1}^{n} x_i$. Symbol $\arg\max_{\mathbf{x}\in\mathcal{K}}$ denotes, as is usual, argument of the maxima, i.e. that value $\mathbf{x} \in \mathcal{K}$ for which the maximum is attained. It is worth of noticing that solution(s) of (3) coincide(s) with the solution of the problem to find

$$\arg\max_{\mathbf{x}\in\mathcal{K}} \frac{1}{n-1} \sum_{i=1}^{n} \left| x_i - \overline{x}_n \right|. \tag{4}$$

It can be easily shown that solution of our task is not necessarily unique. Moreover, it is clear that the normalization either by $1/n$ or $1/(n-1)$ etc. does not play the role on the results. Following assertion, which will be of key importance for our genetic algorithm, has been proven in [1].

Assertion 1 *Assume the above mentioned setup. Then solution(s) of* (3) *coincide with one (or more) vertex(es) of* $\mathcal{K}$. □

Remark 1 It is important to notice that (3) is equivalent to finding

$$\arg\max_{\mathbf{x}\in\mathcal{K}} \frac{1}{n(n-1)} \sum_{1\le i<j\le n} \left(x_i - x_j\right)^2. \tag{5}$$

From the geometric point of view this means that we are looking for that corner of $\mathcal{K}$, which has the largest distance from the straight line passing through the origin and the point $(1, \ldots, 1)'$. Formal proof can be found in [1].

3 Genetic Algorithm

3.1 General Remarks

Genetic algorithms (GA) are stochastical procedures that provide a random-search based alternative to the traditional optimization methods using powerful search techniques to locate near optimal (and, sometimes, optimal) solutions in complex optimization problems. They can be briefly described as stochastic algorithms whose search methods mimic natural phenomena based on genetic inheritance and selection. GA perform multidirectional search by maintaining a population of potential solutions and assuring knowledge formation and exchange between different directions, see [8] for details. Potential solutions of the problem evolve. More precisely, at each generation better solutions reproduce, while relatively bad solutions eventually die off. GA's have been successfully applied to many real world optimization problems like scheduling processes, travelling salesman problem, etc. For more details see, e.g., [7].

To be able to use any genetic algorithm, it is necessary to define:

- Genetic representation of the data and parameters of the problem.
- Evaluation (fitness) function.
- Genetic operators (crossover, mutation) altering the population.
- Values of the parameters used by the algorithm (population size, number of generations, probabilities to apply genetic operators, selective pressure, etc.).

A genetic algorithm is described by the following scheme:

procedure genetic algorithm
begin
  choose a coding to represent variables
  $t \leftarrow 0$
  initialize population $P(t)$
  evaluate population $P(t)$
  while (**not** terminating condition) **do**
    $t \leftarrow t + 1$
    select $P(t)$ from $P(t-1)$
    modify $P(t)$ using crossover and mutation
    evaluate $P(t)$
  end
end

3.2 Specification for Our Specific Problem

For most applications, in particular when the objective function and the search domain are not too complex, the most critical point is the one related to the encoding of respective objects, while the other can be usually easily accomplished or chosen by empirical evidence.

Representation of variance in the form (5) is the key point allowing us to concentrate only on the vertexes of the cube $\mathcal{K}$. It is evident that there exist a 1–1 mapping between the set of all vertexes of $\mathcal{K}$ and a set of vectors $\alpha = (\alpha_1, \ldots, \alpha_n)' \in \{0, 1\}^n$, where $\alpha_i = 1$ corresponds to the choice $x_i = a_i$ while $\alpha_i = 0$ corresponds to the choice $x_i = b_i$. Thanks to the fact that the assignment of chromosomes and vertexes of $\mathcal{K}$ is natural and the fitness function is given by the variance calculated for given vector $\mathbf{x}$, it is enough to set the population size S, crossover k, mutation probability p_M and stopping rule.

Summarizing, above described representation allows us to attack the problem with a genetic algorithm in which the fitness function is the variance we want to maximize, and a candidate solution is modeled as a combination of boundary points of the intervals whose coding is (natively) binary. The other parameters of the genetic algorithm have been, for our example, chosen empirically as follows:

- Initial generation has been chosen randomly, i.e., the genes were simulated from the alternative distribution $Alt(1/2)$.
- Crossover scheme : single point crossover with $k \approx 0.6n$.
- Mutation probability $p_M \approx 0.01$.
- Fitness $f(\alpha) = var\,\mathbf{x}$, where $\mathbf{x}$ is that vertex of $\mathcal{K}$ that corresponds to the chromosome α.
- Population size $card(S) = 100$.
- Number of generations 300.
- Elitism was used, i.e., the best individual of a generation is cloned with the new one.

Finally, take a look on the sensitivity of the procedure when changing the parameters. Basic conclusion is that the mutation probability considerably influences both the population size $card(S)$ and number of generations. Other parameters do not play so important role. More specifically:

- If we increase mutation probability, we must either considerably increase the number of generations or population size. For example, the choice $p_M = 0.025$ recommended by the literature required either to double the population size or to triple the number of generations.
- Choice of the initial generation does not have substantial impact on the speed to arrive to the optimal solution.
- Crossover scheme does not have an impact on the speed to arrive to the optimal solution. Single point crossover gave us practically the same results as two point or random crossover.

4 Example

The algorithm has been tested on many real and simulated datasets. General conclusion is that the convergence for suitably tuned parameters has been very fast in all considered situations. Moreover, it seems to scale well with the problem dimensionality.

The quality of the solution(s) found by the genetic algorithm has been checked on a Grid Computing Environment of the University of Naples Federico II, which has been used for an exhaustive search leading to the optimal solution. In all the simulations that have been performed the genetic algorithm found the global maximum in less than one thousand of iterations. As an example we have chosen a *general type* dataset consisting of 40 interval observations. The data are reported in Table 1.

According to [5] there does not exists for this type of the data other algorithms enabling to find $\mathbf{x}^{max}$ than exhaustive search. It took us about 4 h on a cluster with 16 multi-kernel processors to reveal that $\mathbf{x}^{max}$ correspond to the point given in Table 2. Notice that we have found the same point using our genetic algorithm in less than several hundred iterations, taking less than a second of CPU on one of the processors.

Table 1 Data

a_i	b_i	a_i	b_i	a_i	b_i	a_i	b_i
−47.50	28.75	40.75	95.00	−10.00	7.75	27.75	57.00
47.25	91.75	−47.00	30.50	−81.50	−72.25	49.75	85.50
38.50	81.50	−95.75	−25.25	−94.00	−18.75	12.75	90.75
−53.50	45.00	−34.00	−28.25	−65.50	22.00	18.50	85.50
−46.50	93.50	−1.25	34.50	−3.25	83.25	−93.00	−83.00
−98.25	−40.75	16.50	81.25	−90.25	−77.00	−95.75	−52.00
−34.50	−28.00	−21.75	39.75	−24.75	−1.25	−66.00	−61.25
51.00	64.00	30.25	80.25	42.50	79.75	−77.50	−42.75
−88.50	−71.50	−14.50	−6.00	−11.00	−5.00	−32.50	−13.50
−93.75	−33.00	−22.50	1.25	−19.50	6.75	−42.25	20.00

Table 2 Solution

−47.50	91.75	81.50	45.00	93.50	−98.25	−34.50	64.00
95.00	−47.00	−95.75	−34.00	34.50	81.25	39.75	80.25
7.75	−81.50	−94.00	−65.50	83.25	−90.25	−24.75	79.75
57.00	85.50	90.75	85.50	−93.00	−95.75	−66.00	−77.50
−88.50	−93.75	−14.50	−22.50	−11.00	−19.50	−32.50	−42.25

5 Conclusions and Perspectives

As has been shown in [3] e.g., finding the upper bound of variance for interval data is a NP-Hard problem. Even if in some cases it is possible relatively quickly find it through exploiting the characteristics of some particular interval variables, a

computationally affordable algorithm for all kind of datasets does not exist. In this paper an heuristics based on a genetic algorithm has been presented that founds the optimum in all the real and simulated data we have been able to analyse with a grid computing infrastructure. Of course, the computational complexity of the problem does not allow the grid to perform exhaustive search when the number of statistical units grows above a certain threshold, however, the behavior of the algorithm seems to be robust with respect to the dimension of the dataset (the number of iterations required to converge does not explode when the number of units increases). Future directions of research, some of which are already in progress, are leading towards:

- Fine tuning of the algorithms.
- Calculations of bounds for more complex quantities like covariance, correlation and regression coefficients.
- Study of the convergence speed of the algorithms.

Acknowledgments Work of the first author paper was supported by grants GAČR 201/09/0755 and MSM 0021620839. Work of the second author was supported by the project S.Co.P.E., Programma Operativo Nazionale 2000/2006, Ricerca Scientifica, Sviluppo Tecnologico, Alta Formazione.

References

1. Antoch, J., Brzezina, M., Miele R.: Analysis of symbolic data. Comput. Stat., first on line **25**(1), 143–153 (2010)
2. Boch, H.H., Diday, E.: Analysis of Symbolic Data. Springer, Berlin (2000)
3. Ferson, S., Ginzburg L., Kreinovich V., Longpré, L., Aviles, M.: Computing variance for interval data is NP-hard. ACM SIGACT News **33,** 108–118 (2002)
4. Ferson, S., Joslyn, C.A., Helton J.C., Oberkampf, W.L., Sentz K.: Summary from the epistemic uncertainty workshop: consensus amid diversity. Reliab. Eng. Syst. Saf. **85**, 355–370 (2004)
5. Ferson, S., Kreinovich V., Hajagos, J., Oberkampf, W., Ginzburg, L.: Experimental uncertainty estimation and statistics for data having interval uncertainty. http://www.ramas.com/intstats.pdf (2007)
6. Gioia, F., Lauro, C.N.: Basic statistical methods for interval data. Stat. Appl. **17**, 75–104 (2005)
7. Goldberg, D.E.: Genetic algorithms in search, optimization and machine learning. Addison-Wesley, Reading, MA (1989)
8. Michalewicz, Z.: Genetic Algorithms + Data Structures = Evolution Programs. Springer, Berlin (1996)
9. Osegueda, R., Kreinovich, V., Potluri, L., Aló, R.: Non-destructive testing of aerospace structures: granularity and data mining approach. In: Proceedings of FUZZ-IEEE, Honolulu, Hawaii, vol. 1. pp. 685–689 (2002)
10. Xiang, G., Ceberio, M., Kreinovich, V.: Computing population variance and entropy under interval uncertainty: linear-time algorithms. Reliable Comput. **13**, 467–488 (2007)

Spatial Visualization of Conceptual Data

Michel Soto, Bénédicte Le Grand, and Marie-Aude Aufaure

Abstract Numerous data mining methods have been designed to help extract relevant and significant information from large datasets. Computing concept lattices allows clustering data according to their common features and making all relationships between them explicit. However, the size of such lattices increases exponentially with the volume of data and its number of dimensions. This paper proposes to use spatial (pixel-oriented) and tree-based visualizations of these conceptual structures in order to optimally exploit their expressivity.

1 Introduction

Information retrieval and navigation have become very difficult in current information systems because of data's volume and lack of structure. Building Galois lattices provides raw data with a structure, through clusters of concepts linked by generalization/specialization relationships. The interest of such concept lattices was studied in previous work [9, 10]. The conceptual navigation layer created by Galois lattices provides users with an additional – structured – abstraction level for their navigation. However, the number of concepts increases exponentially with the size of data and its number of dimensions. In such case, graphical representations such as Hasse diagrams become useless. This article proposes a spatial representation of large Galois lattices through a pixel-oriented and a tree-based visualization.

This paper is organized as follows. Section 2 presents the context of this work, in particular Formal Concept analysis and Galois lattice upon which our methodology relies. Section 3 proposes a pixel-oriented representation of large concept lattices whereas Sect. 4 focuses on a tree-based visualization intended for navigation. Section 5 finally concludes and presents perspectives of this work for the future.

M. Soto (✉)
Laboratoire d'Informatique de Paris 6, 75016 Paris, France,
e-mail: Michel.Soto@lip6.fr

B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-13312-1_40,

2 Context

This methodology is generic and may be applied to any type of data. In this paper its use is illustrated on a dataset called *tourism*, consisting of 126 Web pages about tourism, described by the most significant terms they contain (among 60 possible words such as specific countries).

2.1 Formal Concept Analysis and Galois Lattices

Formal Concept Analysis (FCA) is a mathematical approach to data analysis which provides information with structure. FCA may be used for conceptual clustering as shown in [5, 12].

The notion of Galois lattice to describe a relationship between two sets is the basis of a set of conceptual classification methods. This notion was introduced by [1, 2]. Galois lattices group objects into classes that materialize concepts of the domain under study. Individual objects are discriminated according to the properties they have in common. This algorithm is very powerful as it performs a semantic classification.

2.1.1 Galois Lattices Basic Concepts

Consider two finite sets D (a set of *objects*) and M (the set of these objects' *properties*), and a binary relation R $\subseteq$ DxM between these two sets.

Let *o* be an object of D and *p* a property of M. We have *o*R*p* if the object *o* has the property *p*. According to Wille's terminology [6]:

$$\mathrm{Fc} = (\mathrm{D}, \mathrm{M}, \mathrm{R}) \tag{1}$$

is a formal context which corresponds to a unique Galois lattice, representing natural groupings of D and M elements.

Let P(D) be the powerset of D and P(M) the powerset of M. Each element of the lattice is a couple, also called *concept*, noted (O, A). A concept is composed of two sets O $\in$ P (D) and A $\in$ P(M) which satisfy the two following properties (2):

$$\begin{aligned} &\mathrm{A} = \mathrm{f(O)}, \text{ where } \mathrm{f(O)} = \{\mathrm{a} \in \mathrm{M} | \text{for all o} \in \mathrm{O},\ \mathrm{oRa}\} \\ &\mathrm{O} = \mathrm{f'(A)}, \text{ where } \mathrm{f'(A)} = \{\mathrm{o} \in \mathrm{D} | \text{for all a} \in \mathrm{A}, \mathrm{oRa}\} \end{aligned} \tag{2}$$

O is called the *extent* of the concept and A its *intent*. The extent represents a subset of objects and the intent is made of these objects' common properties.

A partial order on concepts is defined as follows (3):

$$\text{Let C1} = (\mathrm{O1}, \mathrm{A1}) \text{ and C2} = (\mathrm{O2}, \mathrm{A2}), \mathrm{C_1} < \mathrm{C_2} \Leftrightarrow \mathrm{A_2} \subseteq \mathrm{A_1} \Leftrightarrow \mathrm{O_1} \subseteq \mathrm{O_2}. \tag{3}$$

In the context of the *tourism* dataset, each page (url) is an *object* and its *properties* are the most frequent terms it contains. In this case, the Galois lattice then consists of concepts comprising sets of Web pages (*objects*) – in the extents – described by their common terms (common *properties*) – in the intents. For example, pages about *restaurants, gastronomy* in *France* will be gathered in the extent of a concept with these three common properties in the intent.

2.2 Galois Lattices' Interpretation

Galois lattices are very well fitted to showing the various types of connections among the Web pages in the *tourism* dataset (comprised of 126 urls about tourism) as they cluster these pages according to the terms they have in common. However their complexity in size due to the very high number of concepts they contain makes them very difficult to interpret with traditional Hasse diagrams.

The authors of [7] have defined interest measures to reduce the size of large concept lattices and apply their method to healthcare social communities.

The lattice represented on Fig. 1 was computed from the set of 126 Web pages about tourism. The total number of significant terms used to characterize these pages

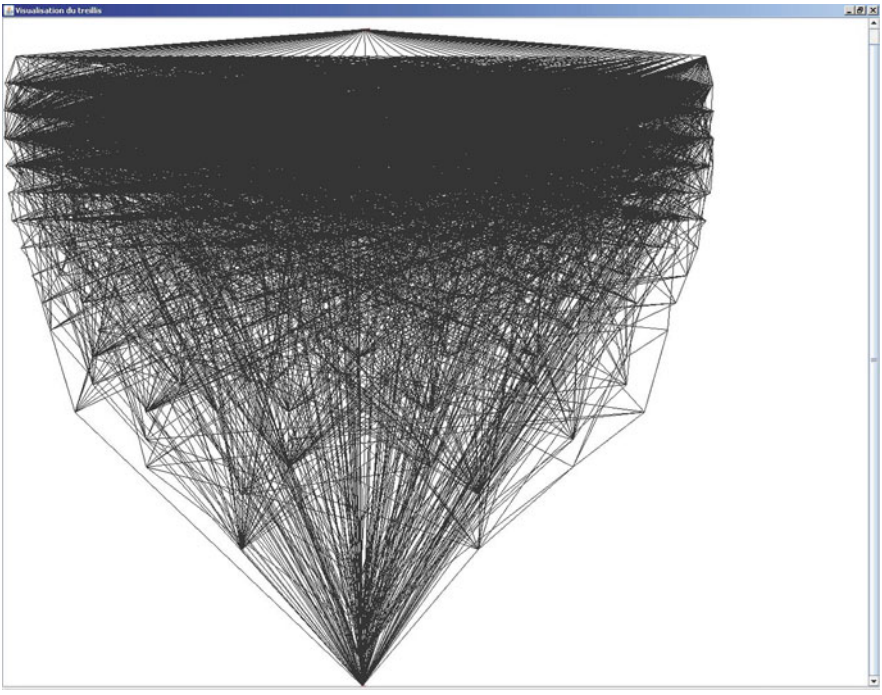


Fig. 1 Large lattice's Hasse diagram

is 60. The lattice contains 2,214 concepts, linked by 7,758 edges; as illustrated by the Fig. 1, its interpretation from the visualization as a Hasse diagram is impossible.

Our approach aims at enhancing Galois lattices interpretation by proposing large lattices visualizations through two different techniques: a pixel-oriented and a tree-based visualization.

2.3 Related Work

The goal of the work presented here is to provide a visual representation of a large Galois lattice, allowing users to interpret it at a glance. We aim at creating a kind of visual *footprint* of the lattice. Among the state of the art, some works generate a map resulting from a query and may represent several hundreds or thousands of documents [4, 11].

In the context of visualization, it seems natural to limit the data space to three dimensions. In order to cope with the dimension issue, three techniques exist: similarity measures (e.g. the vector model), dimension reduction (multidimensional scaling, Principal Components Analysis) and spatial configuration (triangulation, treemaps). Visualization is the most important part as it provides users with an intuitive vision which can be understood and used immediately. Some reduction techniques propose their own visualization; although their interpretation is not always easy, their results may be used as bases for other visualization techniques such as the pixelization [8] presented in the following section.

3 Galois Lattice's Pixel-Oriented Visualization

The methodology used in this section to generate 2D spatial visualizations of Galois lattices is the same as in [3]. The approach consists in representing data as coloured pixels placed in the 2D space along to a Peano–Hilbert curve.

We have applied this method not to the visualization of data itself but of the Galois lattice generated from them. Each concept thus corresponds to a pixel to which a colour must be assigned. Each concept is made of an extent (the list of objects contained in the concept) and of an intent (the list of properties shared by the extension's objects). Each intent's property is a dimension of the concept. A PCA is performed to reduce this number to three (for the Red, Green and Blue colour components). The values of the obtained (X, Y, Z) triples are usually low. In order to get more satisfying values, the inverse Ohta's transform is used as in [3] to approximate the three components of PCA for a natural colour image. This inverse transform is applied to each concept and the values are normalized to be well distributed between 0 and 255. Finally, concepts are ordered according to their RGB vector in order to cluster concepts with similar colours on a straight line – i.e. a 1D space. Each pixel's coordinates are assigned in the 2D space along a Peano–Hilbert

Fig. 2 Pixel-oriented visualization

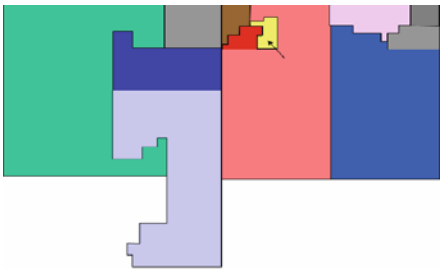


Fig. 3 Clusters identification

curve. This curve places the points of a straight line on a plane by minimizing the Euclidian distance of points which are close on the straight line.

Figure 2 represents the pixelization obtained from the methodology described in Sect. 3 and applied to the Galois lattice of Fig. 1.

From a first visual analysis of this pixelization, some clusters of pixels may be identified as shown on Fig. 3. Each cluster of pixels reflects a cluster of concepts of the lattice. The Euclidian and colorimetric proximity of pixels symbolizes the semantic proximity of the corresponding concepts in the lattice. These clusters show how the lattice's concepts get organized, which was impossible from the Hasse diagram of Fig. 1. This cartography thus allows to consider one or several exploration strategies for the lattice and consequently for the data from which it was computed. In order to check the validity of this interpretation, the portion of the lattice which corresponds to the pixels of the yellow zone (pointed by an arrow on the Fig. 3) was re-built. This zone contains 16 pixels and corresponds to 16 concepts of the Fig. 1.

Given C_Z the set of concepts corresponding to the pixels of the studied zone, we define:

$C_{p\text{-}ext} = \{\text{parent } (c_i)/\text{parent } (c_i) \notin C_z \text{ and } c_i \in C_z\}$ and $C_{f\text{-}ext} = \{\text{child } (c_i)/\text{child } (c_i) \notin C_z \text{ and } c_i \in C_z\}$, with i varying from 1 to card (C_z).

Finally, we define $C'_{p\text{-}ext}$ as the set $C_{p\text{-}ext}$ minus all the parents which have only one child in $C_z$and $C'_{f\text{-}ext}$ the set $C_{f\text{-}ext}$ minus all children which have a single parent in C_z.

The Fig. 4 represents the part of the lattice built from $C'_{p\text{-}ext} \cup C_z \cup C'_{p\text{-}ext}$. On this figure the red nodes belong to C_z and the blue nodes belong to $C'_{p\text{-}ext} \cup C'_{p\text{-}ext}$. The

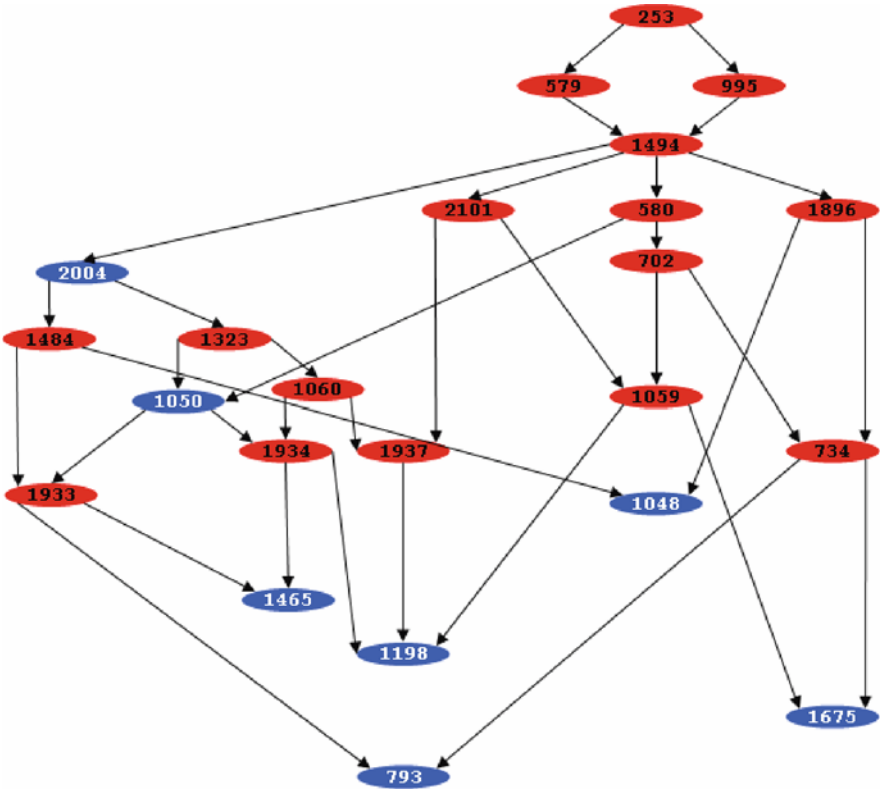


Fig. 4 Portion of Galois Lattice

obtained graph is connex and we may thus conclude that there is a real semantic proximity among C_z's concepts and that this proximity is correctly represented by the cluster of pixels comprised within the studied zone.

This representation provides an overall understanding of the dataset's implicit structure. Figure 3 shows that the *tourism* dataset is structured in six main clusters out of a total of 11. Nevertheless, this representation neither allows to explore the dataset nor to understand the content of the information contained in the dataset. In the next section, we propose a representation providing these two capabilities.

4 Galois Lattice's Tree-Based Visualization

One of the goals stated above is to enhance navigation in large Galois lattices. Exploring concepts in a Hasse diagram is inappropriate when the number of concepts is very high. In this section, we propose to extract a tree of concepts from the Hasse diagram in order to reduce the total number of concepts and provide a hierarchical representation for an intuitive navigation.

4.1 Tree Extraction Algorithm

The tree extracted from the Hasse diagram provides a hierarchical representation of the complex system with different levels of detail (or scales). The result is therefore a hierarchical clustering where clusters may be overlapping (as they are selected concepts from the lattice). The root of the tree contains all objects; the next level groups some objects together in possibly overlapping clusters, the next level is a finer grouping of objects, etc. The number of levels of detail is given by the depth of the tree.

Figure 5 describes the detail of the tree extraction process.

```
The lower bound of the lattice is noted Inf and the upper bound is noted
Sup.
A = allFathers (Sup);
While A ≠{Sup} do
        D = ∅; // set of fathers of A with minimal distance to Sup, where
metric = number of links in the lattice between two nodes)
        A'= ∅; // set of selected parents from A
/** selection of parents of A with minimal distance to Sup **/
        For each concept a∈ A and a≠ Sup do
                F = allFathers (a);
                Da = ∅; // set of fathers of a with minimal distance to
Sup
                For each parent f ∈ F do
                        If distance (f, Sup) = minDistance (F, Sup) then
                                Da = Da ∪{f};
                        End if
                End for
                If card (Da) = 1 then // only 1 parent of a has the mini-
mal distance to Sup
                        Tree = Tree + (a, Da); // adding a and its se-
lected parent in the tree
                        A = A - {a};
                        A' = A' ∪ Da;
                else D = D ∪ Da;
                End if
        End for
// * selecting concepts having an already selected father in A' **/
        For each concept a ∈ A and a ≠ Sup do
                For each parent f ∈ A' do
                        If is Father (f,a) then // f has a father already
selected in A'
                                Tree = Tree + (a,{f}); // adding a and
its father in the tree
                                A = A - {a};
                        Enf if
                End for
        End for

/** selecting random fathers in D **/
        For each concept a ∈ A and a ≠ Sup do
                // randomly select a father of a within D
                f = randomFather (a, allFathers (a), D)
                Tree = Tree + (a, {f}); // adding a and its father in the
tree
                A' = A'∪ {f};
        End for
        A = A';
End While
```

Fig. 5 Tree extraction algorithm

As the clustering algorithm presented here is based on Galois conceptual classification, the generated clusters are conceptually and semantically relevant. This algorithm also exploits the generalisation/specialisation relationship inherent to the Galois lattice.

The construction of the clusters' tree starts from the finest level of detail, i.e. the upper bound of the lattice: the leaf clusters of the tree are thus the most specific concepts of the lattice – i.e. the parent concepts of the upper bound.

For each leaf, one unique parent concept is selected which is a generalisation of the leaf concept. This selection is done according to a hierarchy of criteria in case a concept has several parent concepts in the lattice. These criteria are the following: first, if one of the candidate parents has a lower distance to the upper bound of the lattice (where the distance between two nodes corresponds to the number of links in the lattice between these two nodes), then it is selected as the unique parent. This choice has been made in order to select nodes which bring significant information by filling wider "gaps" between nodes. Moreover, this choice minimizes the number of nodes in the resulting tree.

The following criterion aims at minimizing the overall number of nodes in the final tree: if one of the candidate parents has already been chosen as a unique parent by a sibling node, then it is selected as the unique parent for the current node too. Another criterion is related to potential weights assigned to properties but it has not been used here. Finally if several nodes are still candidate parents, one of them is selected randomly. A unique parent is then selected for each selected concept, and so on until the upper bound of the lattice is reached. At the end of this process, a

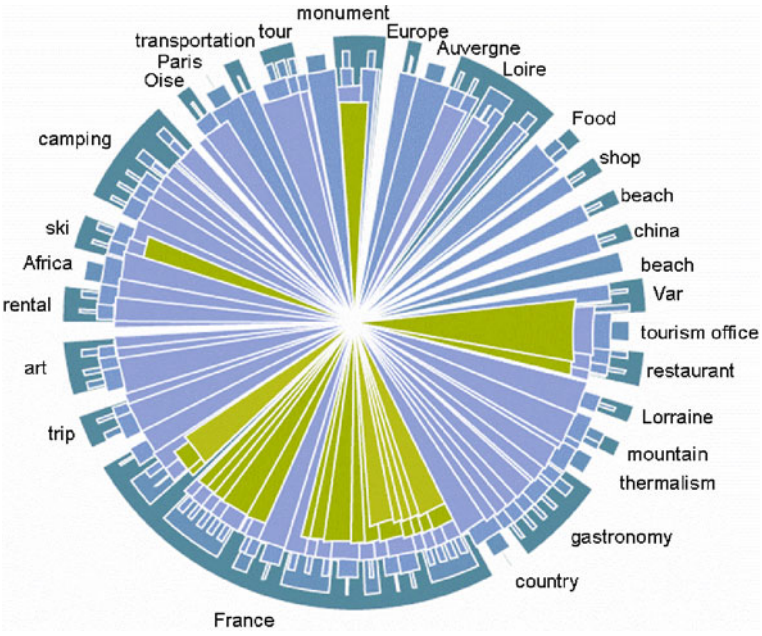


Fig. 6 Clusters visualization

Table 1 Conceptual clustering statistics

- Number of clusters = 272
- Number of concepts in the original lattice = 2,214
- Proportion of eliminated concepts = 87.7%
- Number of levels in the cluster tree = 7
  - Proportion of clusters at level 1 = 0.4%
  - Proportion of clusters at level 2 = 10.7%
  - Proportion of clusters at level 3 = 22.8%
  - Proportion of clusters at level 4 = 30.5%
  - Proportion of clusters at level 5 = 24.6%
  - Proportion of clusters at level 6 = 8.8%
  - Proportion of clusters at level 7 = 2.2%
- Proportion of obvious clusters = 11%
- Proportion of clusters selected because of a minimal distance to the lower bound of the lattice = 18.4%
- Proportion of clusters selected through the sibling relationship = 11%
- Proportion of cluster selected randomly = 59.6%

tree is created. Each level of the tree contains clusters which correspond to a specific level of detail

4.2 Clusters Analysis

Once the tree of clusters is generated, different measures may be computed, e.g. the proportion of concepts of the initial lattice which were not selected to be clusters. An example representation of the tree computed from the *tourism* dataset is shown on Fig. 6 and the corresponding statistics appear in Table 1.

The depth of the tree is interesting because it indicates the number of navigation levels which may be provided to the user. The distribution of clusters at each abstraction level is also studied. If a cluster has no parent, it means that it cannot be generalized. On the other hand, a cluster with no children corresponds to the most specific level.

This representation provides a global understanding of the nature of the information contained in the explored dataset. From the above example we learn that this tourism dataset is mainly related to *France*, *gastronomy*, *camping* and to a region of France named *Loire*. Then users may explore specific clusters by clicking on them.

5 Conclusion

This paper described a methodology to provide a conceptual help for navigation in large and poorly structured datasets, based on the use of Galois lattices. The interpretation of large lattices is impossible with traditional graphical representation.

We proposed a pixel-oriented and a tree-based visualization which are complementary as the former provides an overall understanding of the structuration of the data and the latter provides (a) a semantic understanding of the data thanks to labels associated to the cluster and (b) the capability to navigate within the hierarchy of clusters for information retrieval

In the future, we will conduct experimentations with real end users in order to validate the interpretations obtained with these visualizations and study to what extent they might be automated.

References

1. Barbut, M., Monjardet, B.: Ordre et classification, Algebre et combinatoire, Tome 2. Hachette, Paris (1970)
2. Birkhoff, G.: Lattice Theory, vol. 25, 1st edn. American Mathematical Society, Providence, RI (1940)
3. Blanchard, F., Lucas, L., Herbin, M.: A new pixel-oriented visualization technique through color image. Inf. Vis. **4**(4), 257–265 (2005)
4. Börner, K., Chen, C., Boyak, K.W.: Visualizing knowledge domains. In: Cronin, B. (eds.) Annual review of information science and technology, vol. 37, pp. 179–255. Information Today, Inc., Medford, NJ (2003). Preuss, S., Demchuk, A., Jr., Stuke, M.: Appl. Phys. A **61**
5. Carpineto, C., Romano, G.: Galois: an order-theoretic approach to conceptual clustering. In: Proceeding of the 10th Conference on Machine Learning, Kaufmann, Amherst, MA, pp. 33–40 (1993)
6. Ganter, B., Wille, R.: Formal concept analysis, mathematical foundations. Springer, Berlin (1999)
7. Jay, N., Kohler, F., Napoli, A.: Analysis of social communities with iceberg and stability-based concept lattices. In: ICFCA 2008, LNCS, vol. 4933, pp. 258–272. Springer, Heidelberg (2008)
8. Keim, D.A., Schneidewing, J., Sips, M.: Scalable pixel based visual data exploration. In: Pixelization Paradigm, First Visual Information Expert Workshop, Paris, France, vol. 4370, pp. 12–24. Springer, Berlin (2007)
9. Le Grand, B., Aufaure, M.-A., Soto, M.: Semantic and conceptual context-aware information retrieval. In: The IEEE/ACM International Conference on Signal-Image Technology & Internet-Based Systems (SITIS'2006), pp. 322–332, Hammamet, Tunisie, 17–22 Déc 2006
10. Polaillon, G., Aufaure, M.-A., Le Grand, B., Soto, M.: FCA for contextual semantic navigation and information retrieval in heterogeneous information systems. In: Workshop on Advances in Conceptual Knowledge Engineering, in Conjunction with DEXA 2007, pp. 534–539. IEEE Computer Society, Regensburg, Allemagne, 3–7 Sept 2007
11. Skupin, A., Fabrikant, S.I.: Spatialization methods: a cartographic research agenda for non-geographic information visualization. Cartogr. Geogr. Inf. Sci. **30**(2), 95–115 (2003)
12. Wille, R.: Line diagrams of hierarchical concept systems. Int. Classif. **11**, 77–86 (1984)

Part VIII
Spatial, Temporal, Streaming and Functional Data Analysis

A Test of LBO Firms' Acquisition Rationale: The French Case

R. Abdesselam, S. Cieply and A.L. Le Nadant

Abstract We investigate whether the characteristics of Leveraged Buy-Out (LBO) targets before the deal differ from those of targets that have undergone another type of transfer of shares. Specifically, we examine the size, value, industry, quotation and profitability of French targets involved in transfers of shares between 1996 and 2004. Using two different methods (a classical logit regression and a mixed discriminant analysis), results show that LBO targets are more profitable, that they are more frequently unquoted, and that they more often belong to manufacturing industries in comparison with the targets involved in other types of transfers of shares.

1 Introduction

Leveraged Buy-Outs (LBO) are acquisitions of a significant equity stake of a company by private investors using additional debt financing. Since the evolution of the LBO as a common form of takeover of public or private firms in the 1980s, several companies, hereafter referred to as "LBO firms", specialized in making this type of investment with venture capital raised in the private equity market. This activity in France has experienced an extraordinary increase. From 1997 to 2006, the amounts invested in these transactions increased nine-fold, from 1.259 to 10.164 euro-billion [3].

France is a leader in the LBO market in continental Europe but it is still far behind the United Kingdom and the United States which are the focus of the vast majority of the academic literature (see [6] for a recent overview on LBOs). In this context, we investigate French LBOs in order to provide new evidence on the profile of LBO targets. We test a number of hypotheses derived from LBO firms' acquisition rationale that may explain the French LBO targets' underperformance after the transaction [7, 8, 15, 16]. This analysis allows us to check if LBO firms meet various financial criteria when evaluating an LBO target.

R. Abdesselam (✉)
ERIC EA 3038, University of Lyon 2, 69676 Beron Cedex, France
e-mail: rafik.abdesselam@univ-lyon2.fr

B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-13312-1_41,

2 Theoretical Predictions

To predict the types of targets that are likely to engage in LBOs, we present the specific criteria that are used by LBO firms in their acquisition rationale. LBO firms look for a variety of characteristics in potential investments and are, thus, similar in their basic criteria for takeovers candidates (e.g. mature industries, stable cash flows, low operational risk).

LBO firms generally have two objectives. They seek, first, to maximize their future capital gain from the sale of shares and, second, to minimize the risk of non-payment of the acquisition debt. LBOs create heavy leverage that may be inefficient for firms that expect unstable earnings or plan to engage in new projects. Moreover, heavy leverage may carry with it costs associated with an increased likelihood of insolvency. Since the company's cash flow is used to service the debt, "the most significant risk in an LBO is that the company will not achieve the cash flow necessary to service the large acquisition debt" [18]. Consequently, LBO firms and lenders are most interested in the target's future and past capacity to generate large and steady levels of cash flow. In France, Desbrières and Schatt [7] show that companies undergoing LBOs are the ones which have the greatest ability to remunerate the funds provided by investors and lenders. They find that acquired firms are more profitable than industry average prior to the LBO, which is consistent with the results of Singh [19].

Several characteristics make it possible to define an eligible target for LBO deals. A study by Cressy et al. [5] suggests that LBO firms' skill in investment selection and financial engineering techniques may play a more important role than managerial incentives in raising post LBO performance [11, 12]. A description of financial criteria used by LBO firms to evaluate potential targets follows.

First, one widely accepted conclusion is that the level of financial leverage a firm can bear is a function of its business risk. Business risk is one of the primary determinants of a firm's debt capacity and capital structure [14]. Firms with high degrees of business risk have less capacity to sustain high financial risk, and thus, can use less debt. Firms with risky income streams are less able to assume fixed charges in the form of debt service. Johnson [13] states that firms with more volatile earnings growth may experience more states where cash flows are too low for debt service. For this reason, LBO firms avoid investments in highly cyclical businesses since stability of earnings and cash flow is critical to the success of an LBO. The empirical data developed by Lehn and Poulsen [17] support this view as almost half of their sample of LBOs were in five industries (retailing, textiles, food processing, apparel, and bottled and canned soft drinks) that are all consumer nondurable goods industries for which the income elasticity of demand would be relatively low. Otherwise an LBO target's activity must not require heavy investments. In capital-intensive industries, relatively large amounts of tangible capital assets are required. During the LBO, new investments have to be limited. Moreover, the target expected growth has to be positive but not too high because a high growth rate would create high working capital requirements. The discussion here suggests the following alternative hypothesis. *The likelihood that a target is acquired through an LBO depends on its industry (H_1).*

We expect that LBOs are positively linked with mature and non-cyclical industries and negatively related to the target's industry capital intensity. In particular, we expect that transportation, warehousing and storage (called Transport) is negatively related to LBOs as this industry is cyclical (H1a). On the contrary, wholesale and retail trade industry or hotels and restaurants are indeed cyclical sectors but they are characterized by a low capital intensity. We expect that they are positively linked with LBOs (H1b). We expect that firms in high technology sectors related to LBOs as capital requirements and business risk are high in high-growth firms (H1c). The situation of manufacturing industries is more ambiguous. They are typically very cyclical. But there are important differences among them in how they are affected by a downturn. For instance, the food manufacturing industry is non-cyclical. Otherwise they are rather mature so that growth rates and new investments are limited.

Second, the target profitability ought to be historically high and well controlled. Desbrières and Schatt [7] show that return on equity is higher for LBO targets two years before the deal, and that return on investment is greater two years before and the year preceding the deal. We thus propose the following alternative hypothesis. *The likelihood that a target is acquired through an LBO should be positively related to its profitability (H_2).* Third, only a handful of Public-to-Private transactions (PTP) are completed in France each year because of a number of issues, arising from French corporate ownership structure and legislation [14]. Consequently, the very great majority of French LBOs involve privately held, rather small companies. To test this idea, we propose the following alternative hypotheses. *The likelihood that a target is acquired through an LBO should be negatively related to its quotation on the stock exchange (H_3) and the likelihood that a target is acquired through an LBO should be negatively related to its size and value (H_4).*

3 Sample Selection and Methodology

To test the hypotheses we construct a buy-out sample and a control sample of non-LBO transfers of shares over the period 1996–2004. Our total sample is extracted from the Zephyr database published by Bureau Van Dijk. Since 1996, this database has collected information on various types of deals including mergers and acquisitions, initial public offerings (IPOs), joint ventures and private equity deals, with no minimum deal value. Information concerns the type of deals which can be mergers[1], acquisitions of majority interests (all cases in which the acquirer ends up with 50% or more of the votes of the target), transfers of minority stakes (below 50%), LBOs, or IPOs, which involve targets[2]. Information also concerns the deal value and the deal financing and method of payment. Moreover, Zephyr collects information on the characteristics of each type of actors involved into the deals: targets, buyers and

[1] Mergers are business combinations in which the number of companies decreases after the transaction.

[2] Targets are companies being sold, or companies in which a stake is being sold.

sellers. Some variables are qualitative such as sector, quotation, country. Others are continuous such as firms' size and profitability.

In this database, we select all deals (3,495) corresponding to transfers of ownership rights which involve French targets and which were completed during the period January 1, 1996 – May 5, 2004. The availability of variables limits our sample size to 664 deals which are classified into two groups: LBOs (126 deals) *versus* non-LBOs (538 deals).

To test the hypotheses, we use then compare the results of two decision-making concurentes methods on mixed (qualitative and quantitative) predictors. We can also use decision tree method. The first method is a logistic model, run through SAS system, in which the endogenous variable is the LBO likelihood and the exogenous variables are the targets' characteristics. The second method is a mixed discriminant analysis (MDA) [1], run through SPAD system, which aims to differentiate the two groups of deals according to mixed characteristics of targets. It is a classical discriminant analysis [9, 10] carried out on the principal factors of a Mixed Principal Component Analysis of explanatory mixed variables [2]. Alternatively, the decision tree method could have been used.

The LBO likelihood is the variable we want to explain. The other variables characterize target companies. Some variables are continuous: deal value, target size (total assets and turnover) and target profitability (Return On Equity -ROE- and Return On Assets -ROA-). The deal date is taken into account by introducing a quantitative variable that represents the number of years between the deal date and 1996. The qualitative variables used are the target sector and quotation. The descriptive data of variables are presented in Tables 1 and 2.

Targets belong to different sectors. Among them, the sector of high technology is the most represented one (39.31%), followed by manufacturing industries (30.72%) and services (13.25%). Moreover, 65.36% of the deals (434) involve unquoted targets.

Table 1 Summary statistics of continuous variables

Variable	Label	Frequency	Mean	Std dev.	Min	Max
CDAT	Date (years)	657	3.478	1.668	1	8
DVAL	Deal value (Euro–Mil.)	664	115.382	561.214	0.018	7900.00
TTAS	Total assets (Euro–Mil.)	664	858.809	4516.230	0.025	53228.00
ROE	Return on equity	664	0.416	6.215	−38.320	136.05
ROA	Return on assets	664	−0.112	0.445	−3.030	0.73
TTUR	Turnover (Euro–Mil.)	664	692.874	3361.250	0.013	36351.00

Table 2 Descriptive statistics of qualitative variables

Variable modalities	Frequency	Percent	Cumulative frequency	Percent
LBO likelihood				
LBO	126	18.98	126	18.98
Non-LBO	538	81.02	664	100.00
Target sectors				
Construction	11	1.66	11	1.66
High-tech	261	39.31	272	40.96
Hotel-restaurant	7	1.05	279	42.02
Manufactured	204	30.72	483	72.74
Retail-wholesaling	59	8.89	542	81.63
Services	88	13.25	630	94.88
Transport	25	3.77	655	98.64
Utilities	9	1.36	664	100.00
Target quoted/unquoted				
Quoted	230	34.64	230	34.64
Unquoted	434	65.36	664	100.00

4 Empirical Results

With the logistic model, we find no significant link between the LBO likelihood and the deal value or the size of target, whatever may be the measure of size, total turnover or total assets. This result is not consistent with H4. This may be related to the Zephyr database coverage. The great majority of French LBOs involves privately held firms and Zephyr may cover mainly the largest deals with public information available. The significant explanatory variables of LBOs are: the target sector, quotation and ROA (Table 3). More precisely, the LBO likelihood is positively linked with ROA (consistent with H2) and with manufacturing industries and negatively linked with quotation (consistent with H3) and high technology (consistent with H1a and H1c).

With the mixed discriminant analysis method, results (Table 4) are very significant ($PROBA = 0.0001 < 5\%$). Among the introduced mixed variables, some results are the same as with the logistic regression. LBO targets exhibit higher ROA than other targets (consistent with H3). They are more frequently unquoted (consistent with H3) and belong to manufacturing industries. They belong less than the average to transport industries (consistent with H1a). With the mixed discriminant analysis, we find that LBO targets also exhibit higher ROE (consistent with H2) and belong more often to the sector of Retail and wholesaling (partially consistent with H1b). According to the mixed discriminant analysis and contrary to the logistic model, the high technology industry does not differentiate between the two groups of transfers of shares.

Table 3 Binary logistic model – SAS results

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard error	Wald chi-Square	Pr > ChiSq
Intercept		1	−1.7206	0.3403	25.5619	<.0001
TZCL	Construction	1	0.2514	0.6927	0.1317	0.7167
TZCL	High-tech	1	−1.5692	0.3429	20.9396	<.0001**
TZCL	Hotel-restaurant	1	0.6074	0.9715	0.3909	0.5318
TZCL	Manufactured	1	0.5762	0.2711	4.5174	0.0336*
TZCL	Retail-wholesaling	1	0.4191	0.3738	1.2573	0.2622
TZCL	Services	1	−0.5872	0.3679	2.5482	0.1104
TZCL	Transport	1	−0.6023	0.5460	1.2170	0.2699
TQUO	Quoted	1	−1.1234	0.1677	44.8537	<.0001**
CDAT	Code date	1	−0.0216	0.0670	0.1039	0.7471
DVAL	Deal value (Millions)	1	0.000237	0.000236	1.0099	0.3149
TTAS	Target total assets	1	0.000072	0.000162	0.1970	0.6571
ROE	Return on equity	1	0.0609	0.0589	1.0688	0.3012
ROA	Return on assets	1	2.6568	0.5975	19.7753	<.0001**
TTUR	Target turnover	1	−0.00023	0.000227	1.0405	0.3077

** Significance less or equal than 1%; * Significance]1–5%]

Finally, when we compare the number of observations well classified (Table 5) with each method, we can conclude that the performances of the two methods are quite the same.

5 Discussion and Conclusion

This paper provides an empirical test of four hypotheses about private equity firms' acquisition rationale. The characteristics of companies undergoing LBO transactions have been extensively investigated within the US and the UK but not in continental Europe. This gap in the literature is critical for France as [7, 8, 15, 16] showed that the implications for French LBOs are unique (sources, targets' ex post performance, selection by LBO firms, etc.). Our study examines whether the characteristics of French LBO targets differ from those of firms that have been transferred through another type of deal.

To test the hypotheses we construct a buy-out sample and a control sample of non-LBO transfers of shares over the period 1996–2004. In the first method used, a classical logistic regression, we use a dummy variable to discriminate between the two groups of deals. To check the robustness of our results, we also use a second method, a mixed discriminant analysis which is, to our knowledge, new to the literature on private equity and LBOs.

Results confirm our main theoretical prediction according to which the characteristics of LBO targets differ significantly from the characteristics of other firms that have not been sold through an LBO. More precisely, results show, as expected, that LBO targets are more profitable [7, 8, 16], that they are more frequently unquoted,

Table 4 Mixed discriminant analysis – SPAD results

Fisher's Linear Function: Variables ... Num Labels	Correlations Variables With D.L.F	Parameter Estimate Discriminant Function	Parameter Estimate Regression	Standard Deviation (Res. Type Reg.)	T Value	Proba
LBO						
6 Deal value (Millions)	−0.019	0.0001	0.0000	0.0001	0.54	0.588
10 Target total assets	−0.071	0.0001	0.0000	0.0000	0.72	0.473
11 ROE - Return on equity	0.066	0.0462	0.0141	0.0056	2.52	0.012*
12 ROA - Return on assets	0.197	1.3883	0.4243	0.0853	4.97	0.000**
17 Unquoted	0.240	0.0000	0.0000	0.0000	0.00	0.000**
19 Manufactured ind.	0.244	1.6090	0.4917	0.1895	2.59	0.010**
20 Construction	0.058	1.6514	0.5047	0.3246	1.55	0.120
21 Hotel and restaurant	0.025	1.6442	0.5025	0.3840	1.31	0.191
22 Retail-wholesaling	0.092	1.5460	0.4725	0.2145	2.20	0.028*
23 Services	−0.042	0.1503	0.0459	0.2028	0.23	0.821
24 Utilities	0.010	1.5741	0.4811	0.3584	1.34	0.180
25 Transport	−0.015	0.0001	0.0000	0.0001	0.00	0.000**
Non LBO						
5 Code date	0.016	−0.0073	−0.0022	0.0211	0.11	0.916
13 Target turnover (Millions)	−0.076	−0.0001	−0.0000	0.0000	1.24	0.214
16 Quoted	−0.240	−2.1927	−0.6701	0.0783	8.55	0.000**
18 High tech	−0.272	−0.6560	−0.2005	0.1893	1.06	0.290
INTERCEPT		−0.082825	0.151059	0.1985	0.7611	0.4469

** Significance less or equal than 1%; * Significance]1–5%]

Table 5 Comparison – Classification results

Number of Observations well classified into Group (Percent)				
	Training Sample (80%)		Test Sample (20%)	
Method	(Frequency : 664)	Total	(Frequency : 166)	Total
LOGISTIC	513(81.82%)	627*	121(81.76%)	148*
MDA	539(81.17%)	664	135(81.33%)	166

* Missing values

and that they more often belong to mature and rather non-cyclical industries [17]. Interestingly, we do not identify any sign of abnormality in the selection of French LBO targets by private equity firms over the period 1996–2004. Our results suggest that private equity firms behave in accordance with financial standards when they screen targets for LBO deals. This is not consistent with the study of [16], which finds an unexpected risky profile of French LBO targets. This is also not consistent with Wright et al. [20] who argue that, if we consider LBOs as a vehicle for strategic innovation and renewal that stimulates growth opportunities, then the need for a low business risk of LBO targets becomes less necessary, LBO firms seeking above all to realize entrepreneurial opportunities.

Finally, our analysis relies on data from a single country, France, where the private equity industry has already entered its maturity phase and LBO firms have had the opportunity to accumulate relevant experience. This raises concern about the generalizability of our results to other countries, in particular to those with significantly less developed private equity markets such as, for instance, Italy and Spain. Hence future research might examine and compare the selection of LBO targets in different European countries.

References

1. Abdesselam, R.: Discriminant analysis on mixed predictors. In: Lauro, C., Palumbo, F., Greenacre, M. (eds.) Book Series, Studies in Classification, Data Analysis, and Knowledge Organization, Data Analysis and Classification: From the Exploratory to the Confirmatory Approach. Springer, Heidelberg (2008)
2. Abdesselam, R.: Analyse en Composantes Principales Mixte. Classification: points de vue croisés, RNTI-C-2, Revue des Nouvelles Technologies de l'Information RNTI, Cépadués Editions, pp. 31–41 (2008)
3. AFIC: Rapport sur l'activité du capital investissement en France. http://www.afic.asso.fr/ (2006)
4. AFIC: Le PtoP a-t-il un avenir en France? http://www.afic.asso.fr/ (2004)
5. Cressy, R., Munari, F., Malipiero, A.: Playing to their strenghts? Evidence that specialization in the private equity industry confers competitive advantage. J. Corp. Finance **13**, 647–669 (2007)
6. Cumming, D., Siegel, D.S., Wright, M.: Private equity, leveraged buyouts and governance. J. Corp. Finance **13**, 439–460 (2007)
7. Desbrières, P., Schatt, A.: The impacts of LBOs on the performance of acquired firms: the French case. J. Bus. Finance Account. **29**(5&6), 695–729 (2002)

8. Desbrières, P, Schatt, A.: L'incidence des LBO sur la politique d'investissement et lagestion opérationnelle des firmes acquises: le cas français. Finance Contrôle Stratégie **5**(4), 79–106 (2002)
9. Geoffrey, J. McLachlan: Discriminant Analysis and Data Statistical Pattern Recognition. Wiley-Interscience, New York, NY (2005)
10. Hand, D.: Discrimination and Classification. Wiley, New York, NY (1981)
11. Jensen, M.: Agency costs of free cash flow, corporate finance, and takeovers. Am. Econ. Rev. **76**(2), 323–329 (1986)
12. Jensen, M.: Eclipse of the public corporation. Harv. Bus. Rev. **67**(5), 61–74 (1989)
13. Johnson, S.: An empirical analysis of the determinants of corporate debt ownership structure. J. Finance Quant. Anal. **32**(1), 47–69 (1997)
14. Kale, J., Noe, T., Ramirez, G.: The effect of business risk on corporate capital structure: theory and evidence. J. Finance **46**(5), 1693–1715 (1991)
15. Lehn, K., Poulsen, A.: Leveraged buyouts: wealth created or wealth redistributed? In: Weidenbaum, M., Chilton, K. (eds.) Chapter 4 in Public Policy Towards Corporate Takeovers. Transaction Publishers, New Brunswick, NJ (1988)
16. Le Nadant, A.L.: La Performance des Sociétés Cibles dans les Opérations de LBO: Etude du Marché Françis. Anal. Finan. **116**, 67–85 (1998)
17. Le Nadant, A.L., Perdreau, F.: Financial profile of leveraged buyout targets: some french evidence. Rev. Account Finance **5**(4), 370–392 (2006)
18. Maupin, R.: Financial and stock market variables as predictors of management buyouts. Strateg. Manag. J. **8**(4), 319–327 (1987)
19. Singh H.: Management buyouts: distinguishing characteristics and operating changes prior to public offering. Strateg. Manag. J. **11**(5), 111–129 (1990)
20. Wright, M., Hoskisson, R. E., Busenitz, L.W., Dial, J.: Finance and management buyouts: agency versus entrepreneurship perspectives. Venture Cap. **3**(3), 239–261 (2001)

Kernel Intensity for Space-Time Point Processes with Application to Seismological Problems

Giada Adelfio and Marcello Chiodi

Abstract Dealing with data coming from a space-time inhomogeneous process, there is often the need of semi-parametric estimates of the conditional intensity function; isotropic or anisotropic multivariate kernel estimates can be used, with windows sizes **h**. The properties of the intensities estimated with this choice of **h** are not always good for specific fields of application; we could try to choose **h** in order to have good predictive properties of the estimated intensity function. Since a direct ML approach cannot be followed, we propose an estimation procedure, computationally intensive, based on the subsequent increments of likelihood obtained adding an observation at time. The first results obtained are very encouraging. Some application in statistical seismology is presented.

1 Introduction

When dealing with data coming from a space-time inhomogeneous process, like seismic data, fire data, or even disease data, there is often the need of obtaining reliable estimates of the conditional intensity function, or of the marginal intensity function. According to the field of application, intensity function can be estimated through some assessed parametric model, where parameters are estimated by Maximum Likelihood method and then intensities (conditional or marginal) are estimated using the parameter estimates. In an exploratory context , some kind of nonparametric estimation is required [4]; we could also have this necessity if we need to assess the adequacy of an estimated parametric model; in some other model, like ETAS model [8], or in a clustered intensity function [3], some component of the spatial intensity function is not made explicit and must be estimated from data in a nonparametric way. Often, isotropic or anisotropic kernel estimates can be used, using the Silverman rule to choose the windows sizes **h** [9]. If the purpose of the study is just the estimation of **h**, to choose **h** in order to have good predictive properties

G. Adelfio (✉)
Department of Statistical and Mathematical Sciences, University of Palermo, 90128 Palermo, Italy
e-mail: adelfio@dssm.unipa.it

B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-13312-1_42,

of the estimated intensity function could be useful; for this purpose, a direct ML approach cannot be followed, unless we use a penalizing function.

In [4] the seismicity of the Southern Tyrrhenian Sea is described by the use of Gaussian kernels and the optimum value of $\mathbf{h}$ is chosen such as to minimize the mean integrated square error (MISE) of the estimator $\hat{f}(\cdot)$. A variable bandwidth procedure is proposed in [1].

In the next section a predictive approach for the bandwidth parameters estimation is presented, while in the third section some kind of application to statistical seismology is briefly sketched. Conclusive remarks and directions for future works are provided in the last section.

2 Intensity Function and Predictive Likelihood

Suppose we have a general d-dimensional closed region, Z^d and that one of the dimension is $t \in T$, the time, or however a dimension with a *meaningful ordering* such that $Z^d = S^{d-1} \times T$. Let $\mathcal{P}$ a random collection of k points in Z^d from time t_1 until the time t_k such that $i < j \iff t_i < t_j$ and each observation P_i is constituted by: $\mathbf{z}_i^{\mathrm{T}} = \left\{\mathbf{s}_i^{\mathrm{T}}, t_i\right\}$, $i = 1, 2, \ldots, k$; the conditional intensity function of the process is:

$$\lambda(\mathbf{z}) = \lambda(\mathbf{s}, t|H_t) = \lim_{\Delta t, \Delta \mathbf{s} \to 0} \frac{\mathrm{E}\left[\#(t, t+\Delta t; \mathbf{s}, \mathbf{s}+\Delta \mathbf{s}|H_t)\right]}{\Delta t \, \Delta \mathbf{s}}$$

where H_t is the space-time occurrence history of the process up to time t; Δt and $\Delta \mathbf{s}$ are time and space increments; $\mathrm{E}\left[\#(t, t+\Delta t; \mathbf{s}, \mathbf{s}+\Delta \mathbf{s}|H_t)\right]$ is the history-dependent expected number of events occurring in the volume $[t, t+\Delta t] \times [\mathbf{s}, \mathbf{s}+\Delta \mathbf{s}]$.

Assuming that θ is a vector of smoothing parameters in a semi-parametric context, the log-Likelihood for the point process [7], given the m observed values $\mathbf{z}_i$, is:

$$\log L(\boldsymbol{\theta}) = \sum_{i=1}^{m} \log \lambda(\mathbf{z}_i; \boldsymbol{\theta}) - \int_{T_0}^{T_{max}} \int_{\Omega_{\mathbf{s}}} \lambda(\mathbf{z}; \boldsymbol{\theta}) \, d\mathbf{s} \, dt \tag{1}$$

where $\Omega_{\mathbf{s}}$ is the observed space region and $(T_0 - T_{max})$ is the observed period of time and the intensities $\lambda(\cdot)$ depend on unknown parameters $\boldsymbol{\theta}$ estimated by $\hat{\boldsymbol{\theta}}(H_{t_m}) \equiv \hat{\boldsymbol{\theta}}(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_i, \ldots, \mathbf{z}_m)$.

In the rest of the paper we use the log-likelihood in (1) evaluated at $\hat{\boldsymbol{\theta}}(H_{t_m})$, that is:

$$\log L(\hat{\boldsymbol{\theta}}(H_{t_m}); H_{t_m}) = \sum_{i=1}^{m} \log \lambda(\mathbf{z}_i; \hat{\boldsymbol{\theta}}(H_{t_m})) - \int_{T_0}^{T_{max}} \int_{\Omega_{\mathbf{s}}} \lambda(\mathbf{z}; \hat{\boldsymbol{\theta}}(H_{t_m})) \, d\mathbf{s} \, dt \tag{2}$$

We try to find a trade-off between fitting to observed data and prediction of future data; the context of space-time point processes is different from regression problems, where we can use cross validation techniques, or from time series context, where we can compare the observed value y_{m+1} with an estimated value $\hat{y}_{m+1}$, depending only on the previous m observations. The problem does not arise if we use likelihood computed on different sets of data and with different estimates. We use a variation of the likelihood to measure the ability of the observations until t_m to give information on the next observation. Let:

$$\log L(\hat{\boldsymbol{\theta}}(H_{t_m}); H_{t_{m+1}}) = \sum_{i=1}^{m+1} \log \lambda(\mathbf{z}_i; \hat{\boldsymbol{\theta}}(H_{t_m})) - \int_{T_0}^{t_{m+1}} \int_{\Omega_{\mathbf{S}}} \lambda(\mathbf{z}; \hat{\boldsymbol{\theta}}(H_{t_m}))\, d\mathbf{s}\, dt \tag{3}$$

be the likelihood computed on the first $m+1$ observation but using the estimates based on observation *only until* t_m; e.g. $\lambda(\mathbf{z}; \hat{\boldsymbol{\theta}}(H_{t_m}))$ can be an intensity function computed by an anisotropic kernel method with a multivariate window $\hat{\boldsymbol{\theta}}$ using the first m points. So we use the difference between (3) and (2) to measure the predictive information of the first m observations on the $(m+1)-th$:

$$\delta_l(\hat{\boldsymbol{\theta}}(H_{t_m}); H_{t_{m+1}}) = \log L(\hat{\boldsymbol{\theta}}(H_{t_m}); H_{t_{m+1}}) - \log L(\hat{\boldsymbol{\theta}}(H_{t_m}); H_{t_m}) \tag{4}$$

For the sake of brevity, we report here only essential ideas with few details; briefly we use $\delta_l(\hat{\boldsymbol{\theta}}(H_{t_m}); H_{t_{m+1}})$ to estimate some smoothing parameter $\boldsymbol{\theta}$. In a fashion similar to cross-validation criterion we could choose $\tilde{\boldsymbol{\theta}}(H_{t_m})$ which maximizes a predictive likelihood:

$$FLP_{m_1,m_2}(\hat{\boldsymbol{\theta}}) = \sum_{m=m_1}^{m_2} \delta_l(\hat{\boldsymbol{\theta}}(H_{t_m}); H_{t_{m+1}}) :$$
$$FLP_{m_1,m_2}(\tilde{\boldsymbol{\theta}}) \geq FLP_{m_1,m_2}(\hat{\boldsymbol{\theta}}) \quad \forall \hat{\boldsymbol{\theta}} \in \boldsymbol{\Theta}$$

with $m_2 = k-1$ and maybe m_1 is such that $t_{m_1} - T_0 \approx \frac{T_{max}-T_0}{2}$. Although this aspect will not be introduced in the present paper, we could use the quantities in (4) for diagnostic purposes, in comparison with some previous approach, such as the one in [2].

The method seems to give better kernel estimates of space-time intensity function with respect to classical methods, either using isotropic or anisotropic kernel function.

The solution of the approach is almost objective and data-driven. On the other hand it is computationally expensive [5], although some approximations are here introduced to improve the speed of computation of multiple integrals in the likelihood (based on Gaussian quadrature), very useful when anisotropic kernels are considered.

We applied this technique to seismological data, although it is capable to be applied in quite different contexts.

3 Applications to Seismological Field

One of the most important model in statistical seismology is the ETAS model [8], a self-exciting point process describing earthquakes catalogs as a realization of a branching or epidemic-type point process. The main hypothesis of the model states that all events, both a mainshock or an aftershock, have the possibility of generating offsprings. The conditional intensity function of the ETAS model in a point x, y, t, m is defined by:

$$\lambda(x, y, t, m|\mathcal{H}_t) = J(m)(\mu(x, y) + \sum_{t_j<t} g(t - t_j) f(x - x_j, y - y_j|m_j)) \quad (5)$$

where $J(m)$ is the magnitude distribution, $g(\cdot)$ and $f(\cdot)$ are parametric temporal and spatial functions and $\mu(x, y)$, estimated by semi-parametric approach, describes the spontaneous activity.

In [3] a seismic catalog is described as the realization of a clustered inhomogeneous Poisson process, that is obtained assuming that points of the background seismicity come from a space-time Poisson process (spatially inhomogeneous) and that among these there is a number k of mainshocks that can generate aftershocks sequences, inhomogeneous both in space and times, and with an intensity also linked to the magnitude of the main event. The intensity function is:

$$\lambda(x, y, t; \boldsymbol{\theta}) = \lambda_t \mu(x, y) + K_0 \sum_{j=1;\, t_j<t}^{k} g_j(x, y) \frac{\exp[\alpha(m_j - m_0)]}{(t - t_j + c_j)^{p_j}}$$

where $\boldsymbol{\theta} = (\lambda_t, K_0, c_j, p_j, \alpha)$; t_j and m_j are time of the first event and magnitude of the mainshock of the cluster j, $g_j(x, y)$ is the space intensity of the cluster j and $\mu(x, y)$ is the background one; K_0 and λ_t are the weights of the clustered seismicity and of the background one, respectively; c_j and p_j are parameters of the clusters time distributions to be estimated; $g_j(x, y)$ and $\mu(x, y)$ must be estimated by semi-parametric methods.

3.1 Evaluation of Seismic Gap

A number of statistical models with intensities $\lambda(\cdot; \boldsymbol{\theta})$ have been proposed for representing the intensity function of earthquakes. The parametric models estimation suffers by many drawbacks, often related to the definition of a reliable mathematical

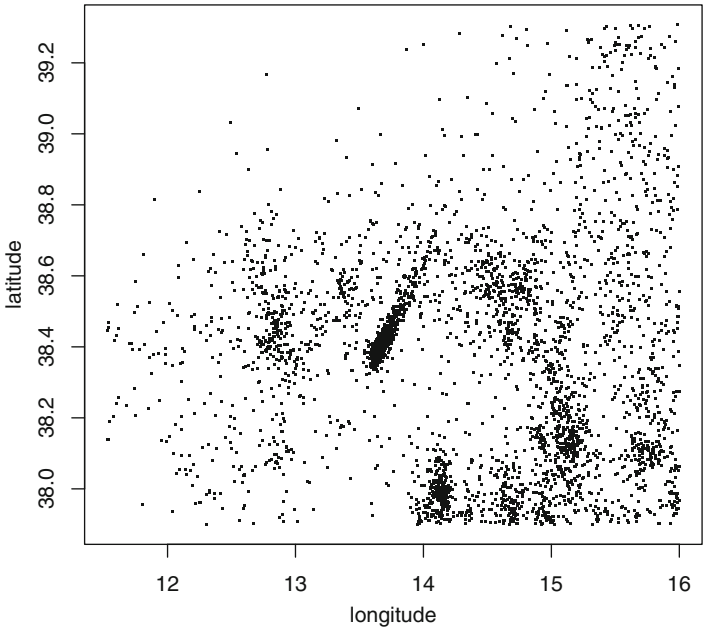


Fig. 1 Epicenters of earthquakes occurred in the South Tyrrhenian Sea from 1981 to 2005, in the region defined by 37.9° $\sim$ 39.31°N and 11.52° $\sim$ 16°E for all depth and magnitude

model from the geophysical theory and to the sensitivity of statistical estimates to the composition of the space-time region under study. Many of the disadvantages of the parametric modelling can be avoided by using also nonparametric techniques, such as those presented in this paper, which provides estimate with few constrains and are supposed to fit well to observed data.

In this paper, given a specific space-time region of interest, a parametric model is compared with a nonparametric one to try to identify the so called seismic gap. A seismic gap can be defined as a segment of an active geologic fault that has not produced seismic events for an unusually long time; gaps are often considered susceptible to future strong earthquakes occurrence and therefore their identification may be useful for predictive purposes.

For this purpose we analyze the seismic activity of the South Tyrrhenian Sea from 1981 to 2005 (epicentral coordinates are showed in Fig. 1).

In Fig. 2 the three dimensional contour-plot of the nonparametric space-time intensity function estimated for the observed seismicity is showed.

The values of the parameters **h** estimated by the proposed approach are reported in Table 1.

Therefore, we compare the ETAS model with intensity function given in (5) and estimated by ML, with the nonparametric one by a graphical approach (see Fig. 3 for latitude-time domain).

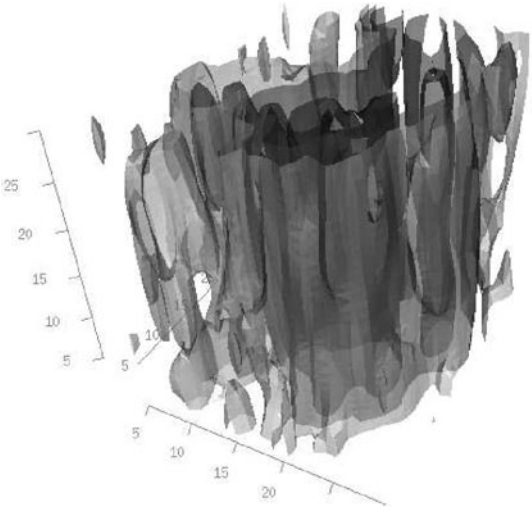


Fig. 2 Three dimensional contour-plot (three levels) of the space-time intensity function for the seismicity of the observed area

Table 1 Values of **h** estimated for the seismicity of the observed area by the proposed approach

h_t (days)	h_x (Degree of longitude)	h_y (Degree of latitude)
60.2083	0.1259	0.0738

Regions with ratio values smaller than one are identified by a brighter grey, while regions with a ratio larger than one are identified by a darker grey: darker areas indicate that the observed seismicity is smaller than those calculated by the estimated space-time ETAS model.

Darker grey area around the source region before large earthquakes that induced a big sequence of events may be observed.

Also in time domain, this approach can identify gaps interval most of all before main event. In Fig. 4 (on the top) temporal kernel estimation is showed, together with quiescent periods, defined as those time intervals for which the intensity $\lambda(\cdot)$ is less than a fixed threshold (in the fig. it is identified by the horizontal broken line); this threshold has been suggested by subjective choices, as discussed in [6]. In this figure, the quiescent periods are indicated by shaded intervals, also reported in the time-magnitude plot (on the bottom) for the events with magnitude greater than 4.5. In particular we observe that this periods seem to occur mostly before events with large magnitude values, that induce a sequence of events and are denoted by peaks of intensity and on the other hand periods that have been identified as quiescent ones do not contain large events.

Although we think that this approach is just a starting analysis to deal with this kind of issue, since it needs further data to be available and the application of more rigorous methodology, it may provide a useful starting-point for studies aimed to seismic hazard evaluation.

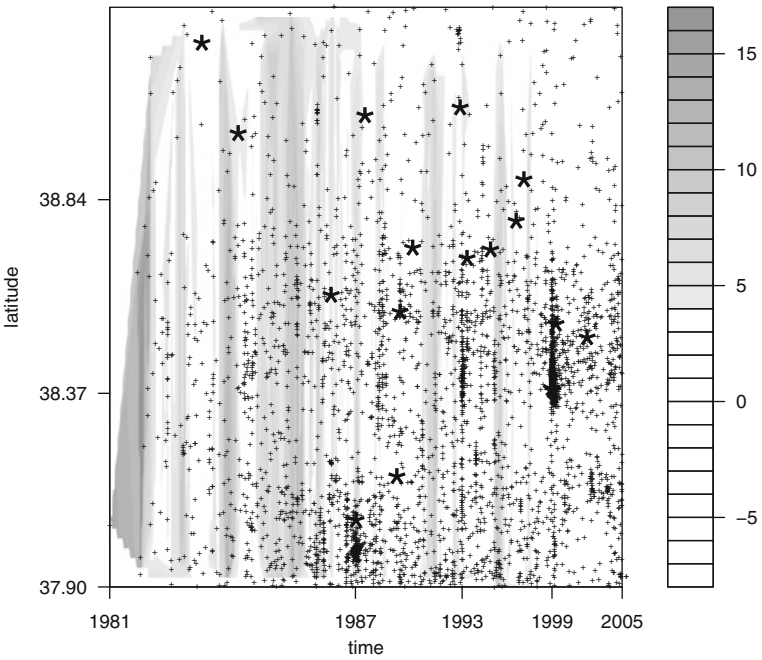


Fig. 3 Ratio between the parametric (ETAS) and nonparametric intensity estimate of the seismic activity of the South Tyrrhenian Sea from 1981 to 2005 (latitude vs time); $*$ symbol is used for events with magnitude greater than 4.5

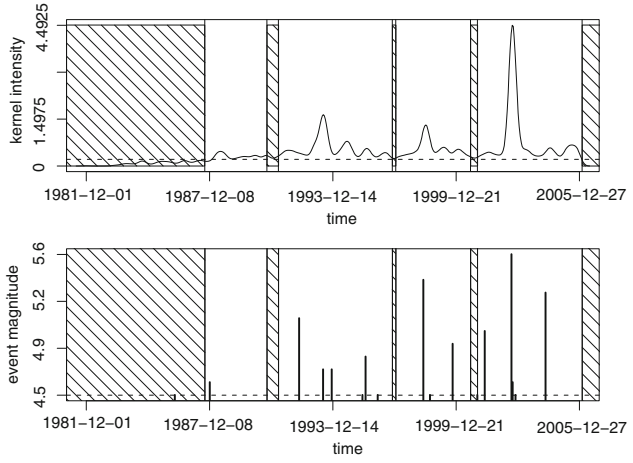


Fig. 4 Nonparametric time intensity estimate of the seismic activity of the South Tyrrhenian Sea from 1981 to 2005 (*on the top*) and quiescent periods for events with magnitude greater than 4.5, represented by vertical spikes (*on the bottom*)

4 Conclusive Remarks

In this paper a nonparametric approach for space-time point processes is introduced. This method is based on a variation of the likelihood function to assess the capability of each observation to give information on the next ones. The introduced approach is used for describing the seismicity of an observed area in the Southern Tyrrhenian Sea.

The nonparametric approach makes possible a reasonable characterization of the observed seismicity, since it does not constrain the process to have predetermined properties.

The estimated model seems to follow adequately the seismic activity of the observed area, characterized by highly variable changes both in space and in time and because of its flexibility, it provides a good fitting to local space-time changes as just suggested by data.

The method is actually still in progress, since we are developing a nonparametric model with variable bandwidth values, to study variations of seismic activity in space and time and to analyze possible correlation between the estimated intensity function and particular distributions of some structural features (i.e. geological structures) of the studied region.

Acknowledgments This paper and the related work have been supported by research fund of University of Palermo and by PRIN funds 2006.

References

1. Adelfio, G.: An analysis of earthquakes clustering based on a second-order diagnostic approach. Studies in Classification, Data Analysis, and Knowledge Organization. Data Analysis and Classification: from the Exploratory to the Confirmatory Approach. Springer, Heidelberg (2008)
2. Adelfio, G., Chiodi, M.: Second-order diagnostics for space-time point processes with application to seismic events. Environmetrics **20**, 895–911 (2009)
3. Adelfio, G., Chiodi, M., De Luca, L., Luzio, D.: Nonparametric clustering of seismic events. In: Data Analysis, Classification and the Forward Search, pp. 397–404. Springer, Berlin (2006)
4. Adelfio, G., Chiodi, M., De Luca, L., Luzio, D., Vitale, M.: Southern-Tyrrhenian seismicity in space-time-magnitude domain. Ann. Geophys. **49**(6), 1245–1257 (2006)
5. Chiodi, M., Adelfio, G.: Semiparametric estimation of conditional intensity functions for spacetime processes. Presented to the scientific meeting of Italian Statistical Society, Cosenza (2008)
6. Choi, E., Hall, P.: Nonparametric approach to analysis of space-time data on earthquake occurrences. J. Comput. Graph. Stat. **8**(4), 733–748 (1999)
7. Daley, D.J., Vere-Jones, D.: An Introduction to the Theory of Point Processes, 2nd edn. Springer, New York, NY (2003)
8. Ogata, Y.: Statistical models for earthquake occurrences and residual analysis for point processes. J. Am. Stat. Assoc. **83**(401), 9–27 (1988)
9. Silverman, B.W.: Density Estimation for Statistics and Data Analysis. Chapman and Hall, London (1986)

Summarizing and Mining Streaming Data via a Functional Data Approach

Antonio Balzanella, Elvira Romano, and Rosanna Verde

Abstract In recent years, the analysis of data streams has become a challenging task since many applicative fields generate massive amount of data that are difficult to store and to analyze with traditional techniques. In this paper we propose a strategy to summarize pseudo periodic streaming data affected by noise and sampling problems, by means of functional profiles. It is a clustering strategy performed in a divide and conquer manner. In the on-line step, a set of summarization structures, collect statistical information on data. Starting from these, in the off-line step, the final clustering structure and the set of functional profiles are computed.

1 Introduction

Recent advances in sensors technology have motivated the development of strategies for the analysis of data generated continuously over time. Application fields include medical information management, climate monitoring and forecasting, telecommunications.

This huge, potentially unbounded, amount of data cannot be entirely stored and the requirement of real time monitoring makes not feasible the usual mining techniques.

Strategies for the so called data streams, present several computational and mining challenges:

- It is no longer possible to process data efficiently by using multiple passes
- The data, after processing, are discarded or archived such to be not easily available
- The memory resources are reduced with reference to the amount of data to process

In such a context, high quality approximated answers can be acceptable, if these are available just when required.

A. Balzanella (✉)
Universitá degli Studi di Napoli Federico II, 80126 Napoli, Italy
e-mail: balzanella2@alice.it

B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-13312-1_43,

Approximation algorithms often use summarizing structures (synopsis) that are incrementally updated with the on-line collecting of data. Some example are: histograms, wavelets, sketches.

In this paper we will focus on summarizing pseudo periodic streaming data. To reach this aim, we detect proper summarizing structures by means of a clustering algorithm which will perform a single pass on data.

We assume that our reference data, are a set of discrete measured values of an unknown function $f(t)$, characterized by groups of measurements which repeat over time with tiny variations. Such variations are risen by noise, sampling frequency which changes over time, informative content.

The novelty of the approach consists in the introduction of functional profiles for the representation of sets of similar item-sets, such to keep into account noise and variable sampling frequency.

To our knowledge, the problem of summarizing pseudo periodic data in data stream framework has been only recently dealt in [6]. The authors propose to build a graph to summarize data. The conceptual schema is to split the incoming data into waves, detected using valley points, to represent these using Piecewise Linear Representation (PLR) and to use waves matching for updating the graph. Especially a new wave, after to be transformed to a segments sequence, is tested for matching to the data in the graph. If no matching is found within an error bound, a new element is added to the graph; if there is full matching a counter for the wave is increased; if only some segment in the PLR representation matches, the unmatched segments are added to the graph.

Our approach shares some idea with [2] since it integrates the micro-cluster technique to perform on-line summarization preserving the locality of the data and the snapshots to recall summary statistics from different time horizons.

This is a divide and conquer approach since a wide set of statistical information on data is on-line collected and starting from these a final summarizing structure is built.

The statistical information are stored into micro-clusters which are continuously updated with the arrival of new data points. Each micro-cluster summarizes set of data selected through a similarity criterion. Their number is chosen to be as wide as possible constrained by the computational and storing resources. An off-line clustering strategy can be performed taking as input the summaries represented by micro-clusters to get higher level clusters and profiles, which can be more easily understood by the user.

The micro-clusters are stored at specific time points, which are referred as snapshots, in order to keep the history of the streams.

Our approach differs from this last one since the input data are subsequences rather than single multidimensional points. We introduce a different definition of micro-cluster where the main concept is a functional representation of a set of subsequences of the stream. Moreover, we aim at finding functionals profiles to summarize the whole data stream.

The rest of the paper is organized as follows. In Sect. 2 we provide the details of the proposed strategy, in Sect. 3 we review the strategy on data, in Sect. 4 conclusions and open problems are discussed.

2 A Functional Approach for Dealing with Streaming Data

Before to introduce our strategy for summarizing a streaming time series we provide some definition and notation.

Let S be a streaming time series defined as an ordered set $S = [S_1, \ldots, S_t, \ldots, S_\infty]$ of unbounded real valued variables observed on a grid $T = [T_1, \ldots, T_t, \ldots, T_\infty]$. It is possible to define the subsequence $Q = \left[q_{t'}, \ldots, q_{t''}\right]$ as a finite sampling of ordered values of S.

Each subsequence is obtained starting from a time-based window W, that is a finite set of elements of T with variable size $\Delta(t)$.

Based on the above definition and according to our procedure, each subsequence Q represents the raw functional form [5]. Since noise is part of the data we determinate a true functional form, called functional subsequence (*f-sub*), which describes the trend of the flowing data, by using regression spline functions. A functional subsequence is defined as follows:

Definition 1 Let Q be a subsequence, and $t^{'} = \xi_1 < \ldots < \xi_j < \ldots < \xi_a = t^{''}$ be a sharing of $[t^{'}, t^{''}]$ in a distinct points, called knots, a functional subsequence (*f-sub*) is a regression spline function obtained by minimizing a linear least square problem. In particular B-spline basis functions are used as basis functions for univariate regression. Formally, a functional subsequence can be written as:

$$f - sub = \sum_{l=1}^{p} \beta_l B_l(t) \tag{1}$$

Where $p = a + splineorder + 1$ are the number of parameters, $\beta = \left(\beta_1, \ldots, \beta_p\right)^{'}$ is the vector of spline coefficients and $B = \left(B_1, \ldots, B_p\right)$ are the B-spline functions [3].

Based on the above definitions our strategy is performed as follows:

1. On-line step
    a. splitting S into subsequences of different length;
    b. identification of functional micro-clusters to collect statistical information on detected subsequences;
2. Off-line step
    a. profiles identification starting from the functional microclusters;

The first challenge consists in defining a criterion to select subsequences.

Evolving streams with pseudo periodicity are composed of waves with various time lengths and key values. The idea is to check the fluctuations of the streaming

time series and to use the change points (minima and maxima points) to select the subsequences.

The detection is performed starting from the time windows on data. These are identified through two threshold values h_1 and h_2 given as input.

Each time window has the beginning time point t' in correspondence of a maximum point M_1 such that $M_1 > h_1$ and the ending t'' in correspondence of the following maximum point M_2 such that $M_2 > h_1$. An additional time point $t' < t* < t''$ such that $S(t*)$ is a minimum point and $S(t*) < h_2$, has to exist to define the time window.

To deal with the evolution of the data flow, the threshold values h_1 and h_2 are incrementally updated using the average value of the past detected change points. For h_1, the updating is performed as follows:

$$h_1 = \alpha \left(\sum_{j=1}^{j=N} M_1 \right) \div N \tag{2}$$

where N is the number of past maxima points and α is an adjustment parameter. The updating of h_2 can be performed in a similar way, using the past minima points.

Starting from the subsequences the central issue is to on-line collect the statistics needed for computing the profiles in the off-line step. The functional micro-clusters are the tool we advice to solve this problem.

A f-microcluster for a set of f-sub is a data structure constituted by the following components:

- the functional prototype $g_c(t)$, which summarizes a set of f-subs;
- the number n_c of allocated f-subs;
- the beginning time point t' and the ending time point t'' for each allocated f-sub.

Starting from a set of k f-micro-clusters, the on-line algorithm tries to allocate each new detected f-sub to an existing f-micro-cluster according to a dissimilarity criterion. If exists a f-micro-cluster such that the computed distance is less than a threshold value, the f-sub is allocated. On the contrary, a new f-micro-cluster is created and the functional subsequence is chosen as prototype. Then, the f-micro-cluster is updated according to the new information provided.

The updating consists in increasing the number of allocated functional subsequences, in storing the beginning time point t' and the ending time point t'' of the subsequence and in computing the new prototype.

Since subsequences, due to not constant sampling frequency, come from windows of different temporal length, we have to introduce a criterion to compute dissimilarity taking into account this issue.

We propose to stretch or shrink each detected subsequence to a new time grid with a common size $\Delta(t_p) = (t''_p - t'_p)$. This is performed before to compute the relative functional subsequence.

Given a subsequence Q obtained from a time based window $W = \left[t', \ldots, t_m, \ldots, t''\right]$, the stretched or shrank subsequence Q will be defined on

a time grid W_2 of size $\Delta(t_p)$ such that it is a set of discrete ordered elements $\left[t'_p, \ldots, t_n, \ldots, t''_p\right]$ of T where $t_n = (t_m - t') * \Delta(t_p)/(t'' - t') + t_{p'}$.

The stretched or shrinked subsequences are used to compute f-subs. Since the latter are defined on $\Delta(t_p)$ the heterogeneity across f-subs is captured by the heterogeneity of the estimated coefficients [1].

In order to deal with query, where it is asked to summarize the data structure before a given time point and to monitor the evolution of the monitored phenomenon, it is proposed to store on disk the set of f-micro-clusters at prefixed time points, these are referred as snapshots according to [2].

In the Off-line analysis of functional micro-clusters step, the summary of the data structure is mined. It is a k-means clustering procedure where the input data are the prototypes stored in the f-microclusters on-line collected. The allocation step assigns each prototype to a class according to the proximity to the mean profile. This is followed by the representation step where the mean profiles are computed as average function of prototypes assigned to the classes at the allocation step weighted by the number of the represented curves n_c, until the convergence of the algorithm.

3 Main Results

In order to show the effectiveness of our strategy we have performed tests on real and synthetic datasets. Here we review the main results for one of these.

It is an univariate time series containing one million data points[4]. It shows high periodicity but never exactly repeats itself. Each observation is generated by independent invocations of the function:

$$\bar{y} = \sum_{i=-3}^{7} \frac{1}{2^i} sin(2\pi(2^{2+i} + rand(2^i))\bar{t}) \tag{3}$$

where $0 \leq \bar{t} \leq 1$.

Original data have been split into 10 sections, here we consider the time series obtained by joining these sections to get a single time series made by one million data points.

To get our summarizing structure we perform our experiments using several parameters sets (Table 1):

In the first two experiments we have evaluated the impact of the initial max and min thresholds, on the capability to detect proper windows.

These are critical parameters since too high or too low values cause the detection of time windows which do not capture the periodicity in data. If too high values for h_1 and too low values for h_2 are chosen, two or more periodic patterns can be included in the same time window and the on-line clustering will generate a functional micro-cluster including, as prototype, the join of two periods. At the opposite, if too low values for h_1 and too high values for h_2 are chosen, the detected time

Table 1 Chosen parameters for the performed tests

Parameter	Experiment 1	Experiment 2	Experiment 3	Experiment 4
Initial h_2 value	-0.22	-0.11	-0.11	-0.11
Initial h_1 value	$+0.22$	$+0.16$	$+0.16$	$+0.16$
α	0.8	0.8	0.8	0.8
Number of observations for each window	100	100	100	100
Number of knots	5	5	5	5
Distance threshold	0.8	0.8	0.6	0.8

window will include only a part of the real period. This generates an unreasonable amount of functional micro-clusters.

Our procedure provides a criterion to adapt the windows cutting points that is effective for adapting to the evolution of the flow of data, however a good initial choice considerably improves the windows detection process.

The values for min and max thresholds evaluated in the first experiment show that the 75% of windows are detected in a proper way, this is becouse the adapting of the thresholds to data occurs slowly due to a wrong choice of parameters. At the opposite, the values used in the second experiment turn out to be effective, capturing correcltly the periodiceness in data.

The parameters for experiment 2, are chosen by building two empirical distributions from a training dataset: a first one for maxima points and a second one for minima points. By extracting the 90th percentile from these distributions, we get the analyzed values.

In the experiment 3 and 4, we have evaluated how the distance threshold affects the summarization quality.

Higher threshold values, strongly reduce the number of detected functional micro-clusters but deteriorate the quality of representation of each functional prototype. Lower values, improve the summarization quality but increase the number of required functional micro-clusters.

On the test dataset, the parameter set of the experiment 3, generates 82 functional micro-clusters. The number of sub-sequences allocated to each functional micro-cluster ranges from 1 to 8. For the parameter set of the experiment 4, we get 32 functional micro-clusters and the number of sub-sequences allocated to each functional micro-cluster ranges from 1 to 31.

Starting from the on-line summarization, we have evaluated the effectiveness of the off-line procedure to get the final clustering structure.

The testing has been performed comparing the representation quality of the profiles got from our procedure to the ones obtained using the k-means algorithm on stocked functional subsequences.

Especially, we evaluate if the mean square error (MSE) of functional profiles obtained starting from the analysis of the functional micro-clusters, well compares

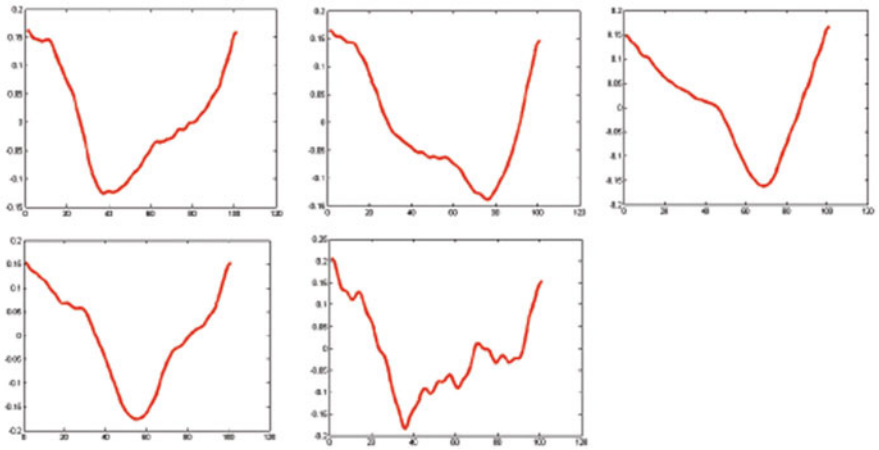


Fig. 1 A set of functional micro-clusters with the allocated functional subsequence

to the MSE of the prototypes of standard clustering algorithm performed having available the whole set of data.

To reach this aim we run the off-line weighted k-means procedure on the f-micro-cluster to get five clusters. The profiles shown in Fig. 1.

To evaluate the quality of representation with reference to the original data we need to generate a partition of the original functional subsequences starting from the partition of the functional micro-clusters prototypes.

Since each f-micro-cluster includes references to the allocated sub-sequences, for testing purposes, we can build a partition of the functional sub-sequences where each cluster includes the f-subs pointed by the f-micro-clusters prototype. Consequently, the MSE of the profiles resulting from weighted k-means, with reference to the f-subs allocated to the cluster, can be computed.

For the clustering on stocked data, the k-means algorithm has been applied on the functional subsequences on-line detected. The main difference is that we do not use the dimensionality reduction performed by on-line updating of functional micro-cluster, but we use the data of each window. The cluster number has been still set to five and the MSE has been computed for the prototypes in each cluster.

The results are evaluable in the following table (Table 1):

The results, confirm that a higher number of functional micro-cluster improve the representation quality however useful results are obtained in both cases.

Table 2 Main results

	Distance threshold = 0.6	Distance threshold = 0.8	k-means
Total MSE	2.9	3.3	2.71

4 Conclusions

In this paper, we have introduced a new strategy for summarizing streaming time series. In order to evaluate the effectiveness of our strategy in terms of accuracy and sensitivity, a study has been conducted on synthetic data.

The approach is useful to underline the sub-structures on sea waves data. The future research direction aims at examining the impact of different criteria to select the windows, as well as in the definition of a criterion to keep constant the number of f-microclusters during the online step.

References

1. Abrham, C., Corillon, P.A., Matzner-Löber, E., Molinari, N.: Unsupervised curves clustering using b-splines. Scand. J. Stat. **30**(3), 581–595 (2003)
2. Aggarwal, C., Han, J., Wang, J., Yu, P.: A framework for clustering evolving data streams. In: VLDB Conference, Berlin, Germany (2003)
3. De Boor, C.: A Practical Guide to Splines. Springer, New York, NY (1978)
4. Pseudo Periodic Synthetic Time Series. In: The UCI KDD Archive Information and Computer Science University of California, Irvine. http://kdd.ics.uci.edu/. Cited 10 Dec 2008
5. Ramsay, J.O., Silverma, B.W.: Functional Data Analysis, 2nd edn. Springer, New York, NY (2005)
6. Tang, L.A., Cui, B., Li, H.Y., Miao, G.S., Yang, D.Q., Zhou, X.B.: PGG: an online pattern based approach for stream variation mnagement. J. Comput. Sci. Technol. **23**(4), 497–515 (2008)

Clustering Complex Time Series Databases

Francesco Giordano, Michele La Rocca, and Maria Lucia Parrella

Abstract Time series data account for a large fraction of the data stored in financial, medical and scientific database. As a consequence, in the last decade there has been an explosion of interest in mining time series data and several new algorithms to index, classify, cluster and segment time series have been introduced. In this paper we focus on clustering of time series from a large database provided by a large Italian electric company, and the power consumption of a specific class of power users, that is the business and industrial customers, is measured. The aim of this paper is to propose an effective clustering technique in the frequency domain where the need of computational and memory resources is much reduced in order to make the algorithm efficient for large and complex temporal data bases.

1 Introduction

Time series analysis has been often associated with the discovery and use of patterns (such as periodicity, seasonality, or cycles), and prediction of future values. One key difference between traditional time series analysis and data mining on time series is the large number of series involved in temporal data mining. Due to the huge amount of data, highly automated analysis techniques become crucial in such applications and classical techniques, based on non-automatic interactive and iterative schemes, become soon impractical. Automatic model building requires both a adequate analysis of all pitfalls in data warehousing and an accurate data pre-processing involving (i) proper time series construction from observed raw data; (ii) automated outlier detection; (iii) right temporal aggregation. Those issues make temporal data mining an area of research that is at the intersection of several disciplines, including statistics, temporal pattern recognition, temporal databases, optimisation, high-performance computing, parallel computing and visualization.

Temporal data mining includes indexing, clustering, classification and segmentation. In this paper the focus will be on clustering. These techniques generally have

F. Giordano (✉)
Department of Economics and Statistics, University of Salerno, 84084 Fisciano, Salerno, Italy,
e-mail: giordano@unisa.it

B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-13312-1_44,

been developed along two main directions. The first is completely based on a parametric model, usually linear, for the time series (see Corduas and Piccolo, [3] inter alia). Given a proper metric, such as the AR-metric of Piccolo [5] which measures the EuclidEan distance between the coefficients of a stationary, gaussian AR(∞) process, a dissimilarity matrix is built and used for clustering. The second approach uses nonparametric techniques such as wavelets, bootstrap, and kernel methods. In this framework stands out the technique proposed in Alonso et al. [1], where the full probability density of the forecasts is estimated by using a resampling method combined with a nonparametric kernel estimator. A measure of discrepancy is then defined and the resulting dissimilarity matrix is used for clustering. In any case the approach, while being nonparametric in its spirit, estimates the forecast density by using the AR-sieve bootstrap which is consistent for linear processes only.

In this paper a novel clustering algorithm based on spectral techniques is proposed and used to cluster complex time series from a large database provided by an Italian electric company. The series refer to the consumption of electricity, measured at intervals of 15 min, of a particular class of users, the business and industrial firms. The database contains 65,245 time series for the year 2006 of length $365 \times 24 \times 4 = 35{,}040$, involving $2{,}286.1848 \times 10^6$ observations.

To motivate the proposed clustering scheme the plots for three different users of the analyzed database are reported (see Figs. 1 and 2). As main features we can distinguish: (i) *Cycles:* time series are generally characterized by periodic components (monthly, weakly, daily and intra-daily cycles); (ii) *Long memory:* high frequency observations raise the question on the fractional integration of the series; (iii) *Stationarity:* apart from the cycles, the plots show substantial stationarity.

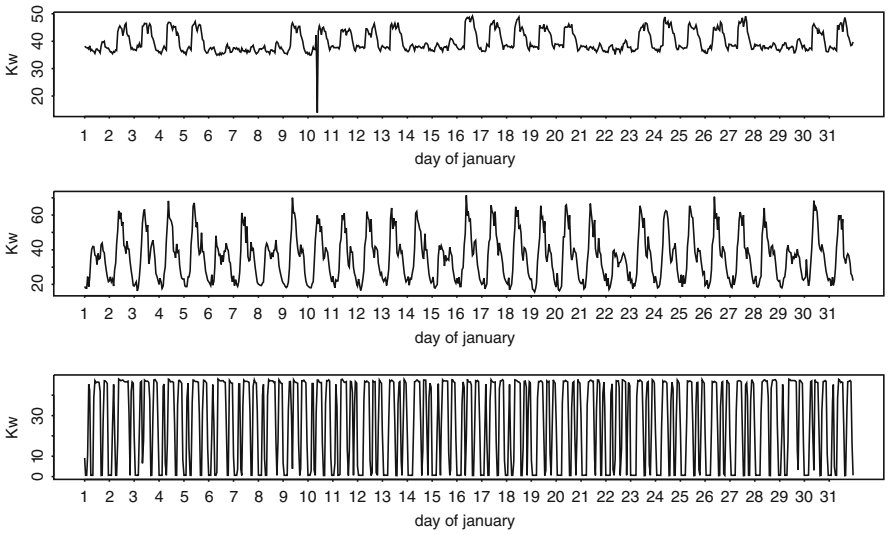


Fig. 1 The load curves for three different users, zooming on january 2006

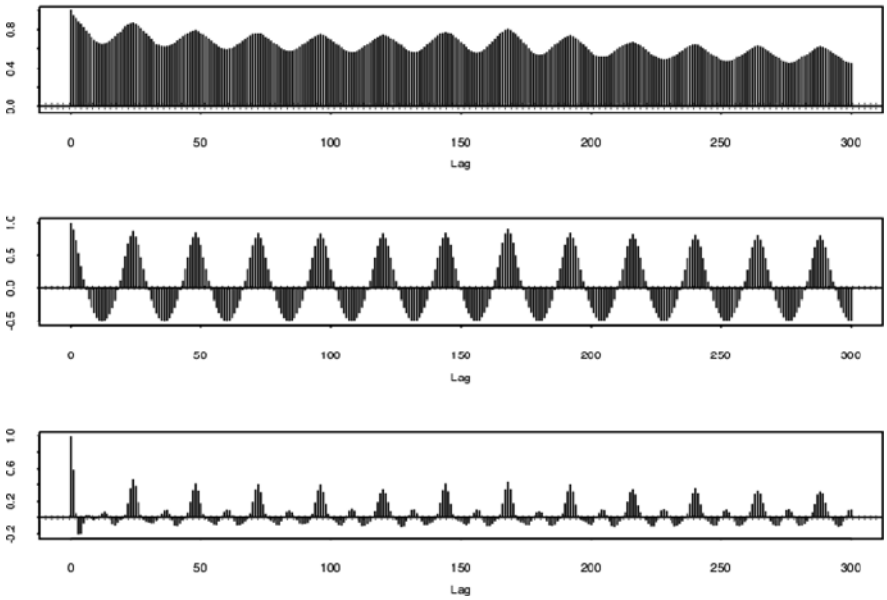


Fig. 2 Estimated ACFs of the series plotted in Fig. 1

The peculiarities highlighted above and the high dimension of the database make the analysis complex, and point towards highly automated procedures, based on a frequency domain analysis, which are able: (i) to efficiently detect the structure of the series and (ii) to reduce the computational burden and the amount of data to be stored.

The paper is organized as follows. In the next section the clustering algorithm is introduced and discussed. In Sect. 3 the results of the application to real data are reported along with some concluding remarks.

2 The Clustering Algorithm

The basic idea of our clustering scheme can be justified as follows. Denote with $\{X_t\}$ the data generating process. By using the Wold representation of a stationary process, we can write

$$X_t - \mu = V_t + Z_t, \tag{1}$$

where Z_t is a linear process, also known as the stochastic component, and V_t is an harmonic process, given by a combination of (say m) sinusoidal functions. In particular,

$$V_t = \sum_{j=1}^{m} A_j \cos(\omega_j t + \phi_j), \qquad 0 \leq \omega_j \leq \pi. \tag{2}$$

The spectrum of the process, denoted with $g_X(\omega)$, identifies the dominant frequencies ω, *i.e.* those explaining large portions of variation in the data. A stationary process as in (1) will possess a *mixed spectrum*, with a discrete component associated to the harmonic process V_t and a continuous component associated to the stochastic process Z_t (for a more detailed and technical definition see, for example, Priestley [6]). Intuitively, the cycles we observe in the energy consumption series are mainly connected with the component V_t. Given the uncorrelation of V_t and Z_t, the variance of the process is $Var(X_t) = Var(V_t) + Var(Z_t)$. We argue that a large portion of the variability in the data is due to the component V_t. Therefore, a clustering procedure for this kind of time series may be naturally based on the explanation of $Var(V_t)$. This is connected with the identification of the discrete component of the spectrum, which therefore become the main step of the clustering procedure.

The clustering procedure is based on the following steps:

1. For each time series (user), denoted with $u = 1, \ldots, T$, estimate the spectrum by using any consistent estimator (see, for example, Priestley, [6])

$$\hat{g}_X^u(\omega_j), \qquad 0 \leq \omega_j = \frac{2\pi j}{n} \leq \pi; \ j = 0, 1, \ldots, m; \ m = \left[\frac{n}{2}\right],$$

   where ω_j are the discrete frequencies and $[x]$ denote the integer part of x.
2. By (2), for each user $u = 1, .., T$, use a *Whittle test* with a Bartlett window to test the following hypotheses (for the details on the test, see Priestley [6]):

$$\begin{aligned} H_0 &: A_j = 0 \qquad j = 1, \ldots, m \\ H_1 &: A_j \neq 0 \qquad \text{at least for one } j. \end{aligned}$$

   Derive the *relevant discrete frequencies* for the user u, as those frequencies ω_j for which H_0 is rejected. Denote these frequencies with ω_j^u, $j = 1, 2, \ldots$.
3. For each user u and a fixed integer h_1, extract *the first most important (relevant) frequencies* $\omega_{(j)}^u$, for $j = 1, \ldots, h_1$, such that:

$$\hat{g}_X^u(\omega_{(1)}^u) \geq \hat{g}_X^u(\omega_{(2)}^u) \geq \ldots \geq \hat{g}_X^u(\omega_{(h_1)}^u),$$

   among which the first, that is $\omega_{(1)}^u$, is called the *dominant frequency* of the user u. For an easier interpretability, convert each frequency $\omega_{(i)}^u$ into the correspondent period $P_{u,i}$ (expressed in hours, days, weeks, etc...). Derive the matrix of the relevant periods $\mathbf{P}$ for all the users (see Fig. 3, on the left, and Remark 2).
4. By using only the first column of the matrix $\mathbf{P}$ as raw data, derive the distribution of the *dominant periods* P_i. Denote with δ_i the percentage of users which have P_i

$$\mathbf{P} = \begin{pmatrix} P_{1,1} & P_{1,2} & \dots & P_{1,h_1} \\ P_{2,1} & P_{2,2} & \dots & P_{2,h_1} \\ \vdots & \vdots & \ddots & \vdots \\ P_{u,1} & P_{u,2} & \dots & P_{u,h_1} \\ \vdots & \vdots & \ddots & \vdots \\ P_{T,1} & P_{T,2} & \dots & P_{T,h_1} \end{pmatrix};$$

Dominant Periods	% users (δ_i)
P_1	δ_1
P_2	δ_2
$\dots$	$\dots$
P_c	δ_c
$\dots$	$\dots$
P_r	δ_r
	100%

;

$$\mathbf{D} = \begin{pmatrix} 0 & 0 & \cdots & \cdots & 1 \\ 1 & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & d_{u,s} & \dots & d_{u,c} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \dots & d_{T,s} & \cdots & d_{T,c} \end{pmatrix}$$

Fig. 3 *On the left,* the matrix **P** of the relevant periods, the first column of which reports the *dominant periods* observed in the database. *In the center*, the distribution of the *dominant periods* observed in the database. *On the right*, the dissimilarity matrix **D**

as dominant period, *i.e.* for which $P_{u,1} = P_i$, for $u = 1, \dots, T$ and $i = 1, \dots, r$. Suppose that $\delta_1 \geq \delta_2 \geq \dots \geq \delta_r$ (see Fig. 3, in the center, and Remark 2).

5. Consider the first c *most observed* dominant periods, *i.e.* the periods P_i for which $\delta_i \geq \delta_c$, for a given c or a given threshold δ_c (see Remarks 1 and 2).
6. For a fixed integer $h_2 \leq h_1$, define the binary matrix **D**, whose generic element $d_{u,s}$ is equal to one if the period P_s appears in the first h_2 positions of the u-th row of matrix **P**, for $u = 1, \dots, T$ and $s = 1, \dots, c$; otherwise $d_{u,s}$ is equal to zero (see Fig. 3, on the right).
   The matrix **D** acts as a dissimilarity matrix. If two rows i and j of the matrix are equal, this means that the load curves of the users i and j are characterized by the presence of the same relevant periodic components, although each load curve might present some other less important periodic component (see also Remark 1).
7. By considering the different combinations of the relevant periods P_i, $i = 1, \dots, c$, derive the 2^c clusters of users by associating the rows (=users) of the matrix **D** with the same sequence of zeroes/ones (see Remark 3).

Remark 1 The parameters h_1, h_2 and c (or equivalently the threshold δ_c) behave like tuning parameters for the clustering procedure and must be fixed in some way. More investigations would be useful in order to determine the influence of such parameters on the results.

Remark 2 Identification of the discrete component from a mixed spectrum, that is identification of the "jumps" in the function $g_X(\omega)$, is particularly difficult when starting from the estimated spectrum. Difficulties arise since a smooth peak could be confounded with a jump (and *vice versa*) by simply modifying the smoothing parameter of the kernel estimator. For this reason, in order to enforce the faith in correctly identifying the discrete part of the spectrum, we consider in step 5 only the first c most observed dominant cycles, that is only the periods P_i with a *strong evidence* in the database.

Remark 3 By construction, the procedure can identify potentially 2^c clusters, but some of them could be empty. This happens when there is an exclusive

disjunction relationship between some of the periodic components, at least in the "relevant positions".

Remark 4 Nonstationarity and fractional integration of the process (long memory) would compromise the spectral estimations, in particular the Whittle test [2]. In order to verify these features on our data, we performed the nonparametric test of Lobato and Robinson [4]. Consider the following hypothesis

$$H_0 : d = 0 \quad \text{vs} \quad H_1 : d \neq 0,$$

where d is the integration parameter of the process. The percentage of rejections of H_0 in the resampling procedure proposed in Lobato and Robinson [4] did not exceed the nominal coverage error (3% against 5%). So the hypothesis H_0 could not be rejected for the analyzed database.

Remark 5 gaussianity of $\{X_t\}$ is generally desirable in spectral analysis. Anyway, dropping the normality condition has little effect on the large sample distributions [7].

3 An Application to a Real Temporal Data Base

The proposed clustering procedure is used to cluster the time series of electricity consumption of business and industrial firms as recorded by a large electric company. For each observed time series, we performed two different estimations of the spectral density. The first was based on the original observations in order to capture the intra-daily periodic components. The second estimation was aimed to identify weekly and monthly cycles and, to avoid masking effects due to high frequency components, time series have been aggregated into daily observations.

The Whittle test was performed in order to identify the relevant frequencies ω_j^u for each time series, as described in the step 2 of the procedure. Each relevant frequency was then converted into the correspondent period P, for an easier interpretation of the results. We selected the following smoothing parameters: (i) we considered $h_1 = 5$ relevant frequencies (or periods). We built the matrix $\mathbf{P}$ of order $64{,}522 \times 5$, as shown in Fig. 3. (ii) With the first column of matrix $\mathbf{P}$, we derived the distribution of the dominant periods and then we selected the first $c = 5$ most observed dominant periods P_i, for $i = 1, \ldots, 5$. (iii) We derived the binary matrix $\mathbf{D}$ by matching the presence of each dominant period P_i in the first $h_2 = 3$ relevant frequencies of each user, that is in the first $h_2 = 3$ columns of the matrix $\mathbf{P}$, for each user $u = 1, \ldots, T$. When collecting the results, we noted that only few frequencies were systematically observed. From the daily data, we identified three different weekly cycles, having period respectively of about 1, 2, and 3 weeks (respectively denoted as W_1, W_2 and W_3). From the hourly observations, we identified a daily cycle of about 24 h and an intra-daily cycle of about 4 h (respectively denoted as D_1 and D_2). The following table summarizes the results, reporting also the percentages of users on the total which present each period as dominant period (δ_i).

Dominant period	Approximate time	% of users on the total (δ_i)
D_1	4 hours	79.8%
D_2	1 day	15.3%
W_3	3 weeks	2.4%
W_2	2 weeks	0.5%
W_1	1 week	0.3%

Cluster	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}
Main Cycles	none	W_1	W_2	W_3	D_1	W_1 D_1	W_2 D_1	W_3 D_1	D_2	W_1 D_2	W_2 D_2	W_3 D_2	D_1 D_2	W_1 D_1 D_2	W_2 D_1 D_2	W_3 D_1 D_2
Users	2265	242	59	545	1880	1080	27	302	16200	603	173	1342	25978	10212	150	1602

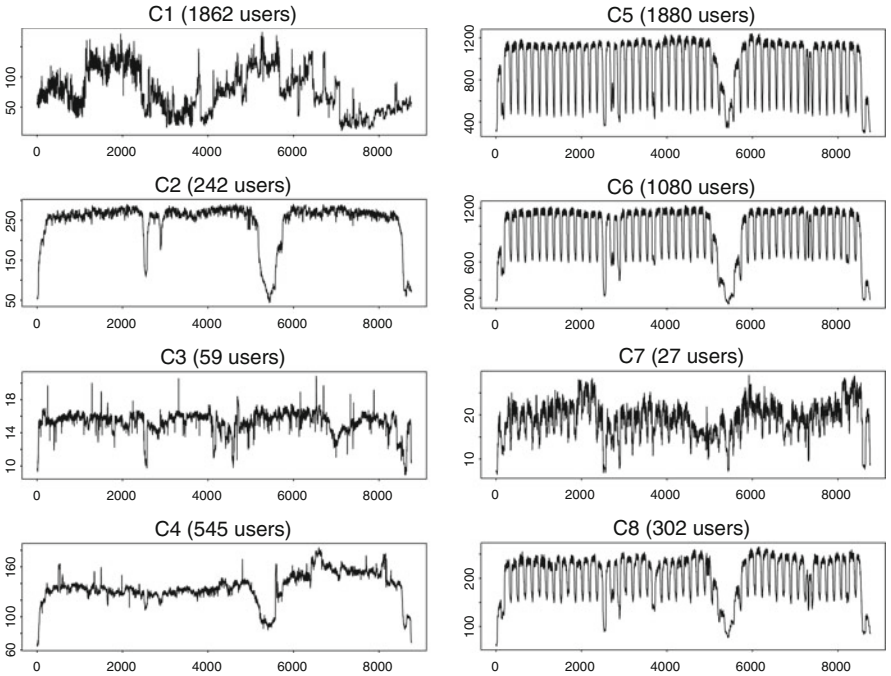


Fig. 4 The aggregated time series representing the identified clusters of users (from 1 to 8)

On the basis of such results, we derived the matrix **D** by assigning to each series a binary code, i.e. by constructing a vector of zeroes and ones depending on whether each frequency wasn't or was present in the three first positions of matrix **P**. Potentially we had to consider $2^5 = 32$ different codes or clusters, but we found that there is an exclusive disjunction relationship between some frequencies and this implies that some codes reported null dimension. Finally we observed $16 + 1$ relevant clusters (we add a cluster C_0 including the series which had more than

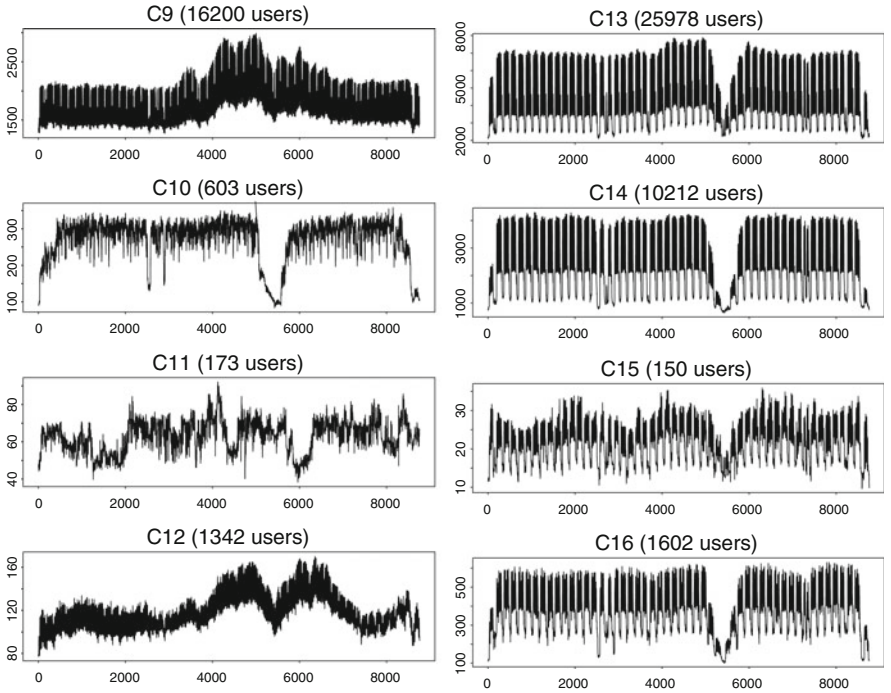


Fig. 5 The aggregated time series representing the identified clusters of users (from 9 to 16)

50% null observations, in order to exclude from the analysis the anomalous series). The following table reports the main characteristics and the dimensions of the 16 clusters.

In Figs. 4 and 5 we show the results. Each cluster is represented by the *global load path*, that is the global energy consumption series for all the users in the cluster. As desired, the 16 patterns shows different characteristics of the series, although each one is characterized by a substantial regular path, even when the cluster has a very high dimension (for example, the clusters C_{13} and C_{14}). This could be considered a very encouraging result. Even summing up thousands of users belonging to the same cluster, a clear periodic pattern is still evident from the plots.

References

1. Alonso, A.M., Berrendero, J.R., Hernández, A., Justel, A.: Time series clustering based on forecast densities. Comput. Stat. Data Anal. **51**, 762–776 (2006)
2. Beran, J., Ghosh, S.: Estimation of the dominating frequency for stationary and nonstationary fractional autoregressive models. J. Time Ser. Anal. **21**, 517–533 (2000)
3. Corduas, M., Piccolo, D.: Time series clustering and classification by the autoregressive metric. Comput. Stat. Data Anal. **52**, 1860–1872 (2008)

4. Lobato, I.N., Robinson, P.: A nonparametric test for I(0). Rev. Econ. Stud. **65**(3), 475–495, Blackwell Publishing (1998)
5. Piccolo, D.: A distance measure for classifying ARIMA models. J. Time Ser. Anal. **11**, 153–164 (1990)
6. Priestley, M.B.: Spectral Analysis and Time Series. Academic Press, London (1981)
7. Walker, A.M.: Some asymptotic results for the periodogram of a stationary time series. J. Aust. Math. Soc. **5**, 107–128 (1965)

Use of a Flexible Weight Matrix in a Local Spatial Statistic

Massimo Mucciardi

Abstract Most of local indices of spatial autocorrelation utilize a classical adjacency matrix as interconnection system. In this paper we attempt to use generalized matrix of spatial weights for measuring local autocorrelation. The work concludes with a comparison of local autocorrelation indices according to different hypotheses of neighborhood.

1 Introduction

The concept of local spatial autocorrelation is based on the idea of spatial outlier, that is a instability point in a spatial process underlying. In these terms, the local spatial autocorrelation expresses itself in the identification of the units characterized by extreme value of variable. These extreme values are identified through the comparison between and the value assumed by the process in the contiguity units [1]. As point out by Unwin A. and Unwin D. [8], the aim of local index is to learn more about each individual datum by relating it in some way to the value observed at neighbouring locations. This can be carried out by using the visualization of the resulting maps as a direct analytical procedure. Moreover, it is possible that within the same dataset, a different degree of spatial autocorrelation could be present; both positive and negative autocorrelation could even exist within the same dataset [2]. In this case, global measures of spatial autocorrelation would fail to pick up these different degree of spatial dependence within data. Consequently, a global statistic might misleadingly indicate that there is no spatial autocorrelation in a dataset, when there is a strong positive autocorrelation in one part of territory and negative autocorrelation in another. Whatever "local statistic" is used, there is a need to define a "local neighbourhood". Most of local spatial indices utilize a classical (0–1) matrix of weight as the interconnection system. In these paper we try to use generalized matrix of spatial weights [4] for measuring local autocorrelation. The

M. Mucciardi (✉)
Department D.E.S.Ma.S. "V. Pareto", University of Messina, Messina, Italy
e-mail: massimo.mucciardi@unime.it

B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-13312-1_45,

paper concludes with a comparison of local autocorrelation indices according to different hypotheses of neighbourhood.

2 Local Measure of Spatial Autocorrelation with S-DSMA Procedure

In a local index of spatial autocorrelation each unit is characterized by one value of the index; it gives the individual contribution of that location in the global spatial autocorrelation measured on all locations. Although there are available different indices in literature, in these paper we focus a "local Moran" statistic only. Local Moran's statistic for each observation may be defined as follows [1]:

$$I_i = \sum_{j,j\neq i}^{n} w_{ij} z_j \tag{1}$$

where the observation z_i and z_j are in standardized form and the weight w_{ij} are in row-standardized form. As observed earlier, the local index is the product of the standardized local value and the weighted mean of the standardized neighboring value. Thus, similarly to the global index, it can be positive, negative or equal to zero. It is negative when there is an association of opposite values at neighboring locations, and positive in the case of spatial association of similar values. At this point, use of binary 0–1 weight is attractive and computationally convenient [8], but there also are several other possible methods. A variety of approaches, from the use of simple 0–1 adjacency, through various measure of distance and length between the zones, have been experimented and is not possible to review them all here. More recently a "general weight matrix" has been proposed according the S-DSMA procedure [4]. We should briefly remember that the procedure determines different types of weight matrices with "threshold distance" h^k : (1) matrix Δ^k whose values δ_{ij} represent the interconnection between the territorial barycenters or centroids of the aerial units; (2) matrix E^k whose values represent ε_{ij} the interconnection between the territorial barycenters or centroids of the aerial units with weights sensitive to the effective distance of each unit; (3) matrix B^k whose values γ_{ij} represent the weights in function of the physical characteristics (surface) of the aerial units only; (4) matrix Ω^k whose values δ_{ij} are obtained introducing a suitable function (mean) on the coefficients obtained from the two matrices E^k and B^k. In this case, factors of "distance" and "surface" in aerial units are evaluated simultaneously. Lastly, a mixed matrix Ω_c^k is possible using it in the case we combine adjacency (0–1 matrix) and surface (B^k) between the units [7]. Therefore, the distinguishing feature of this approach is to obtain a "flexible weights matrix" in relation to the phenomenon under investigation. In the following section we are going to apply a local index with S-DSMA procedure for detecting spatial outlier and/or spatial clusters.

3 Application and Conclusion

The comparison between standard measure of local spatial autocorrelation and local measure obtained using our procedure is performed using no GIS package S-Joint (for more details see Appendix). This software allows us to consider the matrix of binary contiguity W^k (rook case) and the matrix Δ^k, E^k, B^k, Ω^k and Ω_c^k, according to hypothesis of neighborhood. In any case, the comparison will be made considering Ω^k only. The proposed application calculates the local spatial autocorrelation for Italian region localization rate relative to public administration sector [3]. As evidenced by the simultaneous analysis through local indices of autocorrelation (Table 1) and the Moran scatterplot (Fig. 1, the use of S-DSMA (hypothesis Ω) determines a different indices quantification). Consequently, Moran scatterplot seems to be less dispersive, indicating, in our opinion, more evidence of two clusters made respectively by northern regions (with low levels of specialization in the public administration sector) and southern regions (with high levels of specialization in the public administration sector). In conclusion, this paper has developed and applied a new technique for measuring local autocorrelation. Local measure of spatial autocorrelation with S-DSMA procedure, allow us to detect the regions with significant (positive or negative) deviations from the national average, and to determine the intensity of the interactions between neighbouring locations. Finally, we underline that the main disadvantage of traditional local statistics is their strong

Table 1 Comparison between $I_i(\mathrm{W})$ and $I_i(\Omega)$

Region	Z_i	$\sum_{j,j\neq i}^{n} w_{ij} z_j^{\mathrm{a}}$	$\sum_{j,j\neq i}^{n} \omega_{ij} z_j^{\mathrm{b}}$	$I_i(\mathrm{W})$	$I_i(\Omega)$
Piemonte	−1.248	−0.206	−0.603	0.257	0.752*
Valle d'Aosta	1.064	−1.248	−0.894	−1.328	−0.951*
Lombardia	−1.638	−0.969	−0.580	1.588*	0.950*
Trentino	0.194	−1.587	−0.818	−0.308	−0.159
Veneto	−1.535	−0.820	−0.589	1.260*	0.905*
Friuli	−0.549	−1.535	−0.886	0.842	0.486
Liguria	−0.044	−1.132	−0.698	0.050	0.031
Emilia	−1.289	−1.002	−0.449	1.291*	0.578*
Toscana	−0.859	−0.243	−0.462	0.209	0.397**
Umbria	−0.247	−0.166	−0.337	0.041	0.083
Marche	−0.932	−0.245	−0.246	0.229	0.230
Lazio	1.295	0.023	−0.061	0.029	−0.079
Abruzzo	−0.126	0.476	0.067	−0.060	−0.008
Molise	1.064	0.375	0.185	0.399	0.197
Campania	0.438	0.772	0.516	0.338	0.226
Puglia	−0.108	0.764	0.794	−0.082	−0.086
Basilicata	0.790	0.547	0.736	0.432	0.581**
Calabria	1.311	1.070	0.732	1.402*	0.960*
Sicilia	1.599	1.065	0.871	1.703*	1.392*
Sardegna	0.820	1.161	1.295	0.951**	1.062

*$p < 0.05$; **$p < 0.01$.

[a] 1st spatial order, 35 joints.

[b] 1st spatial order – h-distance (379 km), 82 joints.

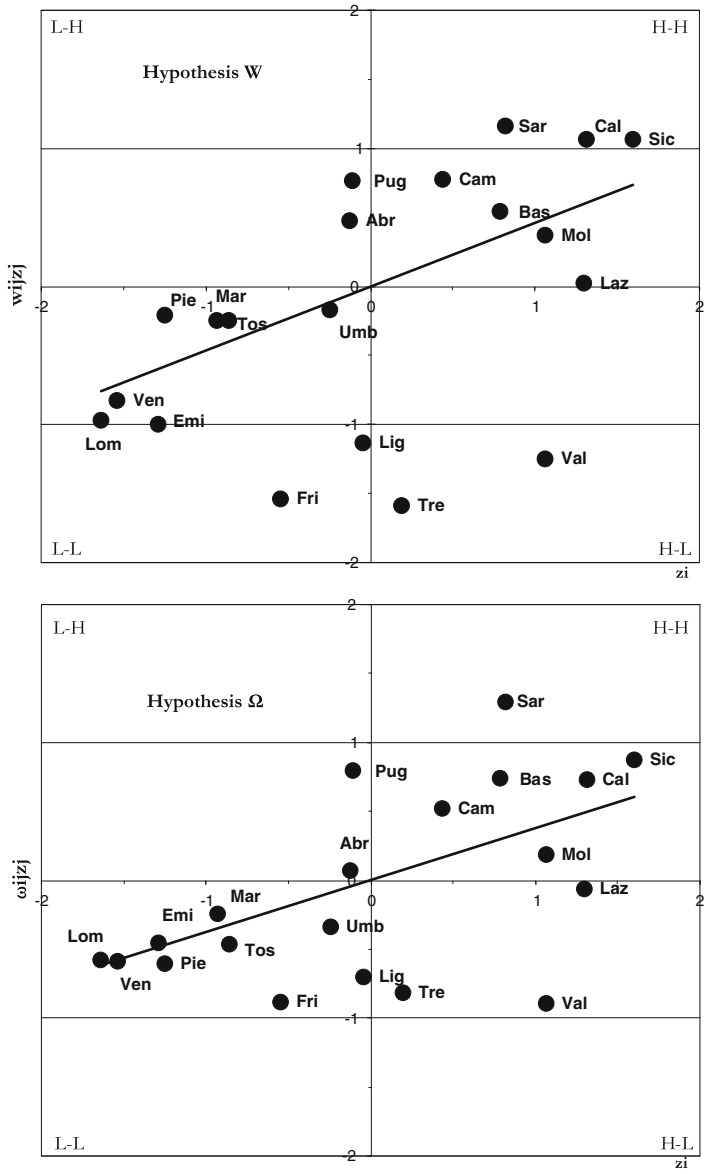


Fig. 1 Dynamic Moran scatterplot for standardized Italian region localization rate: hypothesis W and hypothesis Ω. (H-H indicates high index point with high index neighbors; L-L indicates low index point low index neighbors; H-L indicates high index point with low index neighbors; L-H indicates low index point high index neighbors; the slope of the regression line through the points is proportional to the global Moran's I for the dataset)

Table 2 Output of S-Joint software – matrix W (row-standardized form) for the Italian regions – 1st spatial order, joints 35

Region	PIE	VAL	LOM	TRE	VEN	FRI	LIG	EMI	TOS	UMB	MAR	LAZ	ABR	MOL	CAM	PUG	BAS	CAL	SIC	SAR	Total
PIE		0.33	0.33				0.33														1
VAL	1																				1
LOM	0.25			0.25	0.25			0.25													1
TRE			0.5		0.5																1
VEN			0.25	0.25		0.25		0.25													1
FRI					1																1
LIG	0.33							0.33	0.33												1
EMI			0.2		0.2		0.2		0.2		0.2										1
TOS							0.2	0.2		0.2	0.2	0.2									1
UMB									0.33		0.33	0.33									1
MAR								0.2	0.2	0.2		0.2	0.2								1
LAZ									0.14	0.14	0.14		0.14	0.14	0.14					0.14	1
ABR											0.33	0.33		0.33							1
MOL												0.25	0.25		0.25	0.25					1
CAM												0.2		0.2		0.2	0.2			0.2	1
PUG														0.33	0.33		0.33				1
BAS															0.33	0.33		0.33			1
CAL																	0.33		0.33	0.33	1
SIC																		0.5		0.5	1
SAR												0.25			0.25			0.25	0.25		1

Table 3 Output of S-Joint software – Matrix Ω (row-standardized form) for the Italian regions – 1st spatial order, h-distance (379 km), joints 82

Region	PIE	VAL	LOM	TRE	VEN	FRI	LIG	EMI	TOS	UMB	MAR	LAZ	ABR	MOL	CAM	PUG	BAS	CAL	SIC	SAR	Total
PIE		0.15	0.16	0.13	0.13		0.15	0.15	0.14												1
VAL	0.18		0.17	0.16	0.15		0.19	0.15													1
LOM	0.11	0.09		0.11	0.11	0.09	0.1	0.11	0.1	0.08	0.08										1
TRE	0.1	0.1	0.12		0.13	0.13	0.1	0.11	0.1		0.1										1
VEN	0.09	0.08	0.11	0.12		0.11	0.09	0.11	0.1	0.09	0.1										1
FRI			0.13	0.16	0.16			0.14	0.12	0.14	0.14										1
LIG	0.12	0.13	0.11	0.11	0.1			0.11	0.11	0.11	0.1										1
EMI	0.09	0.07	0.09	0.08	0.09	0.08	0.08		0.1	0.08	0.08	0.08	0.07								1
TOS	0.08		0.09	0.08	0.09	0.07	0.08	0.1		0.09	0.09	0.09	0.08	0.07							1
UMB			0.07		0.08	0.08	0.08	0.09	0.1		0.11	0.1	0.1	0.09	0.09						1
MAR			0.07	0.08	0.08	0.08	0.07	0.08	0.09	0.1		0.09	0.1	0.09	0.08						1
LAZ								0.09	0.1	0.11	0.11		0.11	0.1	0.1	0.09	0.09			0.09	1
ABR								0.09	0.1	0.12	0.12	0.12		0.12	0.12	0.1	0.11				1
MOL									0.09	0.11	0.11	0.11	0.13		0.12	0.11	0.12	0.1			1
CAM										0.1	0.1	0.11	0.12	0.12		0.12	0.12	0.11	0.09		1
PUG												0.15	0.16	0.16	0.18		0.18	0.17			1
BAS												0.13	0.14	0.15	0.16	0.16		0.15	0.11		1
CAL														0.18	0.21	0.21	0.22		0.19		1
SIC															0.32		0.31	0.37			1
SAR												1									1

dependence on the definition of neighbourhood. Therefore, different hypothesis of neighbourhood would be proposed in order to reach a comprehensive analysis of the data.

Appendix: S-Joint Software

S-Joint is a multi-document interface program realized in C++ language that offers several spatial analysis operations in a simple and intuitive way [5, 6]. The current version of this program has been planned to run under Microsoft Windows. The graphic interface was built on the QT library (ver. 3.3.2) by Trolltech. Other libraries used were: OpenGL (3D graphic library), LibQGLViewer (QT OpenGL support library), QWT (graphics extension to the Qt GUI application framework), GDAL and OGR (geospatial data abstraction library). S-Joint can open data files, vector maps (point, line and polygonal maps) and raster maps. It is also possible to create regular grids. The vector maps and the grids are shown in three-dimensional windows in order to give greater flexibility in operations such as, for instance, rotations, translations and zooming. Point and polygonal maps, data files and grids can be used to do spatial analysis studies. Moreover the ESRI shapefile (.shp) format, the MapInfo (.mif), the common separated value (.csv), the ARC/INFO coverage (.adf), the USGS SDTS, the dBase (.dbf) have been implemented. In the current development phase, it is also possible to calculate the traditional spatial autocorrelation index (Moran and Geary), to join-count statistics and the local Moran, to show the Moran scatterplot and to calculate new autocorrelation measures based on a "generalized weight system". S-Joint implements the S-DSMA procedure with a local reweighting [4] in relation to weights system (see Tables 2 and 3). It is also possible to choose other systems of weights depending on the map topology (polygonal maps will have different methods of weights creation compared to point maps). Furthermore, each of these systems provides us with a vast number of options which add some more flexibility in the matrix generation. Among the other S-Joint characteristics, we can take into account the layer management, weight matrix importing/exporting to different formats capability, the possibility of selecting/deselecting the vector maps in order to do local analysis. Another important add-on feature is the program ability to recognize if a map is using a geographical or a projected coordinate system, in order to calculate distances between points and polygonal map areas in different ways. (For more details visit URL: http://ww2.unime.it/scistat/homepages/mucciardi/down/manuale.pdf)

References

1. Anselin, L.: Local indicator of spatial association. Geogr. Anal. **27**, 93–115 (1995)
2. Fotheringham, A.S., Brunsdon, C., Chatlton, M.: Geographically Weight Regression. Wiley, New York, NY (2002)

3. La Rocca, A.: Analisi della Struttura Settoriale dell'Occupazione Regionale: 8° Censimento dell'industria e dei servizi (Contributi Istat). **27** (2004)
4. La Tona, L., Mazza A., Mucciardi M.: A generalized weight matrix for autocorrelated superficial data. In: Cafarelli, B, Lasinio, G.J., Pollice, A. (eds.) Spatial Data Methods for Environmental and Ecological Processes. Wip edizioni, Foggia (2006)
5. Mucciardi M., Bertuccelli, P.: S-joint: a new software for the analysis of spatial data. In: Proceedings of the Intermediate Conference of the Italian Statistical Society, Venice, Italy (2007)
6. Mucciardi, M., Bertuccelli, P.: Exploratory spatial data analysis with s-joint. In: Proceedings of XLIV Scientific Meeting of the Italian Statistical Society, Padova, Italy (2008)
7. Mucciardi, M.: Modelli di Vicinato Combinati per Dati Spaziali Areali. In: Proceedings of 6th National Congress of the Italian Biometric Society, Pisa, Italy, pp. 79–82 (2007)
8. Unwin, A., Unwin, D.: Exploratory spatial data analysis with local statistics. Statistician **3,** 415–421 (1998)

Constrained Variable Clustering and the Best Basis Problem in Functional Data Analysis

Fabrice Rossi and Yves Lechevallier

Abstract Functional data analysis involves data described by regular functions rather than by a finite number of real valued variables. While some robust data analysis methods can be applied directly to the very high dimensional vectors obtained from a fine grid sampling of functional data, all methods benefit from a prior simplification of the functions that reduces the redundancy induced by the regularity. In this paper we propose to use a clustering approach that targets variables rather than individual to design a piecewise constant representation of a set of functions. The contiguity constraint induced by the functional nature of the variables allows a polynomial complexity algorithm to give the optimal solution.

1 Introduction

Functional data [13] appear in applications in which objects to analyse display some form of variability. In spectrometry, for instance, samples are described by spectra: each spectrum is a mapping from wavelengths to e.g., transmittance.[1] Time varying objects offer a more general example: when the characteristics of objects evolve through time, a loss free representation consists in describing these characteristics as functions that map time to real values.

In practice, functional data are given as high dimensional vectors (e.g., more than 100 variables) obtained by sampling the functions on a fine grid. For smooth functions (for instance in near infrared spectroscopy), this scheme leads to highly correlated variables. While many data analysis methods can be made robust to this type of problem (see, e.g., [6] for discriminant analysis), all methods benefit from a compression of the data [12] in which relevant and yet easy to interpret features are extracted from the raw functional data.

F. Rossi (✉)
Institut Télécom, Télécom ParisTech, LTCI – UMR CNRS 5141, 75013 Paris, France
e-mail: Fabrice.Rossi@telecom-paristech.fr

[1] In spectrometry, transmittance is the fraction of incident light at a specified wavelength that passes through a sample.

B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-13312-1_46,

There are well-known standard ways of extracting optimal features according to a given criterion. For instance in unsupervised problems, the first k principal components of a dataset give the best linear approximation of the original data in $\mathbb{R}^k$ for the quadratic norm (see [13] for functional principal component analysis (PCA)). In regression problems, the partial least-squares approach extracts features with maximal correlation with a target variable (see also Sliced Inversion Regression methods [4]). The main drawback of those approaches is that they extract features that are not easy to interpret: while the link between the original features and the new ones is linear, it is seldom sparse; an extracted feature generally depends on many original features.

A different line of thoughts is followed in the present paper: the goal is to extract features that are easy to interpret in terms of the original variables. This is done by approximating the original functions by piecewise constant functions. We first recall in Sect. 2 the best basis problem in the context of functional data approximation. Section 3 shows how the problem can be recast in term of a constrained clustering problem for which efficient solutions are available.

2 Best Basis for Functional Data

Let us consider n functional data, $(s_i)_{1\leq i\leq n}$. Each s_i is a function from $[a, b]$ to $\mathbb{R}$, where $[a, b]$ is a fixed interval common to all functions (more precisely, s_i belongs to $L^2([a, b])$, the set of square integrable functions on $[a, b]$). In terms of functional data, linear feature extraction consists in choosing for each feature a linear operator from $L^2([a, b])$ to $\mathbb{R}$. Equivalently, one can choose a function ϕ from $L^2([a, b])$ and compute $\langle s_i, \phi\rangle_{L^2} = \int_a^b \phi(x)s_i(x)dx$. In an unsupervised context, using e.g., a quadratic error measure, choosing the k best features consists in finding k orthonormal functions $(\phi_i)_{1\leq i\leq k}$ that minimise the following quantity:

$$\sum_{i=1}^{n} \left\| s_i - \sum_{j=1}^{k} \langle s_i, \phi_k\rangle_{L^2}\phi_k \right\|_{L^2}^2 . \tag{1}$$

The $(\phi_i)_{1\leq i\leq k}$ form an orthonormal basis of the subspace that they span: the optimal set of such functions is therefore called the *best basis* for the original set of functions $(s_i)_{1\leq i\leq n}$.

If the ϕ_k are unconstrained, the best basis is given by functional PCA [13]. However, in order for the corresponding feature to be easy to interpret, the ϕ_k should have compact supports, the simple case of $\phi_k = \mathbb{I}_{[u_k,v_k]}$ being the easiest to analyse ($\mathbb{I}_{[u,v]}(x) = 1$ when $x \in [u, v]$ and 0 elsewhere).

The problem of choosing an optimal basis among a set of bases has been studied for some time in the wavelet community [3, 15]. In unsupervised context, the best basis is obtained by minimizing the entropy of the features (i.e., of the coordinates of the functions on the basis) in order to enable compression by discarding the

less important features. Following [12, 14] proposes a different approach, based on B-splines: a leave-one-out version of Eq. (1) is used to select the best B-splines basis. While the orthonormal basis induced by the B-splines does not correspond to compactly supported functions, the dependency between a new feature and the original ones is still localized enough to allow easy interpretation. Nevertheless both approaches have some drawbacks. Wavelet based methods lead to compactly supported basis functions but the basis has to be chosen in a tree structured set of bases. As a consequence, the support of a basis function cannot be any sub-interval of $[a, b]$. The B-spline approach suffers from a similar problem: the approximate supports have all the same lengths leading either to a poor representation of some local details or to a large number of basis functions.

3 Best Basis via Constrained Clustering

3.1 From Best Basis to Constrained Clustering

The goal of the present paper is to select an optimal basis using only basis functions of the form $\mathbb{I}_{(u,v)}$, without restriction on the possible intervals among sub-interval of $[a, b]$.[2] Let us consider $\left(\phi_j = \frac{1}{v_j - u_j}\mathbb{I}_{(u_j, v_j)}\right)_{1 \le j \le k}$ such an orthonormal basis. We assume that the $((u_j, v_j))_{1 \le j \le k}$ form a partition of $[a, b]$. Obviously, we have $\langle \phi_j, s_i \rangle = \frac{1}{v_j - u_j} \int_{u_j}^{v_j} s_i(x)dx$, i.e., the feature corresponding to ϕ_j is the mean value of s_i on $[u_j, v_j]$. In other words, $\sum_{j=1}^{k} \langle s_i, \phi_k \rangle_{L^2} \phi_k$ is a piecewise constant approximation of s_i (which is optimal according to the L^2 norm).

In practice, functional data are sampled on a fine grid with support points $a \le t_1 < \ldots < t_m \le b$, i.e., rather than observing the functions $(s_i)_{1 \le i \le n}$, one gets the vectors $(s_i(t_l))_{1 \le i \le n, 1 \le l \le m}$ from $\mathbb{R}^m$. Then $\langle \phi_j, s_i \rangle$ can be approximated by $\frac{1}{|I_j|} \sum_{l \in I_j} s_i(t_l)$ where I_j is the subset of indexes $\{1, ..., m\}$ such that $t_l \in (u_j, v_j) \Leftrightarrow l \in I_j$. Any partition of $((u_j, v_j))_{1 \le j \le k}$ of $[a, b]$ corresponds to a partition of $\{1, ..., m\}$ in k subsets $(I_j)_{1 \le j \le k}$ that satisfies an ordering constraint: if r and s belong to I_j then any integer $t \in [r, s]$ belongs also to I_j. Finding the best basis means for instance minimizing the sum of squared errors given by Eq. (1) which can be approximated as follows

$$\sum_{i=1}^{n} \sum_{j=1}^{k} \sum_{l \in I_j} \left(s_i(t_l) - \frac{1}{|I_j|} \sum_{u \in I_j} s_i(t_u) \right)^2 = \sum_{j=1}^{k} Q(I_j), \tag{2}$$

where

[2] The notations (u, v) is used to include all the possible cases of open and close boundaries for the considered intervals.

$$Q(I) = \sum_{i=1}^{n} \sum_{l \in I} \left(s_i(t_l) - \frac{1}{|I|} \sum_{u \in I} s_i(t_u) \right)^2 \quad (3)$$

The second version of the error shows that it corresponds to an additive quality measure of the partition of $\{1, ..., m\}$ induced by the $(I_j)_{1 \leq j \leq k}$. Therefore, finding the best basis for the sampled functions is equivalent to finding an optimal partition of $\{1, ..., m\}$ with some ordering constraints and according to an additive cost function. A suboptimal solution to this problem, based on an ascending (agglomerative) hierarchical clustering, is proposed in [9].

3.2 Dynamic Programming

However, an optimal solution can be reached in a reasonable amount of time, as pointed out in [10]: when the quality criterion of a partition is additive and when a total ordering constraint is enforced, a dynamic programming approach leads to the optimal solution (this is a generalization of the algorithm proposed by Bellman for a single function in [2, 16]; see also [1, 8] for rediscoveries/extensions of this early work). The algorithm is simple and proceeds iteratively by computing $F(j, k)$ as the value of the quality measure (from Eq. (2)) of the best partition in k classes of $\{j, ..., m\}$:

1. initialization: set $F(j, 1)$ to $Q(\{j, \ldots, m\})$ for all j
2. iterate from $p = 2$ to k:
   a. for all $1 \leq j \leq m - p + 1$ compute

$$F(j, p) = \min_{j \leq l \leq m-p+1} Q(\{j, \ldots, l\}) + F(l + 1, p - 1)$$

The minimizing index $l = l(j, p)$ is kept for all j and p. This allows to reconstruct the best partition by backtracking from $F(1, k)$: the first class of the partition is $\{1, \ldots, l(1, k)\}$, the second $\{l(1, k) + 1, \ldots, l(l(1, k) + 1, k - 1)\}$, etc. A similar algorithm was used to find an optimal approximation of a single function in [2, 11]. Another related work is [7] which provides simultaneously a functional clustering and a piecewise constant approximation of the prototype functions.

The internal loop runs $O(km^2)$ times. It uses the values $Q(\{j, \ldots, l\})$ for all $j \leq l$. Those quantities can be computed prior to the search for the optimal partition, using for instance a recursive variance computation formula, leading to a cost in $O(nm^2)$. More precisely, we are interested in

$$Q_{i,j,l} = \sum_{r=j}^{l} (s_i(t_r) - M_{i,j,l})^2, \quad (4)$$

where

$$M_{i,j,l} = \frac{1}{l-j+1} \sum_{u=j}^{l} s_i(t_u). \tag{5}$$

For a fixed function s_i, the $M_{i,j,l}$ and $Q_{i,j,l}$ are computed and stored in two $m \times m$ arrays, according to the following algorithm:

1. initialisation: set $M_{i,j,j} = s_i(t_j)$ and $Q_{i,j,j} = 0$ for all $j \in \{1, \ldots, m\}$
2. compute $M_{i,1,j}$ and $Q_{i,1,j}$ for $j > 1$ recursively with:

$$M_{i,1,j} = \frac{1}{j}\left((j-1)M_{i,1,j-1} + s_i(t_j)\right)$$
$$Q_{i,1,j} = Q_{i,1,j-1} + \frac{j}{j-1}(s_i(t_j) - M_{i,1,j})^2$$

3. compute $M_{i,j,l}$ and $Q_{i,j,l}$ for $l > j > 1$ recursively with:

$$M_{i,j,l} = \frac{1}{l-j+1}\left((l-j+2)M_{i,j-1,l} - s_i(t_{j-1})\right)$$
$$Q_{i,j,l} = Q_{i,j-1,l} - \frac{l-j+1}{l-j+2}(s_i(t_{j-1}) - M_{i,j,l})^2$$

This algorithm is applied to each function leading to a total cost of $O(nm^2)$ with a $O(m^2)$ storage. The full algorithm has therefore a complexity of $O((n+k)m^2)$.

3.3 Extensions

As pointed out in [10], the previous scheme can be used for any additive quality measure. It is therefore possible to use e.g., a piecewise linear approximation of the functions on a sub-interval rather than a constant approximation (this is the original problem studied in [2] for a single function). However, additivity is a stringent restriction. In the case of a piecewise linear approximation for instance, it prevents the introduction of continuity conditions: if one searches for the best continuous piecewise linear approximation of a function, then the optimized criterion is no more additive (this is in fact the case for all spline smoothing approaches expect the piecewise constant ones).

In addition, for the general case of an arbitrary quality measure Q there might be no recursive formula for evaluating Q. In this case, the cost of computing the needed quantities might exceed $O(nm^2)$ and reach $O(nm^3)$ or more, depending on the exact definition of Q.

That said, the particular case of leave-one-out is quite interesting. Indeed when the studied functions are noisy, it is important to rely on a good estimate of the

approximation error to avoid overfitting the best basis to the noise. It is straightforward to show that the leave-one-out (l.o.o.) estimate of the total error from Eq. (2) is given by

$$\sum_{i=1}^{n}\sum_{j=1}^{k}\sum_{l\in I_j}\left(\frac{|I_j|}{|I_j|-1}\right)^2\left(s_i(t_l)-\frac{1}{|I_j|}\sum_{u\in I_j}s_i(t_u)\right)^2, \tag{6}$$

when l.o.o. is done on the sampling points of the functions. This is an additive quality measure which can be computed using from the $Q_{i,j,l}$, that is in an efficient recursive way. As shown above, the piecewise constant approximation with k segments is obtained via the computation of the best approximation for all l in $\{1,\ldots,k\}$. It is then possible to choose the best l based on the leave-one-out error estimate at the same cost as the one needed to compute the best approximation for the maximal value of l. This leads to two variants of the algorithm. In the first one, the standard algorithm is applied to compute all the best bases and the best number of segments is chosen via the l.o.o. error estimate (which can be readily computed once the best basis is known). In the second one, we compute the best basis directly according to the l.o.o. error estimate, leveraging its additive structure. It is expected that this second solution will perform better in practice, as it constrains the best basis to be reasonable (see Sect. 4 for an experimental validation). For instance, it will never select an interval with only one point whereas this could be the case for the standard solution. As a consequence, the standard solution will likely produce bases with rather bad leave-one-out performances and tend to select a too small number of segments (see Sect. 4 for an example of this behavior).

4 Experiments

We illustrate the algorithm on the Wine dataset[3] which consists in 124 spectra of wine samples recorded in the mid infrared range at 256 different wavenumbers[4] between 4,000 and 400 cm^{-1}. Spectra number 34, 35 and 84 of the learning set of the original dataset have been removed as they are outliers. As shown on Fig. 1 the function approximation problem is interesting as the smoothness of the spectrum varies along the spectral range and an optimal basis will obviously not consist in functions with supports of equal size. Figure 2 shows an example of the best basis obtained by the proposed approach for $k = 16$ clusters, while Fig. 3 gives the suboptimal solution obtained by a basis with equal length intervals (as used in [14]). The uniform length approach is clearly unable to pick up details such as the peak on the right of the spectra. The total approximation error (Eq. (2)) is reduced from 62.66

[3] This dataset is provided by Prof. Marc Meurens, Université catholique de Louvain, BNUT unit, and available at http://www.ucl.ac.be/mlg/index.php?page=DataBases.

[4] The wavenumber is the inverse of the wavelength.

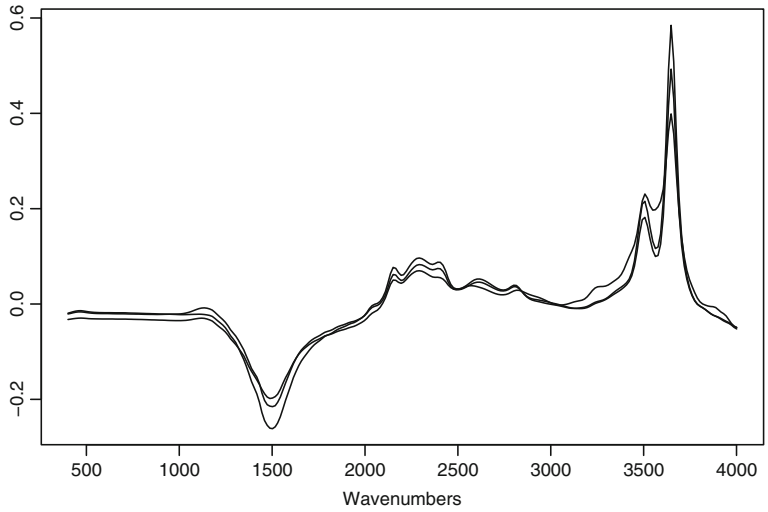


Fig. 1 Three spectra from the Wine dataset

with the uniform approach to 7.74 with the optimal solution. On the same dataset, the greedy ascending hierarchical clustering approach proposed in [9] reaches a total error of 8.55 for a similar running time of the optimal approach proposed in the present paper.

To test the leave-one-out approach, we have first added a Gaussian noise with 0.04 standard deviation (the functions take values in $[-0.265, 0.581]$). Then we look for the best basis up to 64 segments. As expected, the total approximation error

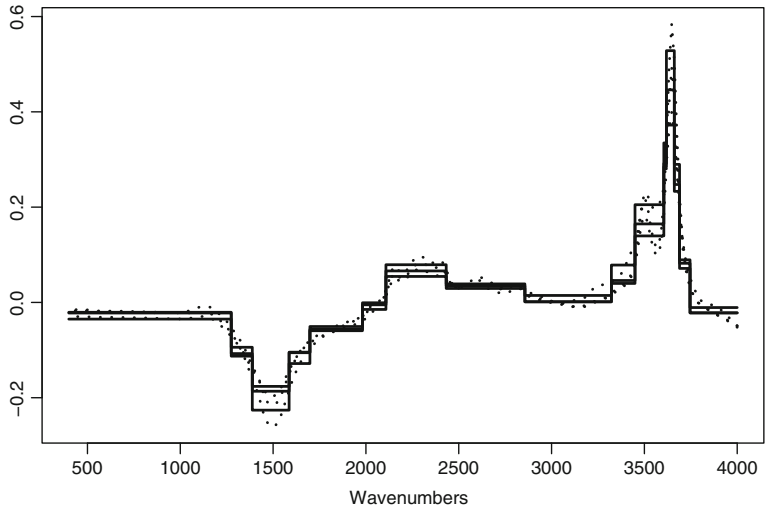


Fig. 2 Example of the optimal approximation results for 16 clusters on the Wine dataset

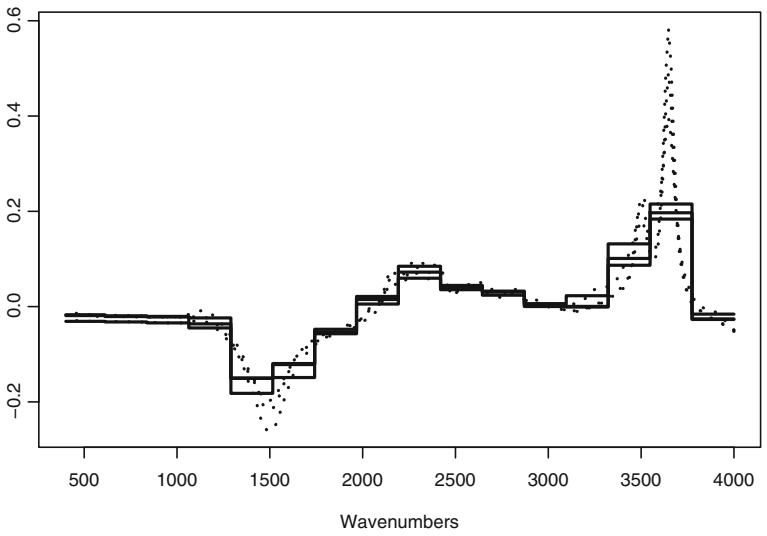


Fig. 3 Example of the uniform approximation results for 16 clusters on the Wine dataset

decreases with the number of segments and would therefore lead to a best basis with 64 segments. Moreover, as explained in the previous Section, the bases are not controlled by a l.o.o. error estimate. As a consequence, the optimization leads very quickly to basis with very small segments (starting at $k = 12$, there is at least one segment with only one sample point in it). Therefore, the l.o.o. error estimate applied to this set of bases selects a quite low number of segments, namely $k = 11$. When the bases are optimized according to the l.o.o. error estimate, the behavior is more smooth in the sense that small segments are always avoided. The minimum value of the l.o.o. estimate leads to the selection of $k = 20$ segments.

Table 1 summarizes the results by displaying the total approximation error on the noisy spectra and the total approximation error on the original spectra (the ground truth) for the three alternatives. The full l.o.o. approach leads clearly to the best results, as illustrated on Figs. 4 and 5.

Those experiments show that the proposed approach is flexible and provides an efficient way to get an optimal basis for a set of functional data. We are currently investigating supervised extensions of the approach following principles from [5].

Table 1 Total squared errors for the Wine dataset with noise

Basis	Noisy data	Real spectra
$k = 64$ (standard approach)	37.28	14.35
$k = 11$ (l.o.o. after the standard approach)	63.19	17.35
$k = 20$ (full l.o.o.)	54.07	12.07

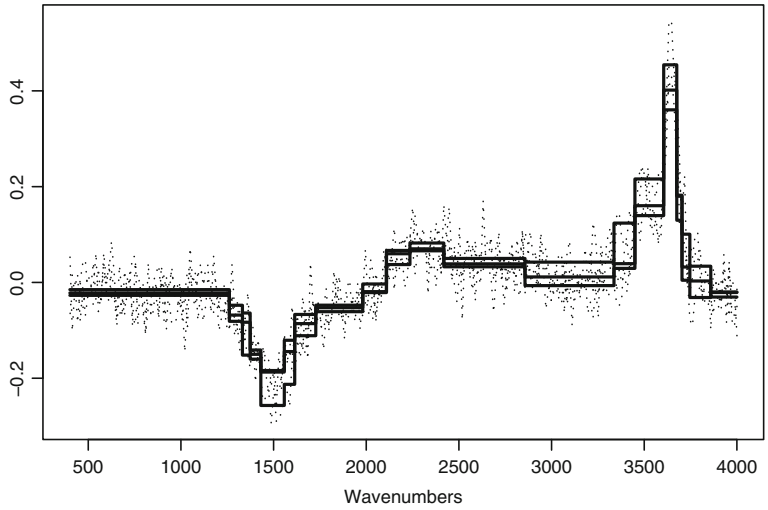


Fig. 4 Best basis selected by leave-one-out with the standard approach combined with loo

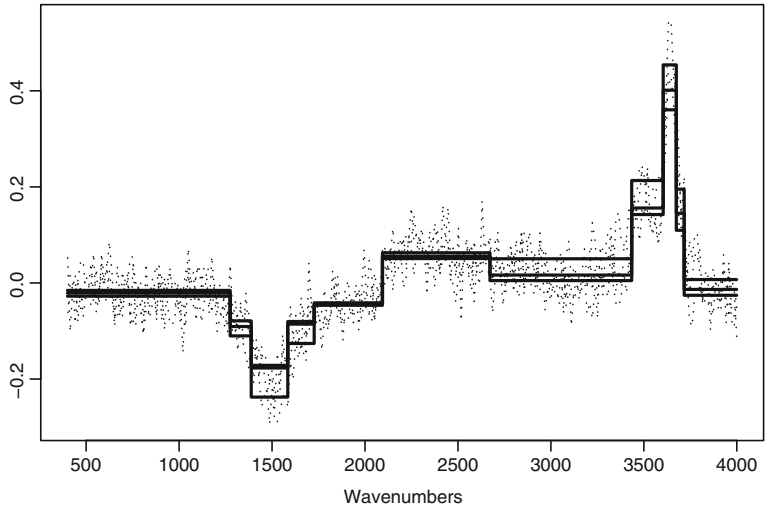


Fig. 5 Best basis selected by leave-one-out with the full loo approach

Acknowledgments The authors thank the anonymous reviewer for the detailed and constructive comments that have significantly improved this paper.

References

1. Auger, I.E., Lawrence, C.E.: Algorithms for the optimal identification of segment neighborhoods. Bull. Math. Biol. **51**(1), 39–54 (1989)

2. Bellman, R.: On the approximation of curves by line segments using dynamic programming. Commun. ACM 4(6), 284 (1961). DOI http://doi.acm.org/10.1145/366573.366611
3. Coifman, R.R., Wickerhauser, M.V.: Entropy-based algorithms for best basis selection. IEEE Trans. Inf. Theory **38**(2), 713–718 (1992)
4. Ferré, L., Yao, A.F.: Functional sliced inverse regression analysis. Statistics **37**(6), 475–488 (2003)
5. François, D., Krier, C., Rossi, F., Verleysen, M.: Estimation de redondance pour le clustering de variables spectrales. In: Actes des 10èmes journés Européennes Agro-industrie et Méthodes statistiques (Agrostat 2008), pp. 55–61. Louvain-la-Neuve, Belgique (2008)
6. Hastie, T., Buja, A., Tibshirani, R.: Penalized discriminant analysis. Ann. Stat. **23**, 73–102 (1995)
7. Hugueney, B., Hébrail, G., Lechevallier, Y.: Réduction de séries temporelles par classification et segmentation. In: Actes des 38ièmes Journées de Statistique de la SFDS. Clamart, France (2006)
8. Jackson, B., Scargle, J., Barnes, D., Arabhi, S., Alt, A., Gioumousis, P., Gwin, E., Sangtrakulcharoen, P., Tan, L., Tsai, T.T.: An algorithm for optimal partitioning of data on an interval. IEEE Signal Process. Lett. **12**(2), 105–108 (2005)
9. Krier, C., Rossi, F., Franois, D., Verleysen, M.: A data-driven functional projection approach for the selection of feature ranges in spectra with ICA or cluster analysis. Chemom. Intell. Lab. Syst. **91**(1), 43–53 (2008)
10. Lechevallier, Y.: Classification automatique optimale sous contrainte d'ordre total. Rapport de recherche 200, IRIA (1976)
11. Lechevallier, Y.: Recherche d'une partition optimale sous contrainte d'ordre total. Rapport de recherche RR-1247, INRIA (1990). http://www.inria.fr/rrrt/rr-1247.html
12. Olsson, R.J.O., Karlsson, M., Moberg, L.: Compression of first-order spectral data using the b-spline zero compression method. J. Chemom. **10**(5–6), 399–410 (1996)
13. Ramsay, J., Silverman, B.: Functional data analysis. Springer Series in Statistics. Springer, New York, NY (1997)
14. Rossi, F., François, D., Wertz, V., Verleysen, M.: Fast selection of spectral variables with b-spline compression. Chemom. Intell. Lab. Syst. **86**(2), 208–218 (2007)
15. Saito, N., Coifman, R.R.: Local discriminant bases and their applications. J. Math Imaging Vis. **5**(4), 337–358 (1995)
16. Stone, H.: Approximation of curves by line segments. Math. Comput. **15**, 40–47 (1961)

Part IX
Bio and Health Science

Plaid Model for Microarray Data: an Enhancement of the Pruning Step

Luigi Augugliaro and Angelo M. Mineo

Abstract Microarrays have become a standard tool for studying gene functions. For example, we can investigate if a subset of genes shows a coherent expression pattern under different conditions. The plaid model, a model-based biclustering method, can be used to incorporate the addiction structure used for the microarray experiment. In this paper we describe an enhancement for the plaid model algorithm based on the theory of the false discovery rate.

1 Introduction

There has been considerable recent interest in the analysis of microarray experiments. A typical microarray experiment investigates thousands of genes, recording their expression level over tens of samples. A *bicluster* identifies a group of genes and an associated group of samples on which the genes are characterized by a similar expression level. This fact may indicate a common biological function. Several clustering methods have been developed in recent years in order to identify a bicluster, such as gene-shaving [3], EMMIX-GENE [6], EMMIX-WIRE [8], spectral biclustering [4] and the plaid model [5], among the others. The plaid model is a model-based clustering method that can be used to study structured microarray experiments, for this reason it is usually preferred to the other methods. Aim of this paper is to present an enhancement of the version of the plaid model algorithm proposed in [9], in order to reduce the uncertainty related to the parameters used in the pruning step to remove ill-fitted genes and samples. To increase the interpretation and the accuracy of the pruning step, we have based the proposed enhancement on the theory of the false discovery rate developed by Benjamini and Hochberg [1].

L. Augugliaro (✉)
Dipartimento di Scienze Statistiche e Matematiche, University of Palermo, 90128 Palermo, Italy
e-mail: augugliaro@dssm.unipa.it

B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-13312-1_47,

2 The Plaid Model

The plaid model, proposed in [5], is defined as sum of a series of additive layers, intended to describe and capture the underlying structure of a gene expression matrix. The first layer, defined *background layer*, is used to take into account the global effects in the data, while any subsequent layer represents additional effects corresponding to the bicluster that exhibits a strong pattern not explained by the general formulation of the plaid model. Let Y_{ij} be the expression level of the i-th gene in the j-th sample, the plaid model is defined as follows:

$$Y_{ij} = \sum_{l=0}^{L} \mu_{ij}^{l} \rho_i^l \kappa_j^l + \varepsilon_{ij}, \tag{1}$$

where l is the layer index starting from zero, the background layer, to L, the number of biclusters, and ε_{ij} is a residual error. The mean parameter μ_{ij}^{l} for the l-th layer is defined as sum of three effects, namely

$$\mu_{ij}^{l} = \mu^l + \alpha_i^l + \beta_j^l, \tag{2}$$

where μ^l is the mean effect and α_i^l, β_j^l are the gene and sample effect in the l-th layer, respectively. Lazzeroni and Owen [5] proposed different variants of the model, but the form (2) is usually the most suitable for the analysis of microarray data. The full model is similar to the model used in a two-way analysis of variance, expected that the two-way interaction between genes and samples is replaced by cluster effects, cluster by gene effects and cluster by sample effects. In this way the plaid model tries to decompose the gene by sample interaction into additive layers that are more easily interpretable. Finally, ρ_i^l and κ_j^l are cluster membership parameters defined for $l \geq 1$; ρ_i^l is equal to one if the i-th gene is an element of the l-th layer, zero otherwise. Similarly, κ_j^l is the cluster membership for the j-th sample and is equal to one if the j-th sample is an element of the l-th layer, zero otherwise.

3 The Proposed Enhancement

Turner et al. [9] proposed a new algorithm for the plaid model which uses binary least squares to fit the cluster membership parameters. According to the authors, the proposed algorithm reduces the level of "false" structure incorporated in the model. A further enhancement was proposed in [10]. The authors proposed a variation in the pruning method, used to remove ill-fitted genes and samples, based on the adjustment of the sum of squares for the associated degrees of freedom. Then, the gene and sample pruning is obtained by means of two parameters, namely τ_1 and τ_2, which can be interpreted as the minimum adjusted R^2 desired for genes and samples.

The problem with this variation is that does not exist a criterion to choose a value for τ_1 and τ_2. In order to remove this uncertainty, we have developed a variation of the plaid model algorithm that is based on the algorithm proposed by Benjamini and Hochberg [1] to control the false discovery rate. This variation requires a new element for the original algorithm proposed in [9], namely, a statistical test. Using a statistical test developed to identify differentially expressed genes, we can overcome another important limitation of the original algorithm, namely, the lack of consideration of the technological features of the GeneChip used for the experiment [7]. This limitation has serious consequences in the accuracy of the layer search step. To overcome this problem, we propose to use a statistical test specific for the platform used for the analysis, such as the statistical test developed by Tusher et al. [11] or the statistical test developed by Efron et al. [2], among others. In the following of this paper, we shall assume that we are working with the modified t-statistic proposed by Tusher et al. [11] and called *relative distance*, namely:

$$d(i) = \frac{\bar{Y}_0(i) - \bar{Y}_1(i)}{s(i) + s_0}, \tag{3}$$

where $\bar{Y}_0(i)$ and $\bar{Y}_1(i)$ are the average levels of expression for the i-th gene under the state 0 and 1, respectively; $s(i)$ is the standard deviation of repeated expression measurements and the correction factor s_0 is introduced in order to minimize the coefficient of variation of $d(i)$ (see [11] for a more complete description). Using the modified t-statistic (3) and the algorithm developed in [1], the proposed variation of the pruning step is the following: let $\hat{Z}^l$ be the residual matrix from the plaid model with l layers and let $\hat{Z}^l(\hat{\rho}^l)$ be the submatrix of $\hat{Z}^l$ defined using the estimated class membership parameters $\hat{\rho}^l$. Using $\hat{\kappa}^l$ as classification factor, for each row of $\hat{Z}^l(\hat{\rho}^l)$ we compute the permuted adjusted p-values (p_i^a), then fixing a significant level of 5%, the proposed pruning rule for the ill-fitted genes is the following

$$\tilde{\rho}_i^l = \begin{cases} 1, \text{ if } \hat{\rho}_i^l = 1 \text{ and } p_i^a \leq 0.05 \\ 0, \text{ otherwise.} \end{cases}$$

In a similar way we can prune the samples that are ill-fitted to the bicluster, namely, let $\hat{Z}^l(\hat{\kappa}^l)$ be the submatrix of $\hat{Z}^l$ defined using the estimated class membership parameters $\hat{\kappa}^l$ and $\hat{\rho}^l$ as classification factor, the pruning rule for the samples is the following

$$\tilde{\kappa}_j^l = \begin{cases} 1, \text{ if } \hat{\kappa}_j^l = 1 \text{ and } p_j^a \leq 0.05 \\ 0, \text{ otherwise.} \end{cases}$$

Different significant levels can be used to define the wideness of a bicluster; for example, we can reduce the significant value used in the proposed enhancement in order to obtain a tighter bicluter. Since these values are chosen using the theory of the false discovery rate, the proposed enhancement reduces the uncertainty related with the parameters used in the pruning step. In Table 1 we have reported the algo-

Table 1 The proposed algorithm for the plaid model

1: Compute $\hat{Z}$ the matrix of residuals from the model so far
2: Compute starting values $\hat{\rho}_{0;i}$ and $\hat{\kappa}_{0;j}$ using one-way k-means clusters
3: $s = 1$
4: Update the layer effects using the submatrix of $\hat{Z}$ indexed by $\hat{\rho}_{s-1;i}$ and $\hat{\kappa}_{s-1;j}$
5: Update the cluster membership parameters
6: Follow steps 4 to 5 for $s = 2, 3, \ldots, S$ iterations
7: Compute the layer effects $\hat{\mu}_{S+1}$, $\hat{\alpha}_{S+1;i}$ and $\hat{\beta}_{S+1;j}$
8: Prune bicluster to remove ill-fitted genes and samples using the rules

$$\tilde{\rho}_i = \begin{cases} 1, & \text{if } \hat{\rho}_{S+1;i} = 1 \text{ and } p_i^a \leq 0.05 \\ 0, & \text{otherwise} \end{cases}$$

$$\tilde{\kappa}_j = \begin{cases} 1, & \text{if } \hat{\kappa}_{S+1;j} = 1 \text{ and } p_j^a \leq 0.05 \\ 0, & \text{otherwise} \end{cases}$$

9: Compute the layer sum of squares
$LSS = \sum_{i,j} (\hat{\mu} + \hat{\alpha}_{S+1;i} + \hat{\beta}_{S+1;j})^2 \tilde{\rho}_i \tilde{\kappa}_j$
10: Permute the matrix Z and follow steps 2 to 9; repeat T times
11: Accept bicluster if LSS is greater than LSS for all permuted runs, otherwise stop
12: Refit all layers in the model R times, then search for next layer

rithm developed in [9] with the proposed enhancement for the pruning step. The index layer is removed in order to simplify the notation.

In the next section we evaluate the proposed enhancement with the original algorithm proposed in [9].

4 Simulation Studies

Aim of this section is to evaluate the behaviour of the proposed enhancement by means of simulation studies. We assume that we have a plaid model with only one layer defined using an expression matrix obtained with a 384-well micro fluidic card and with sample size $n = 30$; in this way, we have an expression matrix with 384 rows and 30 columns. The background layer is generated using a standard normal distribution, while the simulated layer is defined using the following values:

$$\begin{aligned} \mu^1 &= 1 \\ \alpha^1 &= \underbrace{[-2, -1, 0, 1, 2, 3, 4, 3, 2, 1, 0, -1, -2]^T}_{13 \times 1} \\ \beta^1 &= \underbrace{[1, 1, 1, 1, 1, 1, 1, 1, 1, 1]^T}_{10 \times 1}. \end{aligned} \tag{4}$$

These values were chosen in order to evaluate how the size of the estimated layer is related with the parameters τ_1 and τ_2 of the algorithm proposed in [9] and with the adjusted p-values used in the method proposed in this paper. For this reason we have simulated 10,000 expression matrices and for each of them we have applied the considered algorithms.

Table 2 Percentage of layers identified for different values of the pruning parameters τ_1 and τ_2

		No. of layers							No. of layers		
τ_1	τ_2	1	2	3	4	5	τ_1	τ_2	1	2	3
0.5	0.5	0.85	0.14	0.02			0.8	0.5	0.99	0.01	
	0.6	0.87	0.12	0.01				0.6	0.99	0.01	
	0.7	0.56	0.34	0.08	0.02			0.7	0.99	0.01	
	0.8	0.58	0.36	0.05	0.01			0.8	0.73	0.26	0.01
	0.9	0.82	0.17	0.01				0.9	0.94	0.06	
0.6	0.5	0.88	0.11	0.01			0.9	0.5	0.97	0.03	
	0.6	0.88	0.11	0.01				0.6	0.99	0.01	
	0.7	0.52	0.36	0.10	0.02	0.01		0.7	0.98	0.02	
	0.8	0.43	0.43	0.10	0.03			0.8	0.99	0.01	
	0.9	0.28	0.62	0.10				0.9	0.99	0.01	
0.7	0.5	0.91	0.08	0.01							
	0.6	0.91	0.08	0.01							
	0.7	0.88	0.11	0.01							
	0.8	0.38	0.41	0.17	0.04						
	0.9	0.91	0.07	0.02							

In Table 2 we have reported the percentage of number of layers identified using different values of the pruning parameters. This table clearly shows two important aspects of the original plaid model algorithm; the first one is that the method proposed in [9] is characterized by an high level of structural instability when τ_1 assumes values lower than 0.6. In this case the algorithm usually identifies one or two layers. The second aspect that characterizes the original algorithm is the joined effect of τ_1 and τ_2 on the percentage of number of layers identified. This aspect is very important since it is not possible to identify the correct number of layers working with a pruning parameter at once. For example, in this simulation study when τ_1 is fixed at 0.7 the proportion of plaid models identified with a single layer varies from 0.91 to 0.38. Then, we have tried to see if there is an interaction effect between τ_1 and τ_2 by using a linear regression model and a GLM with a binomial distribution for the error structure (in both cases we have considered the proportion of identified single layer models as response variable), but this interaction effect does not seem significant (results not shown).

In Table 3 we have computed the proportion of layers identified using the proposed algorithm for different combinations of the adjusted p-values. In particular, we can see that the proposed algorithm is characterized by a low level of structural instability since the number of layers varies from 3 to 0. Moreover, we have not observed a joined effect between the two adjusted p-values.

Another aspect that we think is important in order to evaluate the performance of a biclustering method is the size of the identified biclusters. This aspect is of particular interest since is closely connected with the real aim of the analysis. For example, a biclustering method that finds clusters too tight in respect of the real structure of the expression matrix, can lead to incorrect conclusions since we are leaving out important samples related with the differentially expressed genes. To evaluate this important aspect, we can see that using the parameter values defined in (4) we obtain the following layer

Table 3 Percentage of layers identified for different values of the adjusted p-values

		No. of layers			
p_i^a	p_j^a	0	1	2	3
0.10	0.10	0.005	0.974	0.018	0.003
	0.05	0.008	0.973	0.018	0.001
	0.01	0.008	0.969	0.022	0.001
0.05	0.10	0.023	0.976	0.001	
	0.05	0.015	0.984	0.001	
	0.01	0.011	0.988	0.001	
0.01	0.10	0.14	0.86		
	0.05	0.14	0.86		
	0.01	0.15	0.85		

$$L_{13\times 10} = \begin{bmatrix} 0&0&0&0&0&0&0&0&0&0 \\ 1&1&1&1&1&1&1&1&1&1 \\ 2&2&2&2&2&2&2&2&2&2 \\ 3&3&3&3&3&3&3&3&3&3 \\ 4&4&4&4&4&4&4&4&4&4 \\ 5&5&5&5&5&5&5&5&5&5 \\ 6&6&6&6&6&6&6&6&6&6 \\ 5&5&5&5&5&5&5&5&5&5 \\ 4&4&4&4&4&4&4&4&4&4 \\ 3&3&3&3&3&3&3&3&3&3 \\ 2&2&2&2&2&2&2&2&2&2 \\ 1&1&1&1&1&1&1&1&1&1 \\ 0&0&0&0&0&0&0&0&0&0 \end{bmatrix}$$

from which we observe that the first two and the last two rows cannot be identified by the considered algorithms since they are confounded with the background layer, which is obtained using a standard normal distribution. For this reason, in order to evaluate the performance of the algorithms to identify the real size of the bicluster, we consider as optimal a bicluster with 9 rows and 10 columns.

In Table 4 we have reported the mean and the variance of the number of rows and columns when we have a single layer identified by the original algorithm. To complete our analysis, we have also reported the proportion of single layers identified by the original algorithm seen in table 2. The conclusion that we can draw from this table is that the algorithm proposed in [9] gains in stability excluding important samples. For example, when we choose $\tau_1 = 0.9$ and $\tau_2 = 0.9$ we have an optimal stability of the algorithm (the proportion of single layers identified is 0.99), but the mean value of the number of rows identified is about two. Table 5 is obtained using the algorithm with the proposed enhancement. We can clearly see that the proposed algorithm overcomes the problem with the accuracy previously observed. When we choose the adjusted p-values equal to 0.01 we have an high level of stability (the proportion of single layers identified is equal to 0.85) and at the same time we have an high level of accuracy: indeed, in this case we have always found the optimal layer.

Table 4 Relationship between stability and accuracy for the original plaid model algorithm

			Rows		Columns					Rows		Columns	
τ_1	τ_2	% of 1 layer	Mean	Var	Mean	Var	τ_1	τ_2	% of 1 layer	Mean	Var	Mean	Var
0.5	0.5	0.85	8.96	0.10	10	0.00	0.8	0.5	0.99	4.87	0.12	10	0.00
	0.6	0.87	8.95	0.10	10	0.00		0.6	0.98	4.88	0.10	10	0.00
	0.7	0.56	8.96	0.16	8.01	0.03		0.7	0.99	4.89	0.10	9.89	0.10
	0.8	0.58	8.96	0.18	6.99	0.02		0.8	0.73	4.01	0.01	8.01	0.02
	0.9	0.82	8.17	0.15	3.26	0.57		0.9	0.94	4.83	0.14	7.01	0.01
0.6	0.5	0.88	8.85	0.13	10	0.00	0.9	0.5	0.97	2.01	0.01	9.06	0.05
	0.6	0.88	8.87	0.12	10	0.00		0.6	0.99	2.04	0.04	9.04	0.04
	0.7	0.52	8.90	0.09	8.02	0.03		0.7	0.98	2.03	0.03	9.03	0.03
	0.8	0.43	8.86	0.12	7	0.00		0.8	0.99	2.05	0.05	8.13	0.19
	0.9	0.28	8.00	0.00	3.04	0.04		0.9	0.99	2.04	0.04	6.97	0.08
0.7	0.5	0.92	6.90	0.10	10	0.00							
	0.6	0.91	6.90	0.09	10	0.00							
	0.7	0.88	6.90	0.09	9.01	0.01							
	0.8	0.38	6.92	0.08	8	0.00							
	0.9	0.91	6.00	0.02	5.98	0.04							

Table 5 Relationship between stability and accuracy obtained using the plaid model with the proposed enhancement

			Rows		Columns	
p_i^a	p_j^a	% of 1 Layer	Mean	Var	Mean	Var
0.10	0.10	0.974	8.66	0.90	10.00	0.00
	0.05	0.973	8.68	0.87	10.00	0.00
	0.01	0.969	8.64	0.95	10.00	0.00
0.05	0.10	0.976	8.64	0.97	9.51	1.83
	0.05	0.984	8.64	0.94	9.52	1.68
	0.01	0.988	8.68	0.96	9.57	1.63
0.01	0.10	0.86	9	0.00	10	0.00
	0.05	0.86	9	0.00	10	0.00
	0.01	0.85	9	0.00	10	0.00

5 Conclusions

Plaid model, a model-based clustering method, is one of the most used method to identify sets of genes characterized by a coherent expression pattern over a limited number of samples. Its usefulness is related with the possibility to study different structured microarray experiments. However, in our application we have observed some difficulties with the choice of the parameters used in the pruning step. These difficulties are related with the lack of a well founded criterion that permits to choose the values used to remove ill-fitted genes and samples from the bicluster. To overcome this problem, in this paper we have proposed an enhancement for the plaid model algorithm developed in [9], which is based on the theory of the false discovery rate. In this way, we obtain a better model interpretation, by reducing the uncertainty related with the parameters used in the pruning step. The second important advantage that we have by using the proposed algorithm is that we have an increase in the accuracy of the size of the identified layers. This is due to the fact that we have related the pruning step to the adjusted p-values of a specific test statistic, namely the relative distance proposed in [11], which also considers the technological features of the GeneChip used for the experiment [7].

Acknowledgments This research has been supported by grants of the University of Palermo.

References

1. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B. **57**(1), 289–300 (1995)
2. Efron, B., Tibshirani, R., Storey, J., Tusher, V.: Empirical Bayes analysis of a microarray experiment. J. Am. Stat. Assoc. **96**(456), 1151–1160 (2001)
3. Hastie, T., Tibshirani, R., Eisen, M. B., Alizadeh, A., Levy, R., Staudt, L., Chan, W. C., Botstein, D., Brown, P.: 'Gene Shaving' as a method for identifying distinct sets of genes with similar expression patterns. Genome Biol. **1**(2), 0003.1–0003.21 (2000)
4. Kluger, Y., Basri, R., Chang, J.T., Gerstein, M.: Spectral biclustering of microarray data: coclustering genes and conditions. Genome Res. **13**(4), 703–716 (2003)

5. Lazzeroni, L., Owen, A.: Plaid models for gene expression data. Stat. Sin. **12**(1), 61–86 (2002)
6. McLachlan, G.J., Bean, R.W., Peel, D.: A mixture model-based approach to the clustering of microarray expression data. Bioinformatics **18**(3), 413–422 (2002)
7. Mineo, A.M., Augugliaro, L., Fede, C., Ruggieri, M.: A statistical calibration method based on non-linear mixed model for Affymetrix probe level data. In: Book of short papers of the CLADAG 2007 meeting, pp. 97–100. EUM, Macerata (2007)
8. Ng, S.K., McLachlan, G.J., Wang, K., Ben-Tovim Jones, L., Ng, S.-W.: A mixture model with random-effects components for clustering correlated gene-expression profiles. Bioinformatics **22**(14), 1745–1752 (2006)
9. Turner, H., Bailey T.C., Krzanowski W.J.: Improved biclustering of microarray data demonstrated through systematic performance tests. Comput. Stat. Data Anal. **48**(2), 235–254 (2005)
10. Turner, H., Bailey, T., Krzanowski, W., Hemingway, C.: Biclustering models for structured microarray data. IEEE/ACM Trans. Comput. Biol. Bioinform. **2**(4), 316–329 (2005)
11. Tusher, V.G., Tibshirani, R., Chu, G.: Significance analysis of microarrays applied to the ionizing radiation response. Proc. Natl. Acad. Sci. USA **98**(18), 5116–5121 (2001)

Classification of the Human Papilloma Viruses

Abdoulaye Baniré Diallo, Dunarel Badescu, Mathieu Blanchette, and Vladimir Makarenkov

Abstract In this study we present a whole-genome phylogenetic classification of the human papilloma viruse (HPV) family. We found that the high risk of carcinogenicity taxa are clustered together. The most likely insertion and deletion (indel) scenarios of HPV nucleotides were computed. We also searched for relationships between the number of indels which occurred during the evolution of the HPV family and the degree of carcinogenicity of considered taxa. Linear and polynomial redundancy analyses (RDA) were carried out to relate the HPV carcinogenicity with the number of insertions, deletions and conservations.

1 Introduction

Human papilloma viruses (HPV) form a family of viruses that are well-known for their genomic diversity [1] and potential to cause cervical cancer. Nowadays, about a hundred of HPVs have been identified and the whole genomes of more than eighty of them have been sequenced [6]. They are double-stranded, circular DNA genomes with sizes close to 8 Kbp with complex evolutionary relationships and a small set of genes. A new HPV is recognized as a new HPV type if its complete genome has been cloned and the DNA sequence of the gene L1 differs by more than 10% from the closest known HPV type [3, 8, 9]. Older classifications grouped HPVs according to their higher or lower risk of cutaneous or mucosal diseases. Most of the studies were based on a single gene (usually E6 or E7) analysis. The latter genes are predominantly found in cancer cells due to the binding of their products to the *p53* tumour suppressor protein and the retinoblastoma gene product, respectively [12]. Diagnostics of 3,607 women with cervical cancer from 25 countries revealed that about 89% of them had squamous cell carcinoma (SQUAM cancer) and about 5% had adenosquamous carcinoma (ADENO cancer) ([9]; see also Fig. 1 below). It is worth noting that more than the half of the infection cases are due to the types 16 and

A.B. Diallo (✉)
Département d'informatique, Université du Québec à Montréal, Montréal, Québec H3C 3P8, Canada
e-mail: diallo.abdoulaye@uqam.ca

B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-13312-1_48,

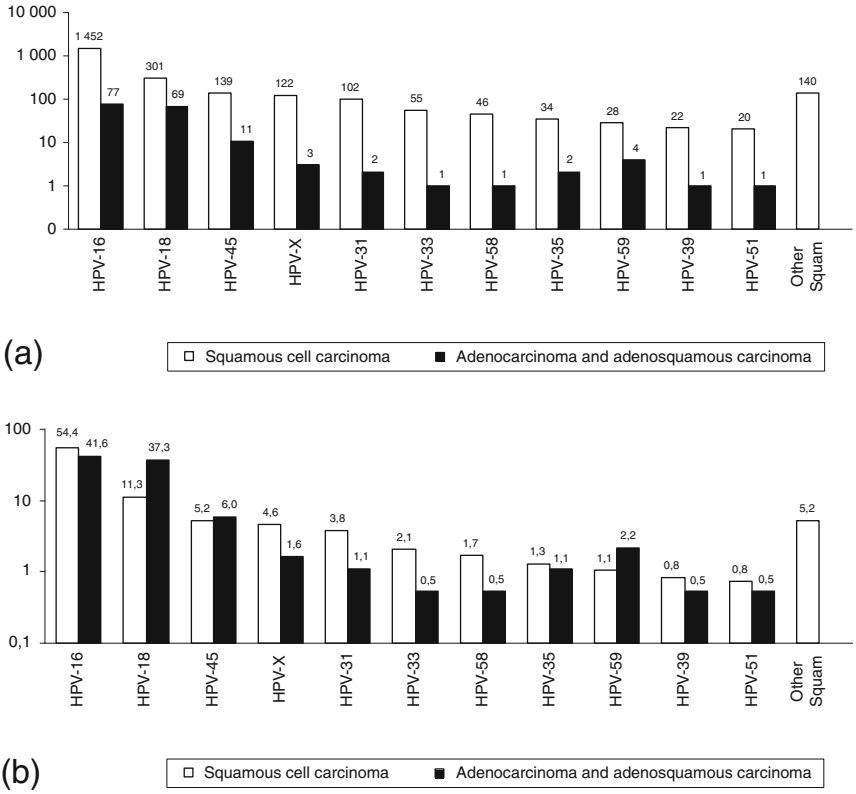


Fig. 1 Distribution of 11 carcinogenic HPVs in terms of the SQUAM and ADENO cancers (drawn using the data from [9]). Total numbers **(a)** and percentages of cases **(b)** are represented on a logarithmic scale. Category "Other Squam" is composed of HPV types being found only for the Squamous cell carcinoma and accounting less than 3% of cases, namely: HPVs-52, 56, 73, 68, 82, 26, 66, 11, 6, HR, 53, 55, 81 and 83. HPVs-35, HR, 68 and X were not considered in this study because their complete genomes were not yet available

18 of HPV ([2]; Fig. 1). Here we studied a whole genome phylogenetic classification of the HPVs and the insertion and deletion (indel) distribution among HPV lineages leading to the different types of cancer. Multiple linear and polynomial regressions and redundancy analyses were used to relate the taxa carcinogenicity with the number of insertions, deletions and conservations, which, in this study, include both conservations and mutations of nucleotides, and estimate the significance of the obtained relationships.

2 Inferring the History of Evolutionary Events

Available genomes of HPVs identified by the ICTV [6] were downloaded and aligned using ClustalW [11]. The alignment length was 10,426 bp. The phylogenetic tree of 83 HPVs (Fig. 2) was inferred using the PHYML method [5] with the

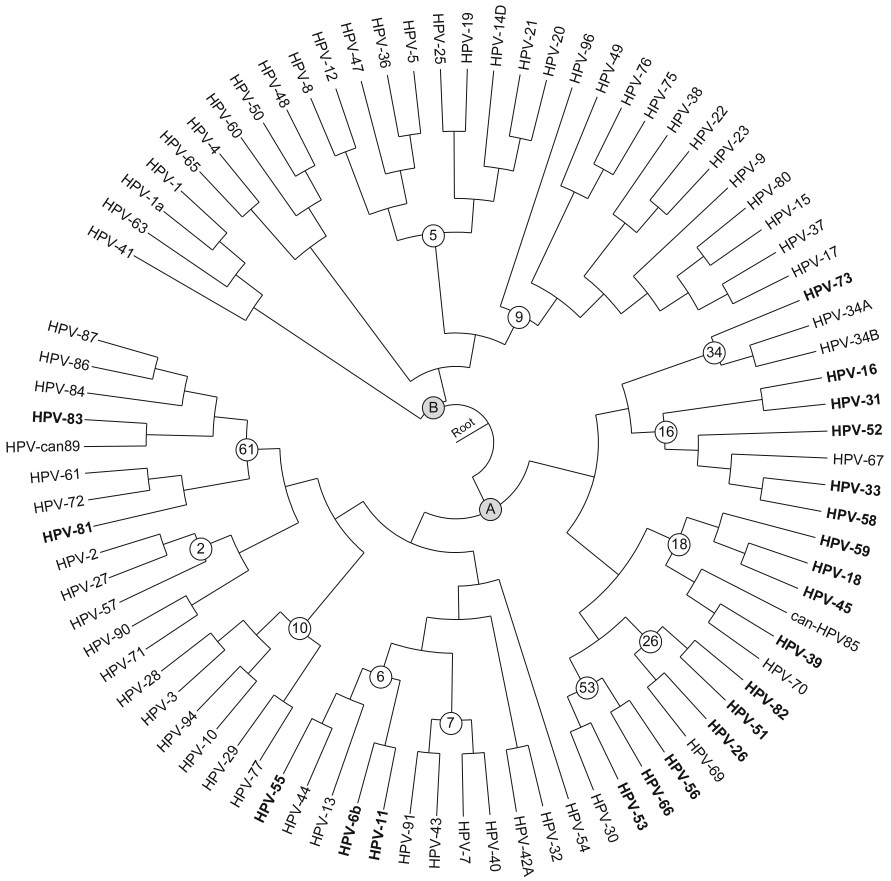


Fig. 2 Phylogenetic tree of 83 HPVs obtained using the PHYML method. The white labelled nodes identify the exiting HPV groups according to the NCBI taxonomy browser and the shaded ones (*A* and *B*) distinguish between the non carcinogenic and carcinogenic families. The 21 carcinogenic HPVs are indicated in *bold* (see Fig. 1)

HKY model of evolution. As suggested in [12], the bovine PV of type 1 was used as an outgroup to root the phylogenetic tree. The bootstrap scores were computed to assess the robustness of the edges. For clarity, they are not indicated in Fig. 2. Mention that they are higher than 80% for most of the edges. In the obtained tree, most of the HPV groups (denoted by numerated nodes) are in agreement with the NCBI/ICTV classifications based on the gene L1. Thus, the evolution of the gene L1 to classify those taxa reflects the whole genome evolution. The most dangerous HPV taxa (see Fig. 2), causing both the ADENO and SQUAM cancers are located in the subtrees rooted by the nodes 16 and 18. The rare HPV-types (less than 3% of all cases; see the caption of Fig. 1) causing the SQUAM cancer only are rather spread around in the large subtree rooted by the node *A* (Fig. 2). We also inferred the phylogenetic trees for all HPV genes and found that, on average, two HPV gene

Table 1 For each of the 15 genes of HPV, this table reports the numbers of the conserved, inserted and deleted regions (and the percentages of nucleotides in these regions) in all lineages of the tree in Fig. 2. Note that the percentages of conservations, insertions and deletions do not sum up to 100% because of the gaps added by the multiple sequence alignment algorithm. These gaps are not explained by deletions but are due to insertions that occurred in the other lineages of the tree

Genes	Cons.	Ins.	Del.	%Cons.	%Ins.	%Del.
E1	12, 111	601	2, 774	91.8	0.3	1
E1A	1, 784	509	320	91.8	1.4	0.6
E2	13, 304	306	3, 460	85.2	0.1	2.2
E4	6, 318	195	2, 117	85.1	0.1	3.8
E5	1, 688	356	503	73.1	2.1	3.1
E5A	208	162	68	79.3	8.2	1.3
E5B	101	31	19	16.3	7.7	0.2
E6	7, 323	613	1, 529	89.0	0.2	1.1
E7	3, 457	0	1, 393	59.4	0	3.9
E8	84	0	0	52.6	0	0
L1	9, 664	314	2, 751	92.7	0.1	1.0
L2	21, 716	494	5, 138	92.3	0.4	2.6
X	484	0	230	43.7	0	1.8
Y	1, 457	54	679	83.2	0.3	2.6
Z	0	0	6	0	0	0.4

phylogenies differ topologically from each other by about 5% (i.e., the Robinson and Foulds distance was used to compare the gene phylogenies).

To quantify the indel distribution, the most likely indel scenarios were computed using a heuristic algorithm described in [4]. For a given phylogenetic tree and the associated multiple sequence alignment, this algorithm computes the set of insertion and deletion events using a tree-based Hidden Markov Model (HMM). Table 1 presents the distribution of the predicted indel and conservation events for all HPV genes. The indel frequencies are higher in the subtrees rooted by the node 61, where only low-risk-carcinogenicity HPVs are located (Fig. 2). The groups located in the subtree rooted by the node *A* have usually a higher percentage of conserved characters on each edge. One can conclude that the organisms of this subtree inherited their carcinogenicity from their least common ancestor. The detailed analysis of the edges of this subtree should be carried out but this goes beyond the scope of this article.

3 Finding Relationships Between the Two Types of Cancer and the Indel/Conservation Distributions in the HPV Genes

We carried out linear and polynomial regressions to establish relationships between the explanatory variables (conservations, insertions and deletions in our case) and response variables (cancer/no cancer outcomes for the SQUAM and ADENO cancers, respectively). To perform the regression, we considered the eight most important HPV genes for the group of 83 HPV viruses (Table 2). The numbers of

conserved, inserted and deleted regions as well as the percentages of characters involved in these evolutionary events, reported in Table 1, formed the matrix of explanatory variables **X**. Two binary variables, consisting of the SQUAM and ADENO cancer outcomes, formed the matrix of response variables **Y**. If a HPV organism can initiate the SQUAM cancer (21 of such HPV organisms were considered) the corresponding value of the first response variable was set to 1, and if it can initiate the ADENO cancer (nine of such HPV organisms were considered) the value of the second response variable was set to 1, otherwise they were set to 0.

Linear and polynomial regressions were carried out separately for the eight genes in Table 2 and for the whole genomic sequences. Generally, both linear and polynomial models were significant: most of the *p*-values for the linear and polynomial regressions were smaller than 0.05 (Table 2). We also performed the test of the difference between the polynomial and linear regressions (last column of Table 2) according to the method discussed in [7]. This test allows one to estimate the possibility of overfitting the data by polynomial regression. If both polynomial and linear models are significant, the significance of the difference between them suggests that the polynomial model is more appropriate in this case, otherwise it suggests the overfitting by polynomial regression and the linear model is preferable.

The results in Table 2 indicate that for the genes E4 and L2 the presence and absence of the SQUAM and ADENO cancers correlate the best with the considered evolutionary events. These two genes should be further analysed by virologists interested by studying the carcinogenic human papilloma viruses. In this way, we carried out linear (for the gene L2 because the difference between the polynomial and linear regressions for this gene was not significant, see Table 2) and polynomial (for gene the E4 because this difference was significant) redundancy analysis (RDA) to find the detailed relationship between the carcinogenic HPVs and the insertion, deletion and conservation (including actual conservations and mutations) events they underwent. RDA [10] allows one to model relationships between the explanatory

Table 2 Percentages of variance accounted for by the linear and polynomial regression for the 8 most important HPV genes and for the whole genomes; *p*-values of the linear and polynomial regressions as well as of their difference are reported. The genes, *E4* and *L2*, for which the best results were obtained, are highlighted. The numbers of taxa available for each gene are shown between the parentheses in the first column

Genes	% of variance for lin. regr.	% of variance for pol. regr.	Lin. regr. *p*-value	Pol. regr. *p*-value	Difference *p*-value
E1 (81)	24.89	41.02	0.01	0.01	0.03
E2 (81)	24.49	41.70	0.01	0.01	0.02
E4 (57)	**32.12**	**58.47**	**0.01**	**0.01**	**0.01**
E5 (20)	39.84	64.98	0.49	0.72	0.71
E6 (81)	31.80	43.42	0.01	0.01	0.08
E7(81)	30.89	38.36	0.01	0.01	0.17
L1(83)	24.74	33.38	0.01	0.01	0.30
L2 (83)	**42.55**	**47.54**	**0.01**	**0.01**	**0.64**
All genes	27.57	36.15	0.02	0.03	0.65

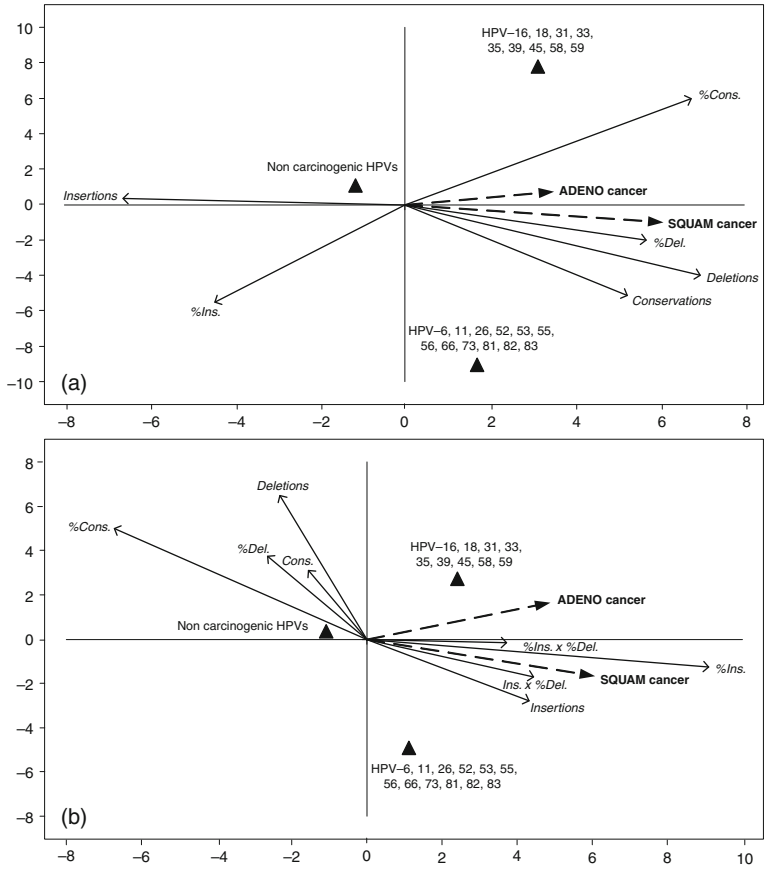


Fig. 3 Linear (case **a** – for the gene L2) and polynomial (case **b** – for the gene E4) RDA biplots for the 83-taxa HPV dataset. Triangles represent the three types of HPVs: viruses causing both types of cancer: HPVs-16, 18, 31, 33, 35, 39, 45, 58 and 59; viruses causing only the SQUAM cancer: HPVs-6, 11, 26, 52, 53, 55, 56, 66, 73, 81, 82 and 83; and, non carcinogenic HPVs. Note that all HPVs causing the ADENO cancer also cause the SQUAM cancer. *Dashed arrows* represent two binary response variables: SQUAM and ADENO cancers. *Solid arrows* represent the numbers of conserved, inserted and deleted regions and the corresponding percentages of the conserved, inserted and deleted nucleotides

variables (conservations, insertions and deletions), response variables (cancer/no cancer outcomes for the SQUAM and ADENO cancers) and considered group of species (83 HPV organisms). For instance, the polynomial RDA, introduced in [7], allows for modeling non-linear relationships between the explanatory and response variables. The correlation biplot [7] was used in this study to represent the relationships between the variables in **X** and **Y**. In such a biplot the angles between the variables from sets **X** and **Y** reflect their correlations; projecting a HPV type (denoted by a triangle in Fig. 3) at right angle on a response variable **y** approximates the value of this HPV type along this variable; projecting a HPV type at

right angle on an explanatory variable **x** approximates the value of this HPV type along this variable. In total, six response variables corresponding to the columns of Table 1 for both genes L2 (Fig. 3a) and E4 (Fig. 3b), and 2 combined variables *%Ins.*x*%Del* and *Ins.*x*%Del* for the gene E4 only (Fig. 3b) were depicted. The two represented combined variables were chosen among all available combined variables because they provided the strongest positive correlations with the SQUAM cancer arrow (Fig. 3b); all other combined variables are not represented in polynomial biplot because they don't correlate strongly, either positively or negatively, with the two response variables depicted by dashed arrows. While observing the biplot ordination diagram for the gene L2 (Fig. 3a), the following main trends can be noticed: both types of carcinogenic HPVs have a greater number of conserved and deleted nucleotides compared to the non carcinogenic HPVs, whereas the non carcinogenic HPVs usually have a higher number of insertions. Also, the presence of the SQUAM cancer is strongly positively correlated with the percentage of deleted nucleotides in the lineages of the gene L2, and both SQUAM and SQUAM cancer types are strongly negatively correlated with the number of insertions. As to the gene E4 (Fig. 3b), the presence of the SQUAM cancer is strongly positively correlated with the percentage of inserted nucleotides as well as with the two depicted combined variables consisting of the products of the percentages of inserted and deleted nucleotides and of the number of insertions and percentage of deletions. Also, the SQUAM cancer HPVs are strongly negatively correlated to the percentages of conserved and deleted nucleotides. Finally, for the gene E4 both types of carcinogenic HPVs have a higher number of insertions compared to the non carcinogenic ones.

4 Conclusion

In this article we studied the classification of the Human Papilloma Viruses (HPV) presumed to be the main cause of the cervical cancer. First, we inferred the PHYML phylogenetic tree [5] of the 83 available HPV organisms (Fig. 2) on the basis of the whole genome phylogenies. We found that all HPV groups (see the 12 HPV subtrees denoted by white nodes in Fig. 2) are monophyletic (i.e., compatible with the current NCBI/ICTV classifications). Then, we inferred the most likely insertion and deletion scenarios for each of the 15 considered HPV genes (Table 1) and found that most of them have more than 90% of the characters conserved throughout the evolution. Multiple linear and polynomial regressions were carried out in order to establish relationships between the conservation, insertion and deletion events and cancer/no cancer outcomes. We found that the presence and absence of both types of cancer correlated the best with the considered evolutionary events in the genes *E4* and *L2*, and the only gene for which the regression p-values were not significant was the gene *E5*. This result warranted additional investigations of the genes *E4* and *L2* consisting of the linear [10] and polynomial [7] RDA conducted for them. RDA biplots drawn for these two genes, shown in Fig. 3, present the detailed relationships between the SQUAM and ADENO cancers, three types of HPV groups

and six selected evolutionary events. Further investigations should be conducted by virologists based on the findings on this study. It would be also interesting to study a model examining the conservation and mutation events separately.

References

1. Antonsson, A., Forslund, O., Ekberg, H., Sterner, G., Hansson, B.G.: The ubiquity and impressive genomic diversity of human skin papillomaviruses suggest a commensalic nature of these viruses. J. Virol. **74**, 11636–11641 (2000)
2. Chan, S.Y., Delius, H., Halpern, A.L., Bernard, H.U.: Analysis of genomic sequences of 95 PV types: uniting typing, phylogeny, and taxonomy. J. Virol. **69**, 3074–3083 (1995)
3. De Villiers, E.M., Fauquet, C., Broker, T.R., Bernard, H.U., Zur Hausen, H.: Classification of papillomaviruses. Virology **324**, 17–27 (2004)
4. Diallo, A.B., Makarenkov, V., Blanchette, M.: Exact and heuristic algorithms for the Indel Maximum Likelihood Problem. J. Comp. Biol. **14**, 446–461 (2007)
5. Guindon, S., Gascuel, O.: A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. **52**, 696–704 (2003)
6. ICTVdB Management: The Universal Virus Database, New York, Bchen-Osmond, C., Columbia University. http://www.ncbi.nlm.nih.gov/ICTVdb/ (2006)
7. Makarenkov, V., Legendre, P.: Nonlinear redundancy analysis and canonical correspondence analysis based on polynomial regression. Ecology **83**, 1146–1161 (2002)
8. Mũnoz, N., Bosch, F.X., Castellsagu, X., Diaz, M., De Sanjose, S., Hammouda, D. et al.: Against which human papillomavirus types shall we vaccinate and screen? The international perspective. Int. J. Cancer. **111**, 278–285 (2004)
9. Mũnoz, N., Bosch, F.X., De Sanjose, S., Herrero, R., Castellsague, X., Shah, K.V. et al.: Epidemiologic classification of human papillomavirus types associated with cervical cancer. N. Engl. J. Med. **348**, 518–527 (2003)
10. Rao, C.R.: Linear statistical inference and its applications. Wiley, New York, NY (1973)
11. Thompson, J.D., Higgins, D.G., Gibson, T.J.: CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**, 4673–4680 (1994)
12. Van Ranst, M., Kaplan, J.B., Burk, R.D.: Phylogenetic classification of human papillomaviruses: correlation with clinical manifestations. J. Gen. Virol. **73**, 2653–2660 (1992)

Toward the Discovery of Itemsets with Significant Variations in Gene Expression Matrices

Mehdi Kaytoue, Sébastien Duplessis, and Amedeo Napoli

Abstract Gene expression matrices are numerical tables that describe the level of expression of genes in different situations, characterizing their behaviour. Biologists are interested in identifying groups of genes presenting similar quantitative variations of expression. This paper presents new syntactic constraints for itemset mining in particular Boolean gene expression matrices. A two dimensional gene expression profile representation is introduced and adapted to itemset mining allowing one to control gene expression. Syntactic constraints are used to discover itemsets with significant expression variations from a large collection of gene expression profiles.

1 Introduction and Motivations

Microarray biotechnology is able to quantitatively measure the expression of a gene in a given biological environment or situation, which is relative to its activity. When considering the expression of a gene in m situations (different cells, times, ...), a so-called gene expression profile (GEP) can be written as a numerical m-dimensional vector, describing the behaviour of the gene. A gene expression matrix (GEM) is a collection of n gene expression profiles (see Table 2) that may be represented as an $n \times m$ numerical table (see Table 1): each line is the profile of a gene.

A widely admitted hypothesis states that genes having a similar expression profile may participate in a same biological function or process [14]. Then, classical clustering methods are used to extract clusters of genes (k-means, hierarchical clustering, see [5] for a survey in GEM analysis). A cluster represents a set of genes that globally have a similar gene expression profile (or line in the table) w.r.t. a similarity measure, e.g. the group $\{g_1, g_2\}$ in Table 1 when considering Euclidean distance. Such genes are said to be co-expressed. However these methods extract *global* patterns and may not be well designed to take into account inherent noise of microarray due to experimental manipulations. Moreover a biological function is

M. Kaytoue (✉)
LORIA, Campus Scientifique, Vandoeuvre-lés-Nancy, France
e-mail: kaytouem@loria.fr

B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-13312-1_49,

Table 1 A gene expression matrix (GEM) composed of five gene expression profiles (GEP)

Gene/situation	S_a	S_b	S_c
g_1	21,050	21,950	1,503
g_2	23,025	24,100	1,708
g_3	62,57	5,057	6,500
g_4	5,392	6,020	7,300
g_5	23,070	22,021	25,548

Table 2 GEP standard vectorial representation. Each vector reflects the behaviour of a gene

Gene	Gene expression profile (GEP)
g_1	(21,050, 21950, 1503)
g_2	(23,025, 24100, 1708)
g_3	(6257, 5057, 6500)
g_4	(5392, 6020, 7300)
g_5	(23070, 22021, 25548)

not necessarily active in all situations of a given dataset. A gene may be involved in several processes/functions, therefore clustering methods should allow overlapping of clusters, which is not often the case.

Biclustering methods, see [8] for a survey in GED analysis, perform a simultaneous clustering of the rows and columns of a table, and thus highlight bi-clusters, or *local* associations between both gene and situation sets, i.e. extracting blocks or "subtables" of similar values. Intuitively, *bi-clusters* are composed of genes sharing similar and *local* expression patterns across a subset of biological situations, e.g. genes of $\{g_1, g_2, g_5\}$ in Table 1 share intuitively a similar *sub-profile* in situations S_a and S_b only. Due to complexity in numerical data, heuristics are used to reduce the result size (or simply enable its computation) and may miss bi-clusters of interest: result is generally composed of one bicluster (the best) or K (a priori chosen) [8]. Therefore overlapping is rarely possible.

It is now part of a growing area to consider symbolic methods for knowledge discovery in Boolean gene expression matrices such as itemset search, association rule extraction, and formal concept analysis, see [10] for a smooth introduction to these methods and [1, 3, 12] for applications to GED analysis. As these methods work on a binary table, the expression matrix is firstly discretized, then local patterns are extracted. Generally, the discretization procedure consists in choosing a threshold t for each gene expression profile (see [1, 12] for examples of threshold calculation). Values higher than t are said to be *over-expressed* and encoded in "1", "0" otherwise. In such derived Boolean expression matrices, whole set of patterns (namely, itemset, assocation rules and formal concepts) is generally tractable but too large to be analysed by human-experts. Some solutions exist to reduce it: the use of a minimal pattern frequency as pruning constraint, (approximative) condensed representations and *a posteriori* extracted pattern clustering [2].

Due to the discretization technique used, symbolic methods [10] extracting local patterns in such derived data may highlight one type of variation of expression only, w.r.t. to the threshold t, i.e. above or below t. In this paper, we propose an interval-

based discretization (Sect. 2) of a GEM. It builds a 2-dimensional vector for each GEP. This transformation allows to apply classical itemset search algorithms to the complex data that are GEM (Sect. 3). We define in Sect. 4 some constraints to fully characterize and control expression *variations* and retain the most variant expression patterns. It dramatically prunes the result, in fact an itemset lattice. Starting from a real-world dataset, Sect. 5 shows that high variations allow biologists to easily discriminate groups of genes and to discover biological processes involving them. Finally, a conclusion draws future researches.

2 Gene Expression Profile Representation

A gene expression profile (GEP) is considered as a vector of numerical values such as (21,050, 21,950, 1,503) for gene g_1 in Table 1 describing the expression of the gene in given situations (S_a, S_b, S_c). Expression values are ranged from 0 (not expressed) to 65,535 (highly expressed), as being monitored with NimbleGen Systems Oligonucleotide Arrays technology[1]. Discretization is based on an interval set T determined either by the expert or statistical methods, e.g. quantile histogram intervals. T is set of intervals that dichotomize the expression value domain into disjunctive intervals, such as $T = \{[0, 10000), [10000, 20000), [20000, 65535)\}$.

Now a gene expression profile can be represented as a 2-dimensional vector $g = \{(a_1, n_1), ..., (a_m, n_m)\}$ where a_k is a biological situation and n_k is the index of an interval in T with $k \in [1, m]$. In the example, $n_k \in \{0, 1, 2\}$ according to $|T| = 3$. This transformation is illustrated from Tables 2 to 3, i.e. from numerical gene expression profiles (GEP) to 2D-GEP. For example, all the values of gene expression profile g_3 are included in the first interval (index 0) of T for each situation S_a, S_b and S_c, therefore it can be represented by the 2D-vector $((S_a, 0), (S_b, 0), (S_c, 0))$.

This representation allows one to mine gene expression data to extract patterns of genes having similar expression values, i.e. in the same interval, in some or all situations. Moreover, this also allows in Sect. 4 to characterize expression variations that biologists wish to highlight. As we can consider that each 2D-GEP is composed of items, i.e. elements of the Cartesian product $S \times N$ where S is a set of situations and N the index set of T, next section formalizes itemset search for 2D-GEP.

Table 3 2D-GEP vectorial representation

Gene	2D-gene expression profil (2D-GEP)
g_1	$((S_a, 2), (S_b, 2), (S_c, 0))$
g_2	$((S_a, 2), (S_b, 2), (S_c, 0))$
g_3	$((S_a, 0), (S_b, 0), (S_c, 0))$
g_4	$((S_a, 0), (S_b, 0), (S_c, 0))$
g_5	$((S_a, 2), (S_b, 2), (S_c, 2))$

[1] http://www.nimblegen.com/

3 Itemset Search

A classical definition of an itemset is the following [10]. Given a set of object O and a set of properties P, an *item* corresponds to a property of an object, and an *itemset*, or a *pattern*, to a set of items: an object is said to own an item. The number of items in an itemset determines its *length*. Its *image* corresponds to set of objects owning all items of the itemset. We call *support* of an itemset the cardinality of its image.

The number of potential itemsets is $2^{|P|}$, i.e. the number of all possible subsets of P. Using condensed representations such as *closed itemsets* [11] allows to reduce this number. An itemset is *closed* if maximal in its equivalence class, i.e. its length is maximal w.r.t. all other itemsets having the same image.

A 2D-GEP is composed of pairs that can be considered as items: a 2D-GEP is an object owning pairs. For example, 2D-GEP g_1 owns property $(S_a, 2)$ and not property $(S_a, 1)$. It is now possible to apply classical itemset search algorithms, e.g. Charm [4]. On the example, the set of object $O = \{g_1, ..., g_5\}$ and $P = \{S_a, S_b, S_c\} \times \{0, 1, 2\}$, remembering that $|T| = 3$. Computation returns 5 closed itemsets, see Table 4, among which $\{(S_a, 2), (S_b, 2), (S_c, 0)\}$. Its image is $\{g_1, g_2\}$, meaning that the genes g_1 and g_2 are co-expressed by sharing similar values in the same interval for all situations (depending on T).

Table 4 Closed itemsets extracted in Table 3

Closed itemsets	Image
$\{(S_a, 2), (S_b, 2), (S_c, 2)\}$	$\{g_5\}$
$\{(S_a, 0), (S_b, 0), (S_c, 0)\}$	$\{g_3, g_4\}$
$\{(S_a, 2), (S_b, 2), (S_c, 0)\}$	$\{g_1, g_2\}$
$\{(S_a, 2), (S_b, 2)\}$	$\{g_1, g_2, g_5\}$
$\{(S_c, 0)\}$	$\{g_1, g_2, g_3, g_4\}$

In Table 4, an illustration of how itemsets extraction is a kind of bi-clustering allowing to take noise into account can be made. $\{(S_a, 2), (S_b, 2), (S_c, 0)\}$ is a *global pattern*, i.e. involving all the situations, and represents a cluster. $\{(S_a, 2), (S_b, 2)\}$ is a *local pattern*, i.e. involving a subset of the situations only, and represents a bi-cluster. Moreover, though the value derived into $(S_c, 2)$ may be an artefact for gene g_5, i.e. should have been $(S_c, 0)$, the group $\{g_1, g_2, g_5\}$ nevertheless exists.

4 Minimal Variation Constraints

A gene expression matrix can contain thousands of genes and dozens of situations. Depending on the choice of the discretization intervals in T, the size of both property and resulting itemset sets may be still very large: *closure constraint* on itemsets is not enough. A classical solution [11] is to retain frequent-closed itemsets, i.e. having a support greater than a given minimal support.

Here, we suggest another possibility. The biologists focus on gene groups presenting similar expression values in some or all situations and having the most important variations of expression simultaneously. Interpretation of variations leads after experimental validations to the discovery of gene functions and biological processes. Large variations are important to discriminate genes responsible for a particular cellular process [14]. As less than 10% of genes have a high variation of expression from a situation to another [5], we can suppose that a large part of the whole set of itemsets presents no or "small" variations of expression (latter defined).

In our context, an itemset $B = \{(a_1, n_1), ..., (a_m, n_m)\}$, where a_k is a biological situation and n_k is the index of an interval in T with $k \in [1, m]$, is a pattern for a group of genes (actually composing the image of B). B is composed of valuations controlling expression in situations (pairs). Numerical syntactic constraints can be designed to characterize variations of expression and retain from the very large collection of itemsets only those having most important variations. The key idea is the following: the itemset $B = \{(S_a, 1), (S_b, 1), (S_c, 1)\}$ presents no variation of expression: all n such as $(s, n) \in B$ are equals, i.e. the expression values are always in the same interval.

Filter to retain itemsets showing variations of expression. In the current and two next paragraphs, we consider (p_i, k_i) and (p_j, k_j) as two distinct pairs of an itemset B, with $i \neq j$. A *variation* is defined as a non null difference between k_i and k_j. Then we can define a *variation constraint*: retaining *variant itemsets* consists in keeping those having at least one variation, i.e. respecting the predicate (1). Others, called *constant itemsets*, are removed. We leave to the reader to check that $\{(S_a, 2), (S_b, 2), (S_c, 2)\}$ is constant and that $\{(S_a, 3), (S_b, 2), (S_c, 2)\}$ is variant.

Filter to control variation amplitude. One may notice that $\{(S_a, 15), (S_b, 2), (S_c, 2)\}$ has unformally higher variations than $\{(S_a, 3), (S_b, 2),(S_c, 2)\}$, because $15 - 2 > 3 - 2$. Thus to have more control on variations, we define an α-*variation constraint*: an α-*variation* is a difference between k_i and k_j of at least α, i.e. $|k_i - k_j| \geq \alpha$. Then an itemset B is α-*variant* if it respects the predicate (2), with $\alpha \geq 0$, i.e. it has at least one α-variation, e.g. $\{(a, 2), (b, 6), (c, 6)\}$ with $\alpha \leq 4$.

Filter to control occurrences of an α-variation. Finally, yet another may notice that $\{(S_a, 15), (S_b, 2), (S_c, 12)\}$ has more variations than $\{(S_a, 15), (S_b, 2), (S_c, 2)\}$. Then the (α, β)-*variation constraint* is defined as follows: an itemset is $(\alpha, \beta) - variant$ if it respects the predicate (3), with $\alpha \geq 0$ *and* $\beta \geq 1$. Intuitively an $(\alpha, \beta) - variant$ itemset presents at least a number β of α-variations, e.g. the itemset $\{(a, 2), (b, 6), (c, 11)\}$ is (4, 3)-variant, while $\{(a, 2), (b, 6), (c, 8)\}$ is not.

$$isVariant(B) \equiv \exists (a_i, k_i) \in B \text{ and } \exists (a_j, k_j) \in B \text{ such as } k_i \neq k_j \quad (1)$$

$$is\alpha Variant(B, \alpha) \equiv \exists (a_i, k_i) \in B \text{ and } \exists (a_j, k_j) \in B \text{ such as } |k_i - k_j| \geq \alpha \quad (2)$$

$$is\alpha\beta Variant(B, \alpha, \beta) \equiv |\{((a_i, k_i), (a_j, k_j)) \text{ with } |k_i - k_j| \geq \alpha\}| \geq \beta \quad (3)$$

α and β are two parameters allowing the biologist to focus on the most important variations. The choice of these parameters strongly depends on the choice T.

5 Experiments

In this section, we show the efficiency of the introduced constraints on real data. Biologists at the UMR IAM (INRA) study the symbiosis between the fungus *Laccaria bicolor* and the tree *Populus*. Thanks to molecular exchanges between root tissues, the productivity of a *Populus* forest may be increased by 30%. The recent sequencing of *Laccaria bicolor* genome predicted more than 20,000 genes [9]. It remains now to study their expression in many environments to understand their functions in the fungal lifestyle. A gene expression matrix is available at the Gene Expression Omnibus at National Centre for Biotechnology Information (NCBI)[2]. It is composed of 22,294 genes in lines and 7 various biological situations in columns, i.e. free-living cells (M81306 and MS238), young (FBe) and mature (FBl) fruiting body cells and fungal cells in association with roots of trees (MPgh, Mpiv, and MD).

Before experimenting our approach, we present a standard k-means result in Fig. 1 with $k = 10$. Most of the clusters prototypes do not present any variation. Increasing the number k of clusters or using fuzzy k-means do not change the problem. In contrary, our method generates lots of prototypes w.r.t. the number of chosen intervals but directly characterizes most informative patterns, i.e. variant. Indeed, an itemset represents an intelligible description of genes composing its image.

Firstly, we work with all the $22,294$ genes and a set of situations $S = \{MP, MD, Fbe, FBl, Myc\}$. MP represents in-symbiosis cells (mean of $MPgh$ and $Mpiv$), and Myc represents mycelium cells (mean of $M81306$ and $MS238$). The expert biologist chooses a set of intervals $T = \{[0, 2000), [20000, 40000), [40000, 65535)\}$ for all situations. In certain data, T may differ for each situation. In our case, microarray manufacturer automatically normalized GED. Search returns 893 itemsets among which 35 that have a minimal length of 4 and that are $(2, 3) - variant$. Two of them are presented in Fig. 2 (a) $B = \{(MD, 0), (FBe, 2), (FBl, 2), (Myc, 0)\}$ and (b) $B = \{(MP, 2), (MD, 2), (FBe, 2), (FBl, 2), (Myc, 0)\}$. These itemsets are patterns of genes shar-

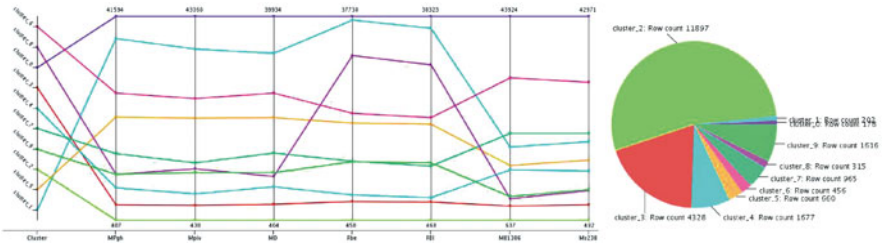


Fig. 1 A k-means algorithm result. *Left*: parallel coordinates. *Right*: cluster cardinality. Designed with Knime (http://www.knime.org). Intuitively, more than 75% of the genes are captured in clusters whose prototype characterizes no variations of expression

[2] http://www.ncbi.nlm.nih.gov/geo/ as series GSE9784

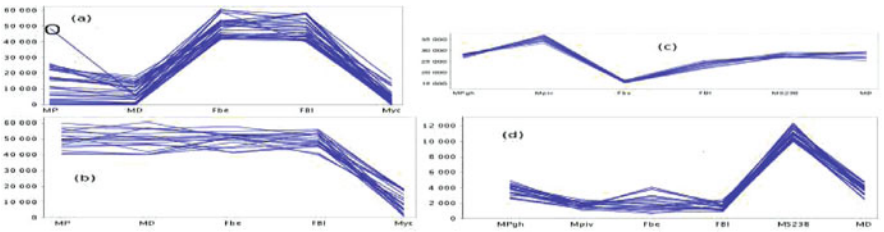


Fig. 2 Each picture is a graphical representation of an itemset. Y-axis contains the situations. X-axis is the expression value axis. Each line represents the numerical vector of the GEP of each gene composing the image of the corresponding itemset

ing the same behaviour. Most of the genes of the 35 itemsets remain today of unknown function. However, some hypothesis can be made. Genes of group (a) may be involved in processes of the fruit body structure: their expression values are high only in Fbe and FBl. Genes of group (b) may play a major role in symbiosis: their expression is high in in-symbiosis and fruit cells and low in free-living mycelium cells: symbiosis is favoured when the fruit is well established.

A second experiment follows the same principle with more situations (6) and more intervals (15). We extract 71, 391 itemsets and retain those of minimal length 4 that are $(4, 2) - variant$: 9, 324 itemsets remain. Most of them have a support less than 10. Then, we also add the minimal frequency constraint: support must be greater than 10. Then 54 itemsets remain and are analysable. The image of two of them is presented in Fig. 2 (c) and (d). Genes of (c) are strongly co-expressed but their function is here again unknown. However, they have been identified as potential proteins of the same type in the yeast species *Candida albicans* by comparing DNA sequences. Genes of (d) may be involved in growth of mycelium.

We have shown two experiments, the first with a few intervals for discretization, i.e. $|T|$ small, and the second with a high number. The choice of the number of intervals and their size is difficult and directly influences the quality (not studied in this paper) and the cardinality of the result (Fig. 3). If $|T|$ is low, the number of itemsets and their quality is generally low w.r.t. a higher $|T|$. If $|T|$ is high, the number of itemsets explodes, but the quality is better, and the filters allow to reduce it (Fig. 4). Genes of Fig. 2 (c) and (d) would have been buried with less similar genes with $|T| = 3$. Instead of manually choosing intervals, some possibilities may be investigated. Firstly, intervals can be computed automatically according to some criterion of optimization. Most of the time, this criterion refers to known class membership of a part of objects [7]. However, most of the genes of interest in *Laccaria bicolor* are unknown, and known genes have generally an expression without variations of expression across the situations we study. Another possibility is to calculate intervals that optimize the support of the resulting itemsets, e.g. [13] with genetic algorithms. However, it is supposed that biological processes can not be characterized by a large amount of genes. Finally, a similar method has been designed without transforming data in [6]. However, it is there harder to consider variations in sense of this paper and should be investigated.

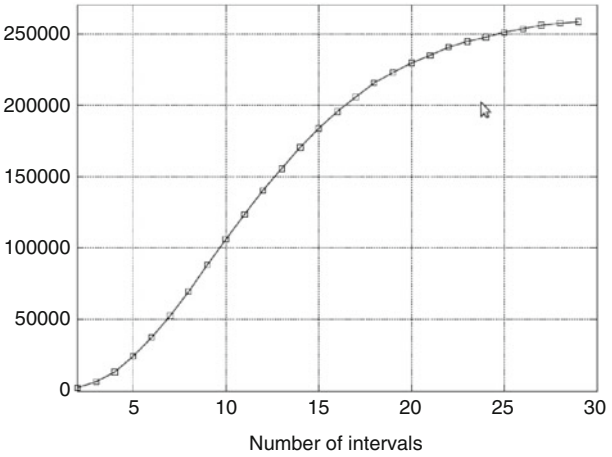


Fig. 3 Number of itemsets when intervals are quantiles

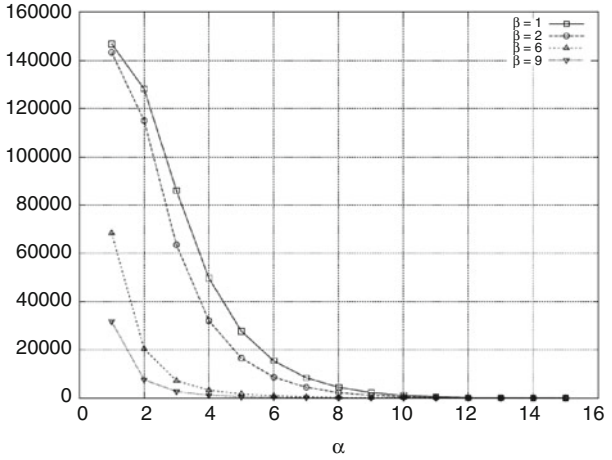


Fig. 4 Number of itemsets w.r.t. α and β and 15-quantiles

6 Conclusion

In this paper, we have presented a method for transforming complex data describing gene expression profiles into standard nominal tables. This enables to apply classical itemset search algorithms to these complex data. Classically, when data are discretized it goes with loss of information. We limit this loss by introducing a two dimensional representation of objects that allows to control the original numerical values, by retaining the most variant patterns and pruning the result.

References

1. Blachon, S., Pensa, R., Besson, J., Robardet, C., Boulicaut, J.-F., Gandrillon, O.: Clustering formal concepts to discover biologically relevant knowledge from gene expression data. In Silico. Biol. **7**(0033), 1–15 (July 2007)
2. Boulicaut, J.-F., Besson, J.: Actionability and formal concepts: a data mining perspective. In: Formal Concept Analysis, LNAI 4933, pp. 14–31. Springer, Heidelberg (2008)
3. Creighton, C., Hanash, S.: Mining gene expression databases for association rules. Bioinformatics **19**(1), 79–86 (2003)
4. Hsiao, C.-J., Zaki, M.J.: Efficient algorithms for mining closed itemsets and their lattice structure. IEEE Trans. Knowl. Data Eng. **17**(4), 462–478 (2005)
5. Jiang, D., Tang, C., Zhang, A.: Cluster analysis for gene expression data: a survey. IEEE Trans. Knowl. Data Eng. **16**(11), 1370–1386 (2004)
6. Kaytoue, M., Duplessis, S., Kuznetsov, S.O., Napoli., A.: Two FCA-based methods for mining gene expression data. In: Formal Concept Analysis, LNAI 5548, pp. 251–266. Springer, Heidelberg (2009)
7. Kurgan, L., Cios, K., Kurgan, L.A., Cios, K.J., Member, S.: Caim discretization algorithm. IEEE Trans. Knowl. Data Eng. **16,** 145–153 (2004)
8. Madeira, S.C., Oliveira, A.L.: Biclustering algorithms for biological data analysis: a survey. IEEE/ACM Trans. Comput. Biol. Bioinform. **1**(1), 24–45 (2004)
9. Martin, F.: The genome of Laccaria bicolor provides insights into mycorrhizal symbiosis. Nature **452**(7183), 88–92 (2008). 68 Co-authors have participated in this paper
10. Napoli, A.: A smooth introduction to symbolic methods for knowledge discovery. In: Cohen, H., Lefebvre, C., (eds.) Handbook of Categorization in Cognitive Science. Elsevier, Amsterdam (2005)
11. Pasquier, N., Bastide, Y., Taouil, R., Lakhal, L.: Discovering frequent closed itemsets for association rules. In: ICDT '99: Proceedings of the 7th International Conference on Database Theory, pp. 398–416. Springer, London (1999)
12. Pensa, R., Besson, J., Boulicaut, J.-F.: A methodology for biologically relevant pattern discovery from gene expression data. In: Proceeding 7th International Conference on Discovery Science, LNAI 3245, pp. 230–241. Springer, Padova (Oct 2004)
13. Salleb-Aouissi, A., Vrain, C., Nortet, C.: Quantminer: A genetic algorithm for mining quantitative association rules. In: IJCAI, Hyderabad, India, pp. 1035–1040 (2007)
14. Stoughton, R.B.: Applications of DNA microarrays in biology. Annu. Rev. Biochem. **74**(1), 53–82 (2005)