Springer Proceedings in Mathematics & Statistics

Adriano Polpo
Francisco Louzada
Laura L. R. Rifo
Julio M. Stern
Marcelo Lauretto *Editors*

Interdisciplinary Bayesian Statistics

EBEB 2014

Springer

Springer Proceedings in Mathematics & Statistics

Volume 118

More information about this series at http://www.springer.com/series/10533

Springer Proceedings in Mathematics & Statistics

This book series features volumes composed of select contributions from workshops and conferences in all areas of current research in mathematics and statistics, including operation research and optimization. In addition to an overall evaluation of the interest, scientific quality, and timeliness of each proposal at the hands of the publisher, individual contributions are all refereed to the high quality standards of leading journals in the field. Thus, this series provides the research community with well-edited, authoritative reports on developments in the most exciting areas of mathematical and statistical research today.

Adriano Polpo • Francisco Louzada
Laura L. R. Rifo • Julio M. Stern
Marcelo Lauretto
Editors

Interdisciplinary Bayesian Statistics

EBEB 2014

Springer

Editors
Adriano Polpo
Federal University of Sao Carlos
Sao Carlos
Brazil

Francisco Louzada
University of Sao Paulo
Sao Carlos
Brazil

Laura L. R. Rifo
Campinas State University
Campinas
Brazil

Julio M. Stern
Dept. of Applied Mathematics
University of Sao Paulo Institute of Mathematics and Statistics
Sao Paulo, São Paulo
Brazil

Marcelo Lauretto
School of Arts, Sciences and Humanities
University of Sao Paulo
Sao Paulo, São Paulo
Brazil

ISSN 2194-1009 ISSN 2194-1017 (electronic)
Springer Proceedings in Mathematics & Statistics
ISBN 978-3-319-12453-7 ISBN 978-3-319-12454-4 (eBook)
DOI 10.1007/978-3-319-12454-4
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014956382

Foreword

It is a great pleasure to preface EBEB 2014, the proceedings of the 12th Brazilian Meeting on Bayesian Statistics, which was attended by the ever-growing community of Brazilian researchers in Bayesian Statistics and by colleagues and collaborators from around the world.

From March 10th to 14th at a beautiful resort in Atibaia, EBEB hosted fine presentations in lecture (33) and poster (39) formats of researchers from institutions in Belgium, Canada, Chile, Finland, India, Italy, Peru, Saudi Arabia, Switzerland, UK, USA, and Brazil. The tradition of meeting participants serving as referees for the selection of papers submitted to be included in this book was maintained. This was not an easy task, as the quality of submission was very high. The EBEB 2014 organizers made sure that graduate students had financial support for active engagement in the meeting activities; presenting their research and interacting with researchers from other institutions. This support is always a sound investment for the development of Bayesian Statistics.

High-quality publications of proceedings, as of EBEB 2012 by AIP (The American Institute of Physics) and of EBEB 2014 by Springer, motivate researchers and enrich the EBEBs. Editing such proceedings is a tradition inspired by the Valencia and MaxEnt series, arguably two of the most influential and long-standing meetings in Bayesian Statistics. This requires a lot of time, great effort, and also knowledge and experience.

Another tradition of the Brazilian Bayesian community is to invite illustrious professors who discuss our research and suggest new directions to explore. First and foremost among those were Dev Basu and Dennis Lindley, as I will comment in the sequel. We look forward to maintaining that tradition at the next EBEBs. We also see as a positive sign of maturity, the increasing number of international researchers that spontaneously attend EBEB, as one expects from prestigious international meetings in well-developed areas of scientific research. This trend is a direct consequence of the existence of high-quality proceedings, and vice-versa, exactly as it is in MaxEnt and used to be in Valencia.

The increasing presence at EBEB of several emergent centers of research in Brazil spreads our Bayesian activities outside the Rio and São Paulo axis. Some of these emerging centers are strongly inserted and well connected internationally, but are

also becoming increasingly capable of self-steering, exhibiting the high degree of autonomy that is so characteristic of genuine scientific research.

My good friend and colleague Josemar Rodrigues received a most deserved accolade at this EBEB. Josa's drive for research produces many papers by himself, his students, and colleagues. Having become a good friend of Professor Basu, Josa was also deeply influenced by him in his move from a good frequentist to a superb Bayesian.

In a way, Dev is a Founding Father of Brazilian Bayesianism. His visit to USP in the eighties left seeds which became important Bayesian trees. This process was also caused by the visit of Dennis, who sadly passed away in 2013. The Brazilian Bayesian community owes gratitude to the two gurus and dedicates this volume to their memory.

Fabio Spizzichino and Claudio Macci's paper in this book is a Brazilian contribution to the celebrations of the 250th anniversary of the publication of (ultimately the reason for us to gather at Atibaia) "An Essay towards Solving a Problem in the Doctrine of Chances" by Our Reverend Thomas Bayes.

São Paulo Sergio Wechsler
January 2015

Preface

This volume of the "Springer Proceedings in Mathematics & Statistics" contains selected articles from EBEB 2014. The home page of the event is available at http://www.ime.usp.br/∼isbra/ebeb/ebeb2014.

- The main promoters of EBEB 2014 were
  - ISBrA—Brazilian chapter of the International Society for Bayesian Analysis (ISBA)
  - INCTMat—National Institute of Mathematical Science and Technology
  - Interinstitutional Graduate Program in Statistics of UFSCar/ICMC-USP
  - Graduate Program in Statistics of IME-USP
  - IME-USP—Institute of Mathematics and Statistics of University of São Paulo
- Organizing Committee
  - Adriano Polpo (UFSCar)
  - Carlos Alberto de Bragança Pereira (IME-USP)
  - Franciso Louzada Neto (ICMC-USP)
  - Julio Stern (IME-USP)
  - Laura Letícia Ramos Rifo (UNICAMP)
  - Marcelo Lauretto (EACH-USP)
  - Teresa Cristina Martins Dias (UFSCar)
- Scientific Committee
  - Adriano Polpo (UFSCar / Brazil)
  - Alexandra M. Schmidt (UFRJ / Brazil)
  - André Rogatko (Cedars-Sinai / US)
  - Carlos Alberto de Bragança Pereira (IME-USP / Brazil)
  - Carlos A. R. Diniz (UFSCar / Brazil)
  - Cassio de Campos (IDSIA / Switzerland)
  - Dani Gamerman (IM-UFRJ / Brazil)
  - Debajyoti Sinha (FSU / US)
  - Francisco Louzada Neto (ICMC-USP / Brazil)
  - Julio Stern (IME-USP / Brazil)
  - Laura Letícia Ramos Rifo (IMECC-UNICAMP / Brazil)
  - Marcelo Lauretto (EACH-USP / Brazil)
  - Márcia D'Elia Branco (IME-USP / Brazil)

  - Marcio Alves Diniz (UFSCar / Brazil)
  - Rosangela H. Loschi (UFMG / Brazil)
  - Victor Fossaluza (IME-USP / Brazil)
- Executive committee
  - Bruno Borcado (Supremum Assessoria / Brazil)
  - Lourdes Vaz da Silva Netto (IME-USP / Brazil)
  - Sylvia Regina A. Takahashi (IME-USP / Brazil)

The organizers thank the first president of ISBrA, Professor Sergio Wechsler, for the Foreword and the reviewers for their careful work in selecting the papers. This book is a consequence of the great and hard work of all people involved in the organization of EBEB: colleagues, administrators and the convention center/hotel employers.

We also thank the support of the following promoters:

Contents

Contributors

Plinio L. D. Andrade Institute of Mathematics and Statistics, University of São Paulo, São Paulo, Brazil

Anderson Ara Departamento de Estatística, Universidade Federal de São Carlos, São Carlos, SP, Brazil

Yiqi Bao Department of Statistics, Federal University of São Carlos-UFSCar, São Carlos, SP, Brazil

Leonardo S. Bastos Program for Scientific Computing (PROCC) - Oswaldo Cruz Foundation, RJ, Rio de Janeiro, Brazil

Jorge L. Bazán Instituto de Ciências Matemáticas e de Computação. USP, São Carlos, SP, Brazil

Heleno Bolfarine Instituto de Matemática e Estatística, Universidade de São Paulo, Cidade Universitária - São Paulo, SP, Brazil

Cassio P. de Campos Dalle Molle Institute for Artificial Intelligence, Manno, Switzerland

Queen's University, Belfast, UK

Vicente G. Cancho Institute of Mathematics and Computer Science, University of São Paulo-USP, São Carlos, SP, Brazil

ICMC, University of Sao Paulo, Sao Carlos, Brazil

Luis Carvalho Boston University, Boston, MA, USA

Luis E. Carvalho Department of Mathematics and Statistics, Boston University, Boston, MA, USA

Luiz Max Carvalho Program for Scientific Computing (PROCC) - Oswaldo Cruz Foundation, RJ, Rio de Janeiro, Brazil

Luis M. Castro Department of Statistics, Universidad de Concepción, Concepción, Chile

Mário de Castro Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, SP, Brazil

Dalia Chakrabarty Department of Statistics, University of Warwick, Coventry, UK

Department of Mathematics, University of Leicester, Leicester, UK

Jukka Corander Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland

Giorgio Corani Istituto Dalle Molle di studi sull'Intelligenza Artificiale (IDSIA), Scuola universitaria professionale della Svizzera italiana (SUPSI), Università della Svizzera italiana (USI), Manno, Switzerland

Jasper De Bock Ghent University, SYSTeMS Research Group, Zwijnaarde, Belgium

Gert de Cooman Ghent University, SYSTeMS Research Group, Zwijnaarde, Belgium

Débora Delbem Department of Statistics, Federal University of São Carlos, São Paulo, Brazil

Carlos Diniz Department of Statistics, Federal University of São Carlos, São Paulo, Brazil

Márcio Diniz Ghent University, SYSTeMS Research Group, Zwijnaarde, Belgium

Federal University of Sao Carlos, Sao Carlos, SP, Brazil

Marcio Augusto Diniz Institute of Mathematics and Statistics, University of São Paulo, São Paulo, Brazil

Joao Daniel Nunes Duarte Departamento de Estatistica, ICEx, UFMG, Belo Horizonte, MG, Brazil

Ricardo S. Ehlers Instituto de Ciências Matemáticas e de Computação. USP, São Carlos, SP, Brazil

Víctor Elvira Department of Signal Theory and Communications, Universidad Carlos III de Madrid, Leganés, Spain

Luis Gustavo Esteves Institute of Mathematic and Statistics, University of Sao Paulo, Sao Paulo, Brazil

Guillermo Ferreira Department of Statistics, Universidad de Concepción, Concepción, Chile

José Augusto Fioruci Department of Statistics, Federal University of São Carlos-UFSCar, São Carlos, SP, Brazil

Victor Fossaluza IME-USP, São Paulo, Brazil

Viviana Giampaoli Instituto de Matemática e Estatística, Universidade de São Paulo, Cidade Universitária - São Paulo, SP, Brazil

Hunter Glanz California Polytechnic State University, San Luis Obispo, CA, USA

Jhon F.B. Gonzales Departamento de Estatística, Universidade Federal de São Carlos, Rod, São Carlos, SP, Brazil

Yang Jin Boston University, Boston, MA, USA

Ian Johnston Boston University, Boston, MA, USA

Khader Khadraoui Department of Mathematics and Statistics, Laval University, Quebec City, QC, Canada

Marcelo de Souza Lauretto IME-USP, São Paulo, Brazil

José G. Leite Departamento de Estatística, Universidade Federal de São Carlos, São Carlos, SP, Brazil

Francisco Louzada Institute of Mathematics and Computer Science, University of São Paulo-USP, São Carlos, SP, Brazil

David Luengo Department of Signal Theory and Communications, Universidad Politécnica de Madrid, Madrid, Spain

Claudio Macci Dipartimento di Matematica, Università di Roma Tor Vergata, Roma, Italia

Márcia A. C. Macera DEs, Federal University of Sao Carlos, Sao Carlos, Brazil

Luca Martino Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland

Vinicius Diniz Mayrink Departamento de Estatistica, ICEx, UFMG, Belo Horizonte, MG, Brazil

Brian Alvarez R. de Melo Institute of Mathematic and Statistics, University of Sao Paulo, Sao Paulo, Brazil

Jean Paul Navarrete Department of Statistics, Universidad de Concepción, Concepción, Chile

Melaine Cristina de Oliveira IME-USP, São Paulo, Brazil

Natalia Oliveira Federal University of Sao Carlos, Sao Carlos, SP, Brazil

Elizabeth González Patiño Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, Brazil

Shashi Paul Emerging Technologies Research Centre, De Montfort University, Leicester, UK

Rosineide F. da Paz Universidade Federal de São Carlos, São Carlos, Brazil

Instituto de Ciências Matemáticas e de Computação. USP, São Carlos, SP, Brazil

Lijun Peng Department of Mathematics and Statistics, Boston University, Boston, MA, USA

Carlos A. B. Pereira Instituto de Matemática e Estatística, Universidade de São Paulo, Cidade Universitária - São Paulo, SP, Brazil

Carlos Alberto de Bragança Pereira Institute of Mathematics and Statistics, University of São Paulo, São Paulo, Brazil

Steven Piantadosi Biostatistics and Bioinformatics Research Center, Cedars-Sinai Medical Center, Los Angeles, CA, USA

Adriano Polpo Federal University of Sao Carlos, Sao Carlos, SP, Brazil

Guaraci Requena Institute of Mathematics and Statistics, University of São Paulo, São Paulo, Brazil

Laura L. R. Rifo Institute of Mathematics and Statistics, University of Campinas, Campinas, Brazil

André Rogatko Biostatistics and Bioinformatics Research Center, Cedars-Sinai Medical Center, Los Angeles, CA, USA

Luis E. B. Salasar Departamento de Estatística, Universidade Federal de São Carlos, São Carlos, SP, Brazil

Ricardo Silva Department of Statistical Science and Centre for Computational Statistics and Machine Learning, University College London, London, UK

Andressa Soreira IME-USP, São Paulo, Brazil

Fabio Spizzichino Dipartimento di Matematica G. Castelnuovo, Sapienza Università di Roma, Roma, Italia

Julio Michael Stern Institute of Mathematics and Statistics (IME-USP), University of São Paulo, São Paulo, Brazil

Claudio J. Struchiner Program for Scientific Computing (PROCC) - Oswaldo Cruz Foundation, RJ, Rio de Janeiro, Brazil

Adriano K. Suzuki Instituto de Ciências Matemáticas e de Computação-ICMC, Universidade de São Paulo-Campus de São Carlos, São Carlos, SP, Brazil

Vera. L. D. Tomazella Departamento de Estatística, Universidade Federal de São Carlos, São Carlos, SP, Brazil

Salimeh Yasaei Sekeh Department of Statistics, UFSCar, São Carlos, Brazil

About the Editors

Adriano Polpo is Head (2013—) and Associate Professor (2006—) of Statistics at Universidade Federal de Sao Carlos (UFSCar, Brazil), Brazil. He received his PhD in Statistics from University of Sao Paulo (Brazil). He is President of ISBrA—Brazilian Chapter of the International Society for Bayesian Analysis (2012–2014). Polpo is co-author of more than 25 publications in statistical peer—reviewed journals, books, and book chapters. He has supervised more than 15 PhDs, masters and undergraduates.

Francisco Louzada is a Full Professor of Statistics at the Department of Applied Mathematics and Statistics, University of Sao Paulo (USP, Brazil), Research Productivity Fellow of the Brazilian founding agency CNPq, Level 1B, Director for the Center for Risk Analysis (CER), Deputy Director for the Center for Applied Mathematics and Statistics in Industry (CeMEAI), Director of Technology Transfer and Executive Director of External Relations of the Center for Research, Innovation and Dissemination of Mathematical Science in Industry (CEPID-CeMEAI). He received his PhD. degree in Statistics from the University of Oxford, UK. Louzada is single and joint author of more than 150 publications in statistical peer—reviewed journals, books, and book chapters. He has supervised more than 80 assistant researches, post-docs, PhDs, masters and undergraduates.

Laura Rifo is Doctor Professor (2005—) of Statistics at Universidade Estadual de Campinas—Unicamp, Brazil. She received her PhD in Statistics from University of São Paulo, Brazil. She is Treasurer of ISBrA—Brazilian Chapter of the International Society for Bayesian Analysis (2012–2014). Rifo is co-author of about 15 publications in statistical peer—reviewed journals, books, and book chapters, and more than 40 short films, experiments and softwares in scientific divulgation. She has supervised more than 20 assistant researches, PhDs, masters and undergraduates.

Julio Michael Stern is Full Professor of IME-USP, the Institute of Mathematics and Statistics of the University of Sao Paulo, and Level 1 Research Fellow of CNPq, the Brazilian National Council for Science and Technology. He has a Ph.D. in Operations Research from Cornell University. He was the 2010–2012 President of ISBrA, the Brazilian Chapter of the International Society for Bayesian Analysis, and the Organizer of MaxEnt 2008, the 28th International Workshop on Bayesian Inference

and Maximum Entropy Methods in Science and Engineering. He has published several books and many articles in the areas of Epistemology and Logic; Mathematical Modeling and Operations Research; Statistical Theory and Methods; and Sparse and Structured Systems.

Marcelo de Souza Lauretto is Assistant Professor (2009—) and Vice-Coordinator (2013—) of the Bachelor's Degree in Computer Information Systems, at the School of Arts, Sciences and Humanities of the University of Sao Paulo (USP), Brazil. He received his PhD in Bioinformatics from the University of Sao Paulo. Lauretto is co-author of more than 24 publications in peer—reviewed journals, books, and book chapters. He has supervised more than 18 masters and undergraduates.

Chapter 1
What About the Posterior Distributions When the Model is Non-dominated?

Claudio Macci and Fabio Spizzichino

Abstract Starting from the first inception of philosophical research that had subsequently led to subjective probability and Bayesian statistics, and to date the most recent developments, the probabilistic nature and the related statistical implications of Bayes theorem have been thoroughly discussed. However, the substantial contents of such a formula is very deep and new contributions are still continuing after 250 years. The simplest form of Bayes theorem is met when dominated statistical models are dealt with. This is, in a sense, comfortable, specially as far as parametric models are considered. Actually, most statistical techniques in the frame of parametric inference refer to dominated statistical models. Different problems in the applications, however, can lead to considering non-dominated models. In these cases, some complications and intriguing conclusions can arise. Concerning *non-dominated statistical models*, we devote this note to discussing some mathematical features that may sometimes escape the attention of statisticians. We deal with questions and results that, at a first glance, may appear of almost-exclusive measure-theoretic interest. However, they have a real statistical meaning of their own and the present note aims to stimulate some reflections about this field.

1.1 Introduction

Starting from the first inception of philosophical research that had subsequently led to subjective probability and Bayesian statistics, and to date the most recent developments, the probabilistic nature and the related statistical implications of Bayes theorem have been thoroughly discussed, both from the viewpoints of mathematical formalization and heuristic meanings. However, the substantial contents of such a formula is very deep and new contributions are still continuing after 250 years.

C. Macci (✉)
Dipartimento di Matematica, Università di Roma Tor Vergata , Roma, Italia
e-mail: macci@mat.uniroma2.it

F. Spizzichino
Dipartimento di Matematica G. Castelnuovo, Sapienza Università di Roma, Roma, Italia
e-mail: fabio.spizzichino@uniroma1.it

A. Polpo et al. (eds.), *Interdisciplinary Bayesian Statistics,*
Springer Proceedings in Mathematics & Statistics 118, DOI 10.1007/978-3-319-12454-4_1

The term "Bayes theorem" is in any case connected to the analysis of the relations between the prior and posterior distributions or, in other words, between the state of knowledge available to the analyst before and after a new source of information. However, it can actually take different forms, depending on the specific contexts and different mathematical forms or it may sometimes hide aspects that are relevant from the statistical viewpoint. Thus, new related remarks come out from time to time.

The simplest form of Bayes theorem is met when dominated statistical models are dealt with. This is, in a sense, comfortable, specially as far as parametric models are considered. Actually, most statistical techniques in the frame of parametric inference refer to dominated statistical models.

Different problems in the applications, however, can lead to considering non-dominated models. In these cases, some complications and intriguing conclusions can arise.

Concerning *non-dominated statistical models*, we devote this note to discussing some mathematical features that may sometimes escape the attention of statisticians. We deal with questions and results that, at a first glance, may appear of almost-exclusive measure-theoretic interest. However, they have a real statistical meaning of their own and the present note aims to stimulate some reflections about this field. We hope that our discussion can also be of interest in the field of Bayesian non-parametric analysis where non-dominated models are more common. A well-known example where the statistical experiment is non-dominated is the family of all discrete probability measures on an uncountable set. In this case, we have a conjugate family of prior distributions formed by the Dirichlet processes (see [4]). This is a distinctive feature of Dirichlet processes described in [7], where they take into account the normalization of increasing processes having independent increments used in [11].

From a technical viewpoint, a basic property of dominated models is that the posterior distribution is absolutely continuous w.r.t. the prior distribution. The starting point for our discussion is that such a property can fail in the non-dominated case. We shall then analyze some peculiar aspects of the relations tying the posterior and the prior distributions under non-domination and, in particular, we consider the *absolutely continuous* and *singular* parts of the posterior distribution w.r.t. the prior.

We aim to offer a friendly presentation, as far as possible. But, on the purpose of reaching sound conclusions, a certain amount of notation and technicalities is unavoidable. The discussion will be articulated as follows. In Sect. 2, we recall some basic definitions about statistical models and introduce the notation that will be used in the subsequent sections. Section 3 contains a brief review about notions of domination and non-domination in the statistical models, from different viewpoints. Section 4 will be devoted to the implications of non-domination in Bayesian statistics. Finally, in Sect. 5, we present a brief discussion based on a special class of non-dominated parametric models.

1.2 Basic Definitions and Notation

We recall here the basic terminology and definitions related with Bayesian statistical experiments and fix the notation that will be used in the following. As a reference we mainly rely, e.g., on [5].

In a statistical model, the *sample space* and the *parameter space* will be denoted by $\mathcal{X}$ and $\mathcal{G}$. $\mathcal{B}_\mathcal{X}, \mathcal{B}_\mathcal{G}$ are the σ-algebras over which probability measures are respectively defined. In particular, we consider a *statistical experiment* $(\mathcal{X}, \mathcal{B}_\mathcal{X}, \mathcal{P})$, where $\mathcal{P} = \{P_\theta : \theta \in \mathcal{G}\}$ is the family of *sampling distributions* with P_θ being probability measures over $(\mathcal{X}, \mathcal{B}_\mathcal{X})$.

Starting from $(\mathcal{X}, \mathcal{B}_\mathcal{X}, \mathcal{P})$, one can define infinitely many different Bayesian experiments. For any given probability measure μ on $(\mathcal{G}, \mathcal{B}_\mathcal{G})$, in fact, we can consider the *Bayesian experiment* $(\mathcal{G} \times \mathcal{X}, \mathcal{B}_\mathcal{G} \otimes \mathcal{B}_\mathcal{X}, \Pi_{\mu,\mathcal{P}})$ where $\mathcal{G} \times \mathcal{X}$ is the product space

$$\mathcal{G} \times \mathcal{X} = \{(\theta, x) \,|\theta \in \mathcal{G}; x \in \mathcal{X}\},$$

$\mathcal{B}_\mathcal{G} \otimes \mathcal{B}_\mathcal{X}$ denotes the *product σ-algebra,* generated by the family of all the subsets of the form

$$A \times B = \{(\theta, x) \,|\theta \in A, x \in B\} \text{ for all } A \in \mathcal{B}_\mathcal{G} \text{ and } B \in \mathcal{B}_\mathcal{X},$$

and $\Pi_{\mu,\mathcal{P}}$ is the probability measure, over $(\mathcal{G} \times \mathcal{X}, \mathcal{B}_\mathcal{G} \otimes \mathcal{B}_\mathcal{X})$, defined by the position

$$\Pi_{\mu,\mathcal{P}}(A \times B) = \int_A P_\theta(B)\mu(d\theta) \text{ (for all } A \in \mathcal{B}_\mathcal{G} \text{ and } B \in \mathcal{B}_\mathcal{X}). \tag{1.1}$$

$\Pi_{\mu,\mathcal{P}}$ is then the joint distribution of the pair (*parameter-observation*).

The measure μ is the marginal distribution of $\Pi_{\mu,\mathcal{P}}$ on $(\mathcal{G}, \mathcal{B}_\mathcal{G})$, and it is seen then as a *prior distribution* for the parameter.

The marginal distribution of $\Pi_{\mu,\mathcal{P}}$ over $(\mathcal{X}, \mathcal{B}_\mathcal{X})$, namely

$$P_{\mu,\mathcal{P}}(B) = \Pi_{\mu,\mathcal{P}}(\mathcal{G} \times B) = \int_\mathcal{G} P_\theta(B)\mu(d\theta) \text{ (for all } B \in \mathcal{B}_\mathcal{X}), \tag{1.2}$$

is the *predictive distribution* of the observations.

Finally, the family of *posterior distributions* is the family of the probability measures $\{\mu_\mathcal{P}(\cdot\,|x) : x \in \mathcal{X}\}$ over $(\mathcal{G}, \mathcal{B}_\mathcal{G})$ defined by the equation

$$\Pi_{\mu,\mathcal{P}}(A \times B) = \int_B \mu_\mathcal{P}(A|x)P_{\mu,\mathcal{P}}(dx) \text{ (for all } A \in \mathcal{B}_\mathcal{G} \text{ and } B \in \mathcal{B}_\mathcal{X}). \tag{1.3}$$

We remark the analogy between (1.1) and (1.3) where we have the integrals of conditional probabilities with respect to marginal distributions. We typically consider situations where the family $\{\mu_\mathcal{P}(\cdot\,|x) : x \in \mathcal{X}\}$ exists and it is unique $P_{\mu,\mathcal{P}}$ almost surely with respect to x.

In view of what follows, we remark that for all $C \in \mathcal{B}_\mathcal{G} \otimes \mathcal{B}_\mathcal{X}$, we have

$$\Pi_{\mu,\mathcal{P}}(C) = \int_\mathcal{G} P_\theta(C(\theta, .))\mu(d\theta) = \int_\mathcal{X} \mu_\mathcal{P}(C(., x)|x)P_{\mu,\mathcal{P}}(dx), \tag{1.4}$$

where

$$C(\theta,.) = \{x \in \mathcal{X} : (\theta, x) \in C\} \text{ and } C(.,x) = \{\theta \in \mathcal{G} : (\theta, x) \in C\} \tag{1.5}$$

are the θ-section and the x-section of the set C. Notice that, when $C = A \times B$ (for $A \in \mathcal{B}_{\mathcal{G}}$ and $B \in \mathcal{B}_{\mathcal{X}}$), the equations in (1.4) reduce to (1.1) and (1.3), respectively.

Associated to a given Bayesian experiment $(\mathcal{G} \times \mathcal{X}, \mathcal{B}_{\mathcal{G}} \otimes \mathcal{B}_{\mathcal{X}}, \Pi_{\mu,\mathcal{P}})$ we can also consider over $(\mathcal{G} \times \mathcal{X}, \mathcal{B}_{\mathcal{G}} \otimes \mathcal{B}_{\mathcal{X}})$ the *product measure* $\mu \otimes P_{\mu,\mathcal{P}}$, i.e., the product of the marginal distributions μ and $P_{\mu,\mathcal{P}}$, which is defined by the equation

$$\mu \otimes P_{\mu,\mathcal{P}}(A \times B) = \mu(A) P_{\mu,\mathcal{P}}(B) \text{ (for all } A \in \mathcal{B}_{\mathcal{G}} \text{ and } B \in \mathcal{B}_{\mathcal{X}}). \tag{1.6}$$

Typically, we think of the cases where the sample space $\mathcal{X}$ coincides with a subset of h-dimensional Euclidean space $\mathbb{R}^h$. As to $\mathcal{G}$, we typically use the term *parametric case* when $\mathcal{G} \subset \mathbb{R}^k$ for some finite integer number k.

No problem however arises in letting $\mathcal{X}$ and $\mathcal{G}$ to be *Polish spaces* (i.e., complete and separable metric spaces), equipped with their Borel σ-algebras $\mathcal{B}_{\mathcal{G}}$ and $\mathcal{B}_{\mathcal{X}}$. Such a generalization permits us to extend everything we say here to more general cases for both statistical observations and parameter. In particular, it allows one to consider *non-parametric models*.

From a technical viewpoint, all the functions $\{\theta \mapsto P_\theta(B) : B \in \mathcal{B}_{\mathcal{X}}\}$ in the formula (1.1) are tacitly assumed to be measurable with respect to $\mathcal{B}_{\mathcal{G}}$ in order to guarantee the condition of measurable dependence which is needed to let the integral (1.2) be correctly defined. Notice that this condition is not generally required in Statistics: it is just peculiar of the Bayesian approach, where the joint probability measure $\Pi_{\mu,\mathcal{P}}$ has a fundamental role.

1.3 Bayesian Versus Non-Bayesian Dominated Models

We now recall the concepts of domination for statistical models. Different definitions have been introduced in the literature, but for our purposes it will be enough to compare just the Bayesian concept with the standard non-Bayesian one.

In a non-Bayesian frame, a statistical experiment determined by $\mathcal{P} = \{P_\theta : \theta \in \mathcal{G}\}$ is *dominated* if there exists a σ-finite measure λ on $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$ with respect to which all the sampling distributions are absolutely continuous. Namely, for $B \in \mathcal{B}_{\mathcal{X}}$, one has the implication

$$\lambda(B) = 0 \Rightarrow P_\theta(B) = 0 \text{ (for all } \theta \in \mathcal{G}).$$

When the experiment is dominated, we have a family of densities $\left\{\frac{dP_\theta}{d\lambda} : \theta \in \mathcal{G}\right\}$, with respect to λ and, as a basic implication of this condition, it is then possible to talk about the *likelihood function*.

We recall in fact that, when ν_1, ν_2 are two σ-finite measures on a same measurable space $(\Omega, \mathcal{F})$ with ν_1 absolutely continuous with respect to ν_2, then the *Radon–Nikodym theorem* ensures the existence of a *density* $\frac{d\nu_1}{d\nu_2}$, namely of a $\mathcal{F}$-measurable function such that

$$\nu_1(A) = \int_A \frac{d\nu_1}{d\nu_2}(\omega)\nu_2(d\omega) \text{ (for all } A \in \mathcal{F}).$$

More generally we have

$$\nu_1(A) = \nu_1^{(ac)}(A) + \nu_1^{(sg)}(A) \text{ (for all } A \in \mathcal{F}),$$

where $\nu_1^{(ac)}$ is absolutely continuous w.r.t. ν_2, and $\nu_1^{(sg)}$ and ν_2 are mutually singular. Namely, we have

$$\nu_1^{(ac)}(A) = \int_A \frac{d\nu_1^{(ac)}}{d\nu_2}(\omega)\nu_2(d\omega) \text{ (for all } A \in \mathcal{F})$$

for a density $\frac{d\nu_1^{(ac)}}{d\nu_2}$ (by the Radon–Nikodym theorem), and

$$\nu_1^{(sg)}(A) = \nu_1^{(sg)}(A \cap D) \text{ (for all } A \in \mathcal{F})$$

for a set $D \in \mathcal{F}$ such that $\nu_2(D) = 0$. The pair $(\nu_1^{(ac)}, \nu_1^{(sg)})$ is uniquely determined by ν_1 and ν_2. Obviously, if ν_1 absolutely continuous with respect to ν_2, $\nu_1 = \nu_1^{(ac)}$ and $\nu_1^{(sg)}$ is the null measure.

Notice that the above definition of dominated experiment only concerns the family of sampling distributions. The concept of domination in the Bayesian setting, on the contrary, refers to the pair $(\mu, \mathcal{P})$ and it is seen as a property of the joint probability measure $\Pi_{\mu,\mathcal{P}}$, over $(\mathcal{G} \times \mathcal{X}, \mathcal{B}_\mathcal{G} \otimes \mathcal{B}_\mathcal{X})$. The product measure $\mu \otimes P_{\mu,\mathcal{P}}$ also enters into play. More precisely, one can formulate the following.

Definition (see [5, p. 28])
The experiment $(\mathcal{G} \times \mathcal{X}, \mathcal{B}_\mathcal{G} \otimes \mathcal{B}_\mathcal{X}, \Pi_{\mu,\mathcal{P}})$ is *dominated* when

$$\Pi_{\mu,\mathcal{P}} \text{ is absolutely continuous with respect to } \mu \otimes P_{\mu,\mathcal{P}}. \tag{1.7}$$

In other words

$$\mu \otimes P_{\mu,\mathcal{P}}(C) = 0 \Rightarrow \Pi_{\mu,\mathcal{P}}(C) = 0.$$

We remark that, by (1.4) and (1.5), $\Pi_{\mu,\mathcal{P}}(C) = 0$ is equivalent to each of both conditions:

$$\mu(\{\theta \in \mathcal{G} : P_\theta(C(\theta, .)) = 0\}) = 1; \tag{1.8}$$

$$P_{\mu,\mathcal{P}}(\{x \in \mathcal{X} : \mu_\mathcal{P}(C(., x)|x) = 0\}) = 1. \tag{1.9}$$

The condition (1.7) guarantees that the posterior distributions $\{\mu_{\mathcal{P}}(\cdot\,|x) : x \in \mathcal{X}\}$ are absolutely continuous w.r.t. the prior distribution μ and also that the sampling distributions $\{P_\theta : \theta \in \mathcal{G}\}$ are such w.r.t. the predictive distribution $P_{\mu,\mathcal{P}}$. See also, in this respect, Sect. 4.

In particular, for all choices of the prior distribution μ, the density of $\mu_{\mathcal{P}}(\cdot\,|x)$ w.r.t. μ is just given ($P_{\mu,\mathcal{P}}$ almost surely with respect to x) by the Bayes formula:

$$\mu_{\mathcal{P}}(A|x) = \frac{\int_A \frac{dP_\theta}{d\lambda}(x)\mu(d\theta)}{\int_{\mathcal{G}} \frac{dP_\theta}{d\lambda}(x)\mu(d\theta)} \text{ (for all } A \in \mathcal{B}_{\mathcal{G}}). \tag{1.10}$$

On the contrary, when the statistical experiment is non-dominated, we cannot rely on the formula (1.10) anymore and the posterior distributions can have singular parts with respect to the prior distribution.

Typically, parametric inference problems are modeled on dominated statistical experiments (the reader can think of the different examples based on exponential families), while non-dominated statistical experiments can emerge more naturally for nonparametric inference problems. Some examples of non-dominated statistical experiments will be briefly discussed in the last section of this note.

The concept of dominated statistical experiment appeared in the literature between the forties and the fifties of last century (see, e.g., [1] and [6]) as a strong condition of regularity for an experiment. In particular, it plays a crucial role to give a characterization of classical sufficiency. We recall that, given a statistical experiment $\mathcal{P} = \{P_\theta : \theta \in \mathcal{G}\}$, a σ-algebra $\mathcal{S} \subset \mathcal{B}_{\mathcal{X}}$ is *sufficient* when the conditional probabilities $\{P_\theta(A|\mathcal{S}) : A \in \mathcal{B}_{\mathcal{X}}\}$ do not depend on θ. Then, if $\mathcal{P} = \{P_\theta : \theta \in \mathcal{G}\}$ is dominated, as very well-known, $\mathcal{S} \subset \mathcal{B}_{\mathcal{X}}$ is sufficient if and only if, for all $\theta \in \mathcal{G}$, the Neyman factorization

$$\frac{dP_\theta}{d\lambda}(x) = h(x)k(x,\theta)$$

holds, for some $\mathcal{B}_{\mathcal{X}}$ measurable function h and $\mathcal{S} \otimes \mathcal{B}_{\mathcal{G}}$ measurable function k. We also recall that $\mathcal{S} \subset \mathcal{B}_{\mathcal{X}}$ is Bayes-sufficient if, for any prior distribution μ, there exists a family of the posterior distributions $\{\mu(\cdot\,|x) : x \in \mathcal{X}\}$ such that $\{x \mapsto \mu(A|x) : A \in \mathcal{B}_{\mathcal{G}}\}$ are measurable with respect to $\mathcal{S}$. For models with conditionally i.i.d. observations, the existence of a fixed-dimensional sufficient statistics is equivalent to the existence of a finitely parametrized conjugate family of priors (see, e.g., Sect. 9.3 in [3]). Finitely parametrized conjugate families, in a "weak" sense emerge in more general situations for which sufficiency in the common sense cannot be defined. This is the case of the very well-known *Kalman filter* and, more generally, in some situations of *stochastic filtering* in discrete time (see, e.g., the discussion in [12]). The concept of conjugate families, in the weak sense, could also be introduced in the analysis of non-dominated parametric models.

1.4 Non-dominated Bayesian Experiments and the Lebesgue Decomposition of $\Pi_{\mu,\mathcal{P}}$

As an immediate implication of the above definitions, a Bayesian experiment can be non-dominated only if we can find $C \in \mathcal{B}_{\mathcal{G}} \otimes \mathcal{B}_{\mathcal{X}}$ such that $\mu \otimes P_{\mu,\mathcal{P}}(C) = 0$ and $\Pi_{\mu,\mathcal{P}}(C) > 0$, i.e., (1.8) and (1.9) fail. Namely, it is possible that the statistical observation carries information of deterministic type about parameters, giving rise to the possibility that the posterior distribution contains some singular components w.r.t. the prior distribution. In such cases, one can be then interested in analyzing the Lebesgue decomposition of $\Pi_{\mu,\mathcal{P}}$ with respect to $\mu \otimes P_{\mu,\mathcal{P}}$. Also the decomposition of $\{\mu_{\mathcal{P}}(\cdot|x) : x \in \mathcal{X}\}$ w.r.t. μ and the one of the sampling distributions $\{P_\theta : \theta \in \mathcal{G}\}$ w.r.t. the predictive distribution $P_{\mu,\mathcal{P}}$ are objects of interest and the analysis of the relations among these decompositions is in order, in this frame.

In this respect, a precise result allows us to analyze the type of such relations ([8], Proposition 1; see also Proposition 2 for a different formulation). This result gives in fact more insight on the Lebesgue decompositions of the posterior distributions w.r.t. the prior distribution, and on the Lebesgue decompositions of the sampling distributions w.r.t. the predictive distribution. On this purpose we write down the Lebesgue decomposition of $\Pi_{\mu,\mathcal{P}}$ w.r.t. $\mu \otimes P_{\mu,\mathcal{P}}$, namely

$$\Pi_{\mu,\mathcal{P}}(E) = \int_E g_{\mu,\mathcal{P}}(\theta, x)\mu \otimes P_{\mu,\mathcal{P}}(d\theta, dx) + \Pi_{\mu,\mathcal{P}}(E \cap D_{\mu,\mathcal{P}}), \tag{1.11}$$

for all $E \in \mathcal{B}_{\mathcal{G}} \otimes \mathcal{B}_{\mathcal{X}}$, where $g_{\mu,\mathcal{P}}$ is the density of the absolutely continuous part of $\Pi_{\mu,\mathcal{P}}$ w.r.t. $\mu \otimes P_{\mu,\mathcal{P}}$, and $D_{\mu,\mathcal{P}}$ is a set such that $\mu \otimes P_{\mu,\mathcal{P}}(D_{\mu,\mathcal{P}}) = 0$ (for instance $D_{\mu,\mathcal{P}}$ is the support of the singular part of $\Pi_{\mu,\mathcal{P}}$). We remark that the decomposition in (1.11) depends on the choice of the prior distribution μ, and therefore both $g_{\mu,\mathcal{P}}$ and $D_{\mu,\mathcal{P}}$ depend on the choice of μ. The objects appearing in (1.11) are the basic ingredients in the analysis of the Lebesgue decompositions cited above. The precise result reads as follows.

Theorem 1.1 *(i) The Lebesgue decomposition of $\mu_{\mathcal{P}}(\cdot|x)$ with respect to μ is ($P_{\mu,\mathcal{P}}$ almost surely with respect to x)*

$$\mu_{\mathcal{P}}(A|x) = \int_A g_{\mu,\mathcal{P}}(\theta, x)\mu(d\theta) + \mu_{\mathcal{P}}(A \cap D_{\mu,\mathcal{P}}(., x)|x) \text{ (for all } A \in \mathcal{B}_{\mathcal{G}}).$$

(ii) The Lebesgue decomposition of P_θ with respect to $P_{\mu,\mathcal{P}}$ is (μ almost surely with respect to θ)

$$P_\theta(B) = \int_B g_{\mu,\mathcal{P}}(\theta, x)P_{\mu,\mathcal{P}}(dx) + P_\theta(B \cap D_{\mu,\mathcal{P}}(\theta, .)) \text{ (for all } B \in \mathcal{B}_{\mathcal{X}}).$$

Roughly speaking the main message concerning these decompositions goes as follows:

> " Any absolutely continuous part depends on the other absolutely continuous parts only, and any singular part depends on the other singular parts only"

As an immediate consequence of this result (see Corollary in [8]), condition (1.7) is equivalent to each of the two conditions

$$\mu(\{\theta \in \mathcal{G} : P_\theta \text{ absolutely continuous with respect to } P_{\mu,\mathcal{P}}\}) = 1, \tag{1.12}$$

and

$$P_{\mu,\mathcal{P}}(\{x \in \mathcal{X} : \mu_{\mathcal{P}}(\cdot\,|x) \text{ absolutely continuous with respect to } \mu\}) = 1. \tag{1.13}$$

The following statements point out the connection between the concepts of dominations presented above:

1. Whatever is μ, (1.7) holds if $\mathcal{P}$ is dominated;
2. Whatever is $\mathcal{P}$, (1.7) holds if μ is discrete (i.e., μ is concentrated on an at most countable set).

The statement 1 is well-known and can be seen as a consequence of Theorem 1.2.3 in [5]. However, it can also be seen as a consequence of the equivalence between (1.7) and (1.13), together with the Bayes formula (1.10). The statement 2, in its turn, is a consequence of the equivalence between (1.7) and (1.12). It contains a sound statistical implication: condition (1.7) can hold or can fail, according to the different choices of the prior distribution μ.

1.5 Mixed Distributions for Additive Noise and Related Decompositions

As mentioned in the introduction, non-dominated statistical models are natural in the non-parametric setting. As far as the parametric case is concerned, they are not so common, on the contrary. Non-dominated experiments can however emerge, in a very natural case, in the frame of models with an additive noise. This happens when the probability law of the noise is a mixture between a discrete and a continuous distribution.

Let $\mathcal{G} \subset \mathbb{R}^k$ be a finite-dimensional parameter space and assume that it is uncountable (otherwise we would have domination in any case). Moreover, under each P_θ (for $\theta \in \mathcal{G}$), the observable random variable X has the form

$$X = K(\theta) + \varepsilon, \tag{1.14}$$

where K is a measurable function defined on $\mathcal{G}$ with an uncountable range, and ε is interpreted as the (random) *noise*. The distribution function F_ε of ε has the form

$$F_\varepsilon(y) = pF_d(y) + (1-p)F_{ac}(y) \tag{1.15}$$

where F_d and F_{ac} are discrete and absolutely continuous distributions functions, respectively, and $p \in (0,1)$.

The specific form of Eq. (1.11) and the message contained in Theorem 1.1 can be discussed for this class of models. For the sake of simplicity, we can think of the case where $\mathcal{G} = \mathcal{X} = \mathbb{R}$, K in (1.14) is the identity function, and F_{ac} and F_d in (1.15) are

$$F_{ac}(y) = \int_{-\infty}^{y} f(x)dx \text{ (for all } y \in \mathbb{R}) \tag{1.16}$$

for a continuous density f, and

$$F_d(y) = \begin{cases} 1 & \text{if } y \geq 0 \\ 0 & \text{if } y < 0. \end{cases} \tag{1.17}$$

Namely, F_d is the distribution function of the (degenerate) random variable taking value 0 with probability one. We have in mind a situation where, with positive probabilities $(1 - p)$ and p, the noise can respectively be present or missing in an observation, but such a circumstance is not directly detectable for the experimenter. Therefore, the statistical observation can carry a piece of deterministic information about the unobservable value θ. This example actually demonstrates some main aspects of the arguments in the previous section. A slight different presentation of this example can be found in [8, Sect. 4]. Related with example, see also [2, p. 30] for a discussion on the "likelihood principle" and Macci and Polettini [10, Sect. 3] for the study of the Bayes factor in a hypothesis testing problem.

We start with the expression of the sampling distributions that, for the present model, is given by

$$P_\theta(B) = p1_B(\theta) + (1 - p) \int_B f(x - \theta)dx \text{ (for all } B \in \mathcal{B}_\mathcal{X}). \tag{1.18}$$

In our analysis, the case of interest is when the prior distribution μ is absolutely continuous, with density m. Then, as one could expect, the posterior distribution $\mu(\cdot\,|x)$ should asses a positive probability to x (and only to x) provided $m(x) > 0$. In fact there is the possibility that the observation has not been affected by any noise, and thus the sampled value x coincides with θ. This intuition is confirmed by the expression of the posterior distribution presented in Eq. (1.19) below, which assesses to the value x the positive probability in (1.20). In order to derive formula (1.19), the expression of the predictive distribution is preliminarily needed. By (1.2) and some manipulations we get

$$P_{\mu,\mathcal{P}}(B) = \int_B \left\{ pm(x) + (1 - p) \int_\mathcal{G} f(x - \theta)m(\theta)d\theta \right\} dx \text{ (for all } B \in \mathcal{B}_\mathcal{X}).$$

Moreover, combining (1.3) and the latter expression, one can obtain

$$\mu(A|x) = \frac{pm(x)1_A(x) + (1 - p) \int_A f(x - \theta)m(\theta)d\theta}{pm(x) + (1 - p) \int_\mathcal{G} f(x - \theta)m(\theta)d\theta} \text{ (for all } A \in \mathcal{B}_\mathcal{G}). \tag{1.19}$$

Note that the singular components of the sampling distributions (1.18) reflect into the presence of the singular components for the posterior distributions. In fact the term $p1_B(\theta)$ in (1.18) triggers the term

$$\frac{pm(x)1_A(x)}{pm(x)+(1-p)\int_{\mathcal{G}} f(x-\theta)m(\theta)d\theta}$$

in (1.19). In particular, for $A=\{x\}$, we have

$$\mu(\{x\}|x)=\frac{pm(x)}{pm(x)+(1-p)\int_{\mathcal{G}} f(x-\theta)m(\theta)d\theta}>0 \text{ when } m(x)>0. \quad (1.20)$$

Interestingly, this probability is influenced by the behavior of the prior density m over all the parameter space $\mathcal{G}$.

We now turn to the Lebesgue decomposition of $\Pi_{\mu,\mathcal{P}}$ with respect to $\mu\otimes P_{\mu,\mathcal{P}}$. We remind that such decomposition is generally given by the formula (1.11). In the present case, the ingredients appearing in such formula respectively become

$$g_{\mu,\mathcal{P}}(\theta,x)=\frac{(1-p)f(x-\theta)}{pm(x)+(1-p)\int_{\mathcal{G}} f(x-\eta)m(\eta)d\eta}$$

and

$$D_{\mu,\mathcal{P}}=\{(\theta,x)\in\mathcal{G}\times\mathcal{X}|\theta=x\}.$$

We said above that the statistical observation can carry some piece of " deterministic" information about the parameter (the term $p1_B(\theta)$ in Eq. (1.18) above) and, when the sample value is x, we must asses positive probability to x when $m(x)$ is positive (see Eq. (1.20) above). This happens because the statistical observation can coincide with the parameter, and therefore the structure of the set $D_{\mu,\mathcal{P}}$ is not surprising.

Let us now focus attention on what happens when we have $n\geq 2$ statistical observations $X_1,\ldots,X_n$ conditionally independent and identically distributed, given θ. Thus, we can write

$$X_i=\theta+\varepsilon_i \text{ (for } i=1,\ldots,n)$$

with $\varepsilon_1,\ldots,\varepsilon_n$ i.i.d. with distribution function F_ε defined by (1.15), (1.16), and (1.17). In particular, we have $P(\varepsilon_i=0)=p$ for all $i=1,\ldots,n$.

In view of what follows, for any possible vector of sampled values $\underline{x}=(x_1,\ldots,x_n)\in\mathcal{X}^n$, we can consider the distinct values $x_{i_1},\ldots,x_{i_k}$, say, for some $\{i_1,\ldots,i_k\}\subset\{1,\ldots,n\}$, with their respective multiplicities. Moreover, let $C_n^{(1)}$ be the subset of vectors having all distinct entries, i.e., n distinct values with multiplicity 1, and let $C_n^{(2)}$ be the subset of vectors having at most one repetitions, i.e., $k<n$ distinct values and only one entry $\hat{x}$, say, (among $(x_1,\ldots,x_n)$) has multiplicity strictly greater than 1. Then, if we denote the predictive distribution for the case of n observations by $P_{\mu,\mathcal{P}_n}$, it is possible to check by some computations that

$$P_{\mu,\mathcal{P}_n}\left(C_n^{(1)}\cup C_n^{(2)}\right)=1.$$

Moreover, as far as the posterior distributions $\{\mu_{\mathcal{P}_n}(\cdot\,|\underline{x}):\underline{x}\in\mathcal{X}^n\}$ are concerned, it is possible to prove that we have the following situations:

1. If $\underline{x} \in C_n^{(1)}$, then the posterior distribution assesses an absolutely continuous part (w.r.t. the prior distribution) and some probability masses on the distinct values $x_1, \ldots, x_n$;
2. If $\underline{x} \in C_n^{(2)}$, then the posterior distribution assesses probability 1 to the value $\hat{x}$.

One could expect that, for an increasing number of observations, the probability increases to observe repetitions in the sample. Such repetitions make the posterior distributions to collapse and degenerate into the repeated value, as stated in item 2 above. A closer inspection of the structure of the model reveals that

$$P_{\mu,\mathcal{P}_n}\left(C_n^{(2)}\right) = 1 - P_{\mu,\mathcal{P}_n}\left(C_n^{(1)}\right) = 1 - [(1-p)^n + np(1-p)^{n-1}],$$

which is nondecreasing with n. Thus

$$P_{\mu,\mathcal{P}_n}\left(\{\underline{x} \in \mathcal{X}^n : \mu_{\mathcal{P}_n}(\cdot\,|\underline{x}) \text{ and } \mu \text{ are mutually singular}\}\right)$$

is nondecreasing with n. This result also holds for more general situations (see, e.g., Eq. (4.2) in [9]).

References

1. Bahadur, R.R.: Sufficiency and statistical decision functions. Ann. Math. Stat. **25**, 423–462 (1954)
2. Berger, J.O., Wolpert, R.L.: The Likelihood Principle, 2nd edn. , IMS Lecture Notes, Hayward (1988)
3. DeGroot, M.H.: Optimal Statistical Decisions. McGraw-Hill, New York (1970)
4. Ferguson, T.S.: A Bayesian analysis of some nonparametric problems. Ann. Stat. **1**, 209–230 (1973)
5. Florens, J., Mouchart, M., Rolin, J.: Elements of Bayesian Statistics. Marcel Dekker, New York (1990)
6. Halmos, P.R., Savage, L.J.: Application of the Radon–Nikodym theorem to the theory of sufficient statistics. Ann. Math. Stat. **20**, 225–241 (1949)
7. James, L.F., Lijoi, A., Prünster, I.: Conjugacy as a distinctive feature of the Dirichlet process. Scand. J. Stat. **33** (2006), 105–120 (2006)
8. Macci, C.: On the Lebesgue decomposition of the posterior distribution with respect to the prior in regular Bayesian experiments. Stat. Probab. Lett. **26**, 147–152 (1996)
9. Macci, C.: On Bayesian experiments related to a pair of statistical observations independent conditionally on the parameter. Stat. Decisions **16**, 47–63 (1998)
10. Macci, C., Polettini, S.: Bayes factor for non-dominated statistical models. Stat. Probab. Lett. **53**, 79–89 (2001)
11. Regazzini, E., Lijoi, A., Prünster, I.: Distributional results for means of normalized random measures with independent increments. Ann. Stat. **31**, 560–585 (2003)
12. Runggaldier, W.J., Spizzichino, F.: Finite-dimensionality in discrete-time nonlinear filtering from a Bayesian statistics viewpoint. In: Germani, A. Stochastic Modelling and Filtering (Rome, 1984) (Lecture Notes in Control and Information Science No. 91), 161–184. Springer, Berlin (1987)

Chapter 2
Predictive Inference Under Exchangeability, and the Imprecise Dirichlet Multinomial Model

Gert de Cooman, Jasper De Bock and Márcio Diniz

Abstract Coherent reasoning under uncertainty can be represented in a very general manner by coherent sets of desirable gambles. In this framework, and for a given finite category set, coherent predictive inference under exchangeability can be represented using Bernstein coherent cones of multivariate polynomials on the simplex generated by this category set. This is a powerful generalisation of de Finetti's representation theorem allowing for both imprecision and indecision. We define an inference system as a map that associates a Bernstein coherent cone of polynomials with every finite category set. Many inference principles encountered in the literature can then be interpreted, and represented mathematically, as restrictions on such maps. We discuss two important inference principles: representation insensitivity—a strengthened version of Walley's representation invariance—and specificity. We show that there is a infinity of inference systems that satisfy these two principles, amongst which we discuss in particular the inference systems corresponding to (a modified version of) Walley and Bernard's imprecise Dirichlet multinomial models (IDMMs) and the Haldane inference system.

2.1 Introduction

This chapter deals with predictive inference for categorical variables. We are therefore concerned with a (possibly infinite) sequence of variables X_n that assume values in some finite set of categories A. After having observed a number $\check{n}$ of them, and having found that, say $X_1 = x_1$, $X_2 = x_2$, ... , $X_{\check{n}} = x_{\check{n}}$, we consider some subject's belief

G. de Cooman (✉) · J. De Bock · M. Diniz
Ghent University, SYSTeMS Research Group, Technologiepark–Zwijnaarde 914,
9052 Zwijnaarde, Belgium,
e-mail: gert.decooman@UGent.be

J. De Bock
e-mail: jasper.debock@UGent.be

M. Diniz
e-mail: marcio.diniz@UGent.be

A. Polpo et al. (eds.), *Interdisciplinary Bayesian Statistics,*
Springer Proceedings in Mathematics & Statistics 118, DOI 10.1007/978-3-319-12454-4_2

model for the next $\hat{n}$ variables $X_{\check{n}+1}, \ldots X_{\check{n}+\hat{n}}$. In the probabilistic tradition—and we want to build on this tradition in the context of this chapter—this belief can be modelled by some conditional predictive probability mass function $p^{\hat{n}}(\cdot\,|x_1, \ldots, x_{\check{n}})$ on the set $A^{\hat{n}}$ of possible values for these next variables. These probability mass functions can be used for prediction or estimation, for statistical inferences, and in decision making involving the uncertain values of these variables. In this sense, predictive inference lies at the heart of statistics, and of learning under uncertainty.

What connects these predictive probability mass functions for various values of $\check{n}$, $\hat{n}$ and $(x_1, \ldots, x_{\check{n}})$ are the requirements of *temporal consistency* and *coherence*. The former requires that when $n_1 \leq n_2$, $p^{n_1}(\cdot\,|x_1, \ldots, x_{\check{n}})$ can be obtained from $p^{n_2}(\cdot\,|x_1, \ldots, x_{\check{n}})$ through marginalisation; the latter essentially demands that these conditional probability mass functions should be connected with temporally consistent unconditional probability mass functions through Bayes's Rule.

A common assumption about the variables X_n is that they are *exchangeable*. De Finetti's famous representation theorem [4, 11] then states that the temporally consistent and coherent conditional and unconditional predictive probability mass functions associated with a countably infinite exchangeable sequence of variables in A are completely characterised by[1] a unique probability measure on the Borel sets of the simplex of all probability mass functions on A, called its *representation*.

This leads us to the central problem of predictive inference: since there is an infinity of such probability measures on the simplex, which one does a subject choose in a particular context, and how can a given choice be motivated and justified? The subjectivists of de Finetti's persuasion would answer that this question needs no answer: a subject's personal predictive probabilities are entirely his, and temporal consistency and coherence are the only requirements he should heed. Proponents of the logicist approach to predictive inference would try enunciating general inference principles in order to narrow down, and hopefully eliminate entirely, the possible choices for the representing probability measures on the simplex. Our point of view holds a compromise between the subjectivist and logicist positions: it should be possible for a subject to make assessments for certain predictive probabilities, and to combine these with certain inference principles he finds reasonable. Although this is not the topic of the present chapter, the inference systems we introduce in Sect. 2.6 provide an elegant framework and tools for making conservative predictive inferences that combine (local) subjective probability assessments with (general) inference principles.

This idea of *conservative probabilistic inference* brings us to a central idea in de Finetti's approach to probability [13]: a subject should be able to make certain probability assessments, and we can then consider these as bounds on so-called precise probability models. Calculating such most conservative but tightest bounds is indeed what de Finetti's fundamental theorem of prevision [13, 19] is about. The theory of imprecise probabilities [25, 28, 30] looks at conservative probabilistic inference precisely in this way: how can we calculate as efficiently as possible the

[1] ... unless the observed sequence has probability zero.

consequences—in the sense of most conservative tightest bounds—of making certain probability assessments. One advantage of imprecise probability models is that they allow for *imprecision*, or in other words, the use of *partial* probability assessments using bounding *inequalities* rather than equalities. In Sect. 2.2, we give a concise overview of the relevant ideas, models and techniques in the field of imprecise probabilities.

The present chapter, then, can be described as an application of ideas in imprecise probabilities to predictive inference. Its aim is to study—and develop a general framework for dealing with—coherent predictive inference using imprecise probability models. Using such models will also allow us to represent a subject's indecision, which we believe is a natural state to be in when knowing, or having learned little, about the problem at hand. It seems important to us that theories of learning under uncertainty in general, and predictive inference in particular, start out with conservative, very imprecise and indecisive models when little has been learned, and become more precise and decisive as more observations come in.

Our work here builds on, but manages to reach much further than, an earlier chapter by one of the authors [9]. The main reason why it does so, is that we are now in a position to use a very powerful mathematical language to represent imprecise-probabilistic inferences: Walley's [28] coherent sets of desirable gambles. Here, the primitive notions are not probabilities of events, nor expectations of random variables. The focus is rather on the question whether a gamble, or a risky transaction, is desirable to a subject—strictly preferred to the zero transaction, or status quo. And a basic belief model is now not a probability measure or lower prevision, but a *set of desirable gambles*.

Let us briefly summarise why, in the present chapter, we work with such sets as our basic uncertainty models for doing conservative probabilistic inference. Most importantly, and as we shall see in Sects. 2.2 and 2.3, marginalisation and conditioning are especially straightforward, and there are no issues whatsoever with conditioning on sets of (lower) probability zero. Furthermore, sets of desirable gambles provide an extremely expressive and general framework: It encompasses and subsumes as special cases both classical (or 'precise') probabilistic inference and inference in classical propositional logic [7].

So, now that we have argued why we want to use sets of desirable gambles to extend the existing probabilistic theory of predictive inference, let us explain in some detail how we intend to go about doing this. The basic building blocks are introduced in Sects. 2.2–2.8. As already indicated above, we give an overview of relevant notions and results concerning our imprecise probability model of choice—coherent sets of desirable gambles—in Sect. 2.2. In particular, we explain how to use them for conservative inference as well as conditioning; how to derive more commonly used models, such as lower previsions and lower probabilities, from them; and how they relate to precise probability models.

In Sect. 2.3, we explain how we can describe a subject's beliefs about a sequence of variables in terms of predictive sets of desirable gambles, and the derived notion of predictive lower previsions. These imprecise probability models generalise the

above-mentioned predictive probability mass functions $p^{\hat{n}}(\cdot\,|x_1,\ldots,x_{\hat{n}})$, and they constitute the basic tools we shall be working with. We also explain what are the proper formulations for the above-mentioned temporal consistency and coherence requirements in this more general context.

In Sect. 2.4, we discuss a number of inference principles that we believe could be reasonably imposed on predictive inferences, and we show how to represent them mathematically in terms of predictive sets of desirable gambles and lower previsions. *Representation insensitivity* means that predictive inferences remain essentially unchanged when we transform the set of categories, or in other words that they are essentially insensitive to the choice of representation—the category set. Another inference principle we look at imposes the so-called *specificity* property: when predictive inference is specific, then for a specific question involving a restricted number of categories, a more general model can be replaced by a more specific model that deals only with the categories of interest, and will produce the same relevant inferences [2].

The next important step is taken in Sect. 2.5, where we recall from the literature [8, 10] how to deal with exchangeability when our predictive inference models are imprecise. We recall that de Finetti's representation theorem can be significantly generalised. In this case, the temporal consistent and coherent predictive sets of desirable gambles are completely characterised by a set of (multivariate) polynomials on the simplex of all probability mass functions on the category set. This set of polynomials must satisfy a number of properties, which taken together define the notion of *Bernstein coherence*. It serves completely the same purpose as the representing probability measure: it completely determines, and conveniently and densely summarises, all predictive inferences. This is the reason why the rest of the developments in the chapter are expressed in terms of such Bernstein coherent sets of polynomials.

We introduce coherent inference systems in Sect. 2.6 as maps that associate with any finite set of categories a Bernstein coherent set of polynomials on the simplex of probability mass functions on that set. The inference principles in Sect. 2.4 impose connections between predictive inferences for different category sets, so we can represent such inference principles mathematically as restrictions on coherent inference systems, which is the main topic of Sect. 2.7.

The material in Sects. 2.8–2.10 shows, by producing explicit examples, that there are quite a few different types—even uncountable infinities—of coherent inference systems that are both representation insensitive and specific. We discuss the vacuous inference system in Sect. 2.8, the family of IDMM inference systems in Sect. 2.9 and the Haldane inference system in Sect. 2.10.

In the Conclusion (Sect. 2.11) we point to a number of surprising consequences of our results, and discuss avenues for further research.

2.2 Imprecise Probability Models

In this section, we give a concise overview of imprecise probability models for representing, and making inferences and decisions under, uncertainty.

We shall focus on sets of desirable gambles as our uncertainty models of choice, because they are the most powerful, expressive and general models at hand, because they are very intuitive to work with—though unfortunately less familiar to most people not closely involved in the field—and very importantly, because they avoid problems with conditioning on sets of (lower) probability zero. For more details, we refer to Refs. [1, 5, 8, 21, 28]. We shall, of course, also briefly mention derived results in terms of the more familiar language of (lower) previsions and probabilities.

We consider a variable X that assumes values in some possibility space A. We model a subject's beliefs about the value of X by looking at which gambles on this variable the subject finds *desirable*, meaning that he strictly prefers them to the zero gamble—the status quo. This is a very general approach, that extends the usual rationalist and subjectivist approach to probabilistic modelling to allow for indecision and imprecision.

A *gamble* is a (bounded) real-valued function f on A. It is interpreted as an uncertain reward $f(X)$ that depends on the value of X, and is expressed in units of some predetermined linear utility. It represents the reward the subject gets in a transaction where first the actual value x of X is determined, and then the subject receives the amount of utility $f(x)$—which may be negative, meaning he has to pay it. Throughout the chapter, we shall use the device of writing $f(X)$ when we want to make clear what variable the gamble f depends on. *Events* are subsets of the possibility space A. With any event $B \subseteq A$ we can associate a special gamble $\mathbb{I}_B$, called its *indicator*, which assumes the value 1 on B and 0 elsewhere.

We denote the set of all gambles on A by $\mathcal{G}(A)$. It is a linear space under point-wise addition of gambles, and point-wise multiplication of gambles with real numbers. For any subset $\mathcal{A}$ of $\mathcal{G}(A)$, posi($\mathcal{A}$) is the set of all positive linear combinations of gambles in $\mathcal{A}$: posi($\mathcal{A}$) := $\{\sum_{k=1}^{n} \lambda_k f_k : f_k \in \mathcal{A}, \lambda_k \in \mathbb{R}_{>0}, n \in \mathbb{N}\}$. Here, $\mathbb{N}$ is the set of natural numbers (without zero), and $\mathbb{R}_{>0}$ is the set of all positive real numbers. A *convex cone* of gambles is a subset $\mathcal{A}$ of $\mathcal{G}(A)$ that is closed under positive linear combinations, meaning that posi($\mathcal{A}$) = $\mathcal{A}$. For any two gambles f and g on A, we write '$f \geq g$' if $(\forall x \in A) f(x) \geq g(x)$, and '$f > g$' if $f \geq g$ and $f \neq g$. A gamble $f > 0$ is called *positive*. A gamble $g \leq 0$ is called *non-positive*. $\mathcal{G}_{>0}(A)$ denotes the convex cone of all positive gambles, and $\mathcal{G}_{\leq 0}(A)$ the convex cone of all non-positive gambles.

We collect the gambles that a subject finds desirable—strictly prefers to the zero gamble—into his *set of desirable gambles*, and we shall take such sets as our basic uncertainty models. Of course, they have to satisfy certain rationality criteria:

Definition 1 [Coherence] A set of desirable gambles $\mathcal{D} \subseteq \mathcal{G}(A)$ is called *coherent* if it satisfies the following requirements:

D1. $0 \notin \mathcal{D}$;
D2. $\mathcal{G}_{>0}(A) \subseteq \mathcal{D}$;
D3. $\mathcal{D} = \text{posi}(\mathcal{D})$.

Requirement D3 turns $\mathcal{D}$ into a *convex cone*. Due to D2, it includes $\mathcal{G}_{>0}(A)$; by D1–D3, it *avoids non-positivity*:

D4. if $f \leq 0$ then $f \notin \text{posi}(\mathcal{D})$, or equivalently $\mathcal{G}_{\leq 0}(A) \cap \text{posi}(\mathcal{D}) = \emptyset$.

$\mathcal{G}_{>0}(A)$ is the smallest coherent subset of $\mathcal{G}(A)$. This so-called *vacuous model* therefore, reflects minimal commitments on the part of the subject: if he knows absolutely nothing about the likelihood of the different outcomes, he will only strictly prefer to zero those gambles that never decrease his wealth and have some possibility of increasing it.

Let us suppose that our subject has a coherent set $\mathcal{D}$ of desirable gambles on A, expressing his beliefs about the value that a variable X assumes in A. We can then ask what his so-called *updated* set $\mathcal{D}\rfloor B$ of desirable gambles on B would be were he to receive the additional information—and nothing more—that X actually belongs to some subset B of A. The *updating*, or *conditioning*, *rule* for sets of desirable gambles states that:

$$g \in \mathcal{D}\rfloor B \Leftrightarrow g\mathbb{I}_B \in \mathcal{D} \text{ for all gambles } g \text{ on } B. \tag{2.1}$$

It states that the gamble g is desirable to a subject were he to observe that $X \in B$ if and only if the *called-off gamble* $g\mathbb{I}_B$ is desirable to him. This called-off gamble $g\mathbb{I}_B$ is the gamble on the variable X that gives a zero reward—is called off—unless $X \in B$, and in that case reduces to the gamble g on the new possibility space B. The updated set $\mathcal{D}\rfloor B$ is a set of desirable gambles on B that is still coherent, provided that $\mathcal{D}$ is [8]. We refer to Refs. [5, 21, 22] for detailed discussions of updating sets of desirable gambles.

We now use coherent sets of desirable gambles to introduce derived concepts, such as coherent lower previsions and probabilities. Given a coherent set of desirable gambles $\mathcal{D}$, the functional $\underline{P}$ defined on $\mathcal{G}(A)$ by

$$\underline{P}(f) := \sup\{\mu \in \mathbb{R} : \ f - \mu \in \mathcal{D}\} \text{ for all } f \in \mathcal{G}(A), \tag{2.2}$$

is a *coherent lower prevision* [25] [Theorem 3.8.1]. The conjugate upper prevision $\overline{P}$ is defined by $\overline{P}(f) := \inf\{\mu \in \mathbb{R} : \mu - f \in \mathcal{D}\} = -\underline{P}(-f)$. For any gamble f, $\underline{P}(f)$ is called the *l*ower prevision of f, and for any event B, $\underline{P}(\mathbb{I}_B)$ is also denoted by $\underline{P}(B)$, and called the *lower probability* of B. Similarly for upper previsions and upper probabilities.

The coherent conditional model $\mathcal{D}\rfloor B$, with B a non-empty subset of A, induces a *conditional lower prevision* $\underline{P}(\cdot\,|B)$ on $\mathcal{G}(B)$, by applying Eq. 2.2:

$$P(g|B) := \sup\{\mu \in \mathbb{R} : g - \mu \in \mathcal{D}\rfloor B\} = \sup\{\mu \in \mathbb{R} : \ [g - \mu]\mathbb{I}_B \in \mathcal{D}\}$$
$$\text{for all gambles } g \text{ on } B. \tag{2.3}$$

It is not difficult to show [25] that $\underline{P}$ and $\underline{P}(\cdot\,|B)$ are related through the following coherence condition:

$$\underline{P}([g - \underline{P}(g|B)]\mathbb{I}_B) = 0 \text{ for all } g \in \mathcal{G}(B), \tag{GBR}$$

called the *Generalised Bayes Rule*. This rule allows us to infer $\underline{P}(\cdot\,|B)$ uniquely from $\underline{P}$, provided that $\underline{P}(B) > 0$. Otherwise, there are an infinity of coherent lower previsions $\underline{P}(\cdot\,|B)$ that are coherent with $\underline{P}$ in the sense that they satisfy GBR.

Coherent sets of desirable gambles are more informative than coherent lower previsions: a gamble with positive lower prevision is always desirable and one with a negative lower prevision never, but a gamble with zero lower prevision lies on the border of the set of desirable gambles, and the lower prevision does not generally provide information about the desirability of such gambles. If such border behaviour is important—and it is when dealing with conditioning on events with zero (lower) probability [5, 21, 22, 28]—it is useful to work with sets of desirable gambles rather than lower previsions, because as Eqs. 2.1 and 2.3 tell us, they allow us to derive unique conditional models from unconditional ones.

When the lower and the upper prevision coincide on all gambles, then the real functional P defined on $\mathcal{G}(A)$ by $P(f) := \underline{P}(f) = \overline{P}(f)$ for all $f \in \mathcal{G}(A)$ is a *linear prevision*. In the particular case that A is finite, this means that it corresponds to the expectation operator associated with a probability mass function p: $P(f) = \sum_{x\in A} f(x)p(x) := E_p(f)$, where $p(x) := P(\mathbb{I}_{\{x\}})$ for all $x \in A$.

2.3 Predictive Inference

Predictive inference, in the specific sense we are focussing on here, considers a number of variables $X_1, \ldots, X_n$ assuming values in the same category set A—we define a *category set* as any non-empty *finite* set. We start our discussion of predictive inference models in the most general and representationally powerful language: coherent sets of desirable gambles, as introduced in the previous section.

Predictive inference assumes generally that a number $\check{n}$ of observations have been made, so we know the values $\check{\boldsymbol{x}} = (x_1, \ldots, x_{\check{n}})$ of the first $\check{n}$ variables $X_1, \ldots, X_{\check{n}}$. Based on this *observation sample* $\check{\boldsymbol{x}}$, a subject then has a posterior *predictive model* $\mathcal{D}_A^{\hat{n}}\rfloor\check{\boldsymbol{x}}$ for the values that the next $\hat{n}$ variables $X_{\check{n}+1}, \ldots, X_{\check{n}+\hat{n}}$ assume in $A^{\hat{n}}$. $\mathcal{D}_A^{\hat{n}}\rfloor\check{\boldsymbol{x}}$ is a coherent set of desirable gambles $f(X_{\check{n}+1}, \ldots, X_{\check{n}+\hat{n}})$ on $A^{\hat{n}}$. Here, we assume that $\hat{n} \in \mathbb{N}$. On the other hand, we want to allow that $\check{n} \in \mathbb{N}_0 := \mathbb{N} \cup \{0\}$, which is the set of all natural numbers with zero: we also want to be able to deal with the case where no previous observations have been made. In that case, we call the corresponding model $\mathcal{D}_A^{\hat{n}}$ a *prior predictive model*. Of course, technically speaking, $\check{n} + \hat{n} \leq n$.

As we said, the subject may also have a prior, unconditional model, for when no observations have yet been made. In its most general form, this will be a coherent set $\mathcal{D}_A^n$ of desirable gambles $f(X_1, \ldots, X_n)$ on A^n, for some $n \in \mathbb{N}$. Our subject may also have a coherent set $\mathcal{D}_A^{\hat{n}}$ of desirable gambles $f(X_1, \ldots, X_n)$ on A^n, where

$\hat{n} \leq n$; and the sets $\mathcal{D}_A^{\hat{n}}$ and $\mathcal{D}_A^n$ then be related to each other through the following *marginalisation*, or *temporal consistency*, requirement:

$$f(X_1, \ldots, X_{\hat{n}}) \in \mathcal{D}_A^{\hat{n}} \Leftrightarrow f(X_1, \ldots, X_n) \in \mathcal{D}_A^n \text{ for all gambles } f \text{ on } A^{\hat{n}}. \quad (2.4)$$

In this expression, and throughout this chapter, we identify a gambles f on $A^{\hat{n}}$ with its *cylindrical extension* f' on A^n, defined by $f'(x_1, \ldots, x_{\hat{n}}, \ldots, x_n) := f(x_1, \ldots, x_{\hat{n}})$ for all $(x_1, \ldots, x_n) \in A^n$. If we introduce the marginalisation operator $\text{marg}_{\hat{n}}(\cdot) := \cdot \cap \mathcal{G}(A^k)$, then the temporal consistency condition can also be rewritten simply as $\mathcal{D}_A^{\hat{n}} = \text{marg}_{\hat{n}}(\mathcal{D}_A^n) = \mathcal{D}_A^n \cap \mathcal{G}(A^{\hat{n}})$.

Prior (unconditional) predictive models $\mathcal{D}_A^n$ and posterior (conditional) ones $\mathcal{D}_A^{\hat{n}}\rfloor\check{\boldsymbol{x}}$ must also be related through the following *updating* requirement:

$$f(X_{\check{n}+1}, \ldots, X_{\check{n}+\hat{n}}) \in \mathcal{D}_A^{\hat{n}}\rfloor\check{\boldsymbol{x}} \Leftrightarrow f(X_{\check{n}+1}, \ldots, X_{\check{n}+\hat{n}})\mathbb{I}_{\{\check{\boldsymbol{x}}\}}(X_1, \ldots, X_{\check{n}}) \in \mathcal{D}_A^n$$
$$\text{for all gambles } f \text{ on } A^{\hat{n}}, \quad (2.5)$$

which is a special case of Eq. 2.1: the gamble $f(X_{\check{n}+1}, \ldots, X_{\check{n}+\hat{n}})$ is desirable after observing a sample $\check{\boldsymbol{x}}$ if and only if the gamble $f(X_{\check{n}+1}, \ldots, X_{\check{n}+\hat{n}})\mathbb{I}_{\{\check{\boldsymbol{x}}\}}(X_1, \ldots, X_{\check{n}})$ is desirable before any observations are made. This called-off gamble is the gamble that gives zero reward—is called off—unless the first $\check{n}$ observations are $\check{\boldsymbol{x}}$, and in that case reduces to the gamble $f(X_{\check{n}+1}, \ldots, X_{\check{n}+\hat{n}})$ on the variables $X_{\check{n}+1}, \ldots, X_{\check{n}+\hat{n}}$. The updating requirement is a generalisation of Bayes's Rule for updating, and in fact reduces to it when the sets of desirable gambles lead to (precise) probability mass functions [7, 28]. But contrary to Bayes's Rule for probability mass functions, the updating rule 2.5 for coherent sets of desirable gambles clearly does not suffer from problems when the conditioning event has (lower) probability zero: it allows us to infer a unique conditional model from an unconditional one, regardless of the (lower or upper) probability of the conditioning event.

As explained in Sect. 2.2, we can use the relationship 2.2 to derive *prior* (unconditional) *predictive lower previsions* $\underline{P}_A^{\hat{n}}(\cdot)$ on $\mathcal{G}(A^n)$ from the prior sets $\mathcal{D}_A^{\hat{n}}$ through:

$$\underline{P}_A^{\hat{n}}(f) := \sup\{\mu \in \mathbb{R} : f - \mu \in \mathcal{D}_A^{\hat{n}}\} \text{ for all gambles } f \text{ on } A^{\hat{n}},$$

and *posterior* (conditional) *predictive lower previsions* $\underline{P}_A^{\hat{n}}(\cdot\,|\check{\boldsymbol{x}})$ on $\mathcal{G}(A^{\hat{n}})$ from the posterior sets $\mathcal{D}_A^{\hat{n}}\rfloor\check{\boldsymbol{x}}$ through:

$$\underline{P}_A^{\hat{n}}(f|\check{\boldsymbol{x}}) := \sup\{\mu \in \mathbb{R} : f - \mu \in \mathcal{D}_A^{\hat{n}}\rfloor\check{\boldsymbol{x}}\} \text{ for all gambles } f \text{ on } A^{\hat{n}}.$$

We also want to condition predictive lower previsions on the additional information that $(X_{\check{n}+1}, \ldots, X_{\check{n}+\hat{n}}) \in B^{\hat{n}}$, where B is some proper subset of A. Using the ideas in Sects. 2.2, this leads for instance to the following lower prevision:

$$\underline{P}_A^{\hat{n}}(g|\check{\boldsymbol{x}}, B^{\hat{n}}) := \sup\{\mu \in \mathbb{R} : [g - \mu]\mathbb{I}_{B^{\hat{n}}} \in \mathcal{D}_A^{\hat{n}}\rfloor\check{\boldsymbol{x}}\} \text{ for all gambles } g \text{ on } B^{\hat{n}}, \quad (2.6)$$

which is the lower prevision $\underline{P}_A^{\hat{n}}(\cdot\,|\check{\boldsymbol{x}})$ conditioned on the event $B^{\hat{n}}$.

2.4 Principles for Predictive Inference

So far, we have introduced coherence, marginalisation and updating as basic requirements of rationality that prior and posterior predictive inference models must satisfy. In addition to these, we now also consider a number of further conditions, which have been suggested by a number of authors as reasonable properties—or requirements—for predictive inference models.

We shall call *representation insensitivity* the combination of pooling, renaming and category permutation invariance; see Ref. [9] for more information. It means that predictive inferences remain essentially unchanged when we transform the set of categories, or in other words that they are essentially insensitive to the choice of representation—the category set. It is not difficult to see that representation insensitivity can be formally characterised as follows. Consider two category sets A and B such that there is a so-called *relabelling map* $\rho : A \to B$ that is *onto*, such that $B = \rho(A) := \{\rho(x) : x \in A\}$. Then with a sample $\boldsymbol{x}$ in A^n, there corresponds a transformed sample $\rho\boldsymbol{x} := (\rho(x_1), \ldots, \rho(x_n))$ in B^n. And with any gamble f on B^n there corresponds a gamble $f \circ \rho$ on A^n.

Representation insensitivity: For all category sets A and B such that there is an onto map $\rho : A \to B$, all $\check{n}, \hat{n} \in \mathbb{N}$ considered, all $\check{\boldsymbol{x}} \in A^{\check{n}}$ and all gambles f on $B^{\hat{n}}$:

$$\underline{P}_A^{\hat{n}}(f \circ \rho) = \underline{P}_B^{\hat{n}}(f) \text{ and } \underline{\overset{\hat{n}}{A}}(f \circ \rho)\check{\boldsymbol{x}} = \underline{P}_B^{\hat{n}}(f|\rho\check{\boldsymbol{x}}), \tag{RI1}$$

or alternatively, and more generally, in terms of predictive sets of desirable gambles:

$$f \circ \rho \in \mathcal{D}_A^{\hat{n}} \Leftrightarrow f \in \mathcal{D}_B^{\hat{n}} \text{ and } f \circ \rho \in \mathcal{D}_A^{\hat{n}}\rfloor\check{\boldsymbol{x}} \Leftrightarrow f \in \mathcal{D}_B^{\hat{n}}\rfloor\rho\check{\boldsymbol{x}}. \tag{RI2}$$

There is another peculiar, but in our view intuitively appealing, potential property of predictive inferences. Assume that in addition to observing a sample of observations $\check{\boldsymbol{x}}$ of $\check{n}$ observations in a category set A, our subject comes to know or determine in some way that the $\hat{n}$ following observations will belong to a proper subset B of A, and nothing else—we might suppose for instance that an observation of $(X_{\check{n}+1}, \ldots, X_{\check{n}+\hat{n}})$ has been made, but that it is imperfect, and only allows him to conclude that $(X_{\check{n}+1}, \ldots, X_{\check{n}+\hat{n}}) \in B^{\hat{n}}$.

We can then make the following requirement, which uses models conditioned on the event $B^{\hat{n}}$, as introduced through Eqs. 2.1, 2.3 and 2.6.

Specificity: For all category sets A and B such that $B \subseteq A$, all $\check{n}, \hat{n} \in \mathbb{N}$ considered, all $\check{\boldsymbol{x}} \in A^{\check{n}}$ and all gambles f on $B^{\hat{n}}$:

$$\underline{P}_A^{\hat{n}}(f|B^{\hat{n}}) = \underline{P}_B^{\hat{n}}(f) \text{ and } \underline{P}_A^{\hat{n}}(f|\check{\boldsymbol{x}}, B^{\hat{n}}) = \underline{P}_B^{\hat{n}}(f|\check{\boldsymbol{x}}\downarrow_B), \tag{SP1}$$

or alternatively, and more generally, in terms of predictive sets of desirable gambles:

$$f\mathbb{I}_{B^{\hat{n}}} \in \mathcal{D}_A^{\hat{n}} \Leftrightarrow f \in \mathcal{D}_B^{\hat{n}} \text{ and } f\mathbb{I}_{B^{\hat{n}}} \in \mathcal{D}_A^{\hat{n}}\rfloor\check{\boldsymbol{x}} \Leftrightarrow f \in \mathcal{D}_B^{\hat{n}}\rfloor\check{\boldsymbol{x}}\downarrow_B, \tag{SP2}$$

where $\check{\boldsymbol{x}}\downarrow_B$ is the tuple of observations obtained by eliminating from the tuple $\check{\boldsymbol{x}}$ all observations not in B. In these expressions, when $\check{\boldsymbol{x}}\downarrow_B$ is the empty tuple, so when

no observations in $\check{\boldsymbol{x}}$ are in B, the 'posterior' predictive model is simply taken to reduce to the 'prior' predictive model. Specificity [2, 3, 24] means that *the predictive inferences that a subject makes are the same as the ones he would get by focussing on the category set B, and at the same time discarding all the previous observations producing values outside B, in effect only retaining the observations that were inside B!* It is as if knowing that the future observations belong to B allows our subject to ignore all the previous observations that happened to lie outside B.

2.5 Adding Exchangeability to the Picture

We are now, for the remainder of this chapter, going to add two additional assumptions. The *first assumption* is that we are dealing with a *countably infinite sequence* of variables $X_1, \ldots, X_n, \ldots$ that assume values in the same category set A. For our predictive inference models, this means that there is a sequence $\mathcal{D}_A^n$ of coherent sets of desirable gambles on A^n, $n \in \mathbb{N}$. The *second assumption* is that this sequence of variables is *exchangeable*, which means, roughly speaking, that the subject believes that the order in which these variables are observed, or present themselves, has no influence on the decisions and inferences he will make regarding these variables.

In this section, we explain succinctly how to deal with these assumptions technically, and what their consequences are for the predictive models we are interested in. For a detailed discussion and derivation of the results presented here, we refer to Refs. [8, 10].

We begin with some useful notation, which will be employed numerous times in what follows. Consider any element $\boldsymbol{\alpha} \in \mathbb{R}^A$. We consider $\boldsymbol{\alpha}$ as an *A-tuple*, with as many (real) components $\alpha_x \in \mathbb{R}$ as there are categories x in A. For any subset $B \subseteq A$, we then denote by $\alpha_B := \sum_{x \in B} \alpha_x$ the sum of its components over B.

Consider an arbitrary $n \in \mathbb{N}$. We denote by $\boldsymbol{x} = (x_1, \ldots, x_n)$ a generic, arbitrary element of A^n. $\mathcal{P}^n$ is the set of all permutations π of the index set $\{1, \ldots, n\}$. With any such permutation π, we can associate a permutation of A^n, also denoted by π, and defined by $(\pi \boldsymbol{x})_k := x_{\pi(k)}$, or in other words, $\pi(x_1, \ldots, x_n) := (x_{\pi(1)}, \ldots, x_{\pi(n)})$. Similarly, we lift π to a permutation π^t of $\mathcal{G}(A^n)$ by letting $\pi^t f := f \circ \pi$, so $(\pi^t f)(\boldsymbol{x}) := f(\pi \boldsymbol{x})$. The permutation invariant atoms $[\boldsymbol{x}] := \{\pi \boldsymbol{x} : \pi \in \mathcal{P}^n\}$, $\boldsymbol{x} \in A^n$ are the smallest permutation invariant subsets of A^n.

We now introduce the *counting map* $\boldsymbol{T} : A^n \to \mathcal{N}_A^n : \boldsymbol{x} \mapsto \boldsymbol{T}(\boldsymbol{x})$, where the *count vector* $\boldsymbol{T}(\boldsymbol{x})$ is the A-tuple with components $T_z(\boldsymbol{x}) := |\{k \in \{1, \ldots, n\} x_k = z\}|$ for all $z \in A$, and the set of possible *count vectors* for n observations in A is given by $\mathcal{N}_A^n := \{\boldsymbol{m} \in \mathbb{N}_0^A : m_A = n\}$. So, $T_z(\boldsymbol{x})$ is the number of times the category z appears in the sample $\boldsymbol{x}$. If $\boldsymbol{m} = \boldsymbol{T}(\boldsymbol{x})$, then $[\boldsymbol{x}] = \{\boldsymbol{y} \in A^n : \boldsymbol{T}(\boldsymbol{y}) = \boldsymbol{m}\}$, so the atom $[\boldsymbol{x}]$ is completely determined by the single count vector $\boldsymbol{m}$ of all its elements, and is therefore also denoted by $[\boldsymbol{m}]$.

We also consider the linear expectation operator $\mathrm{Hy}_A^n(\cdot\,|\boldsymbol{m})$ associated with the uniform distribution on the invariant atom $[\boldsymbol{m}]$:

$$\mathrm{Hy}_A^n(f|\boldsymbol{m}) := \frac{1}{|[\boldsymbol{m}]|}\sum_{\boldsymbol{x}\in[\boldsymbol{m}]} f(\boldsymbol{x}) \text{ for all gambles } f \text{ on } A^n,$$

where the number of elements $\nu(\boldsymbol{m}) := |[\boldsymbol{m}]|$ in the invariant atom $[\boldsymbol{m}]$ is given by the *multinomial coefficient*:

$$\nu(\boldsymbol{m}) = \binom{m_A}{\boldsymbol{m}} = \binom{n}{\boldsymbol{m}} := \frac{n!}{\prod_{z\in A} m_z!}.$$

This expectation operator characterises a (multivariate) *hyper-geometric distribution* [16] [Sect. 39.2], associated with random sampling without replacement from an urn with n balls of types $z \in A$, whose composition is characterised by the count vector $\boldsymbol{m}$. This hyper-geometric expectation operator can also be seen as a linear transformation Hy_A^n between the linear space $\mathcal{G}(A^n)$ and the generally much lower-dimensional linear space $\mathcal{G}(\mathcal{N}_A^n)$, turning a gamble f on A^n into a so-called *count gamble* $\mathrm{Hy}_A^n(f) := \mathrm{Hy}_A^n(f|\cdot)$ on count vectors.

Next, we consider the simplex Σ_A of all probability mass functions $\boldsymbol{\theta}$ on A: $\Sigma_A := \{\boldsymbol{\theta} \in \mathbb{R}^A : \boldsymbol{\theta} \geq 0 \text{ and } \theta_A = 1\}$. With a probability mass function $\boldsymbol{\theta} \in \Sigma_A$ on A, there corresponds the following *multinomial expectation* operator $\mathrm{Mn}_A^n(\cdot\,|\boldsymbol{\theta})$:[2]

$$\mathrm{Mn}_A^n(f|\boldsymbol{\theta}) := \sum_{\boldsymbol{x}\in A^n} f(\boldsymbol{x})\prod_{z\in A}\theta_z^{T_z(\boldsymbol{x})} \text{ for all gambles } f \text{ on } A^n,$$

which characterises the multinomial distribution, associated with n independent trials of an experiment with possible outcomes in A and probability mass function $\boldsymbol{\theta}$. Observe that $\mathrm{Mn}_A^n(f|\boldsymbol{\theta}) = \sum_{\boldsymbol{m}\in\mathcal{N}_A^n} \mathrm{Hy}_A^n(f|\boldsymbol{m})\nu(\boldsymbol{m})\prod_{z\in A}\theta_z^{m_z} = \mathrm{CoMn}_A^n(\mathrm{Hy}_A^n(f)|\boldsymbol{\theta})$, where we used the so-called *count multinomial expectation* operator:

$$\mathrm{CoMn}_A^n(g|\boldsymbol{\theta}) := \sum_{\boldsymbol{m}\in\mathcal{N}_A^n} g(\boldsymbol{m})\nu(\boldsymbol{m})\prod_{z\in A}\theta_z^{m_z} \text{ for all gambles } g \text{ on } \mathcal{N}_A^n. \tag{2.7}$$

Let us introduce the notation $\mathcal{N}_A := \bigcup_{m\in\mathbb{N}}\mathcal{N}_A^m$ for the set of all possible count vectors, corresponding to samples of at least one observation. $\mathcal{N}_A^0$ is the singleton containing only the null count vector $\boldsymbol{0}$, all of whose components are zero. Then $\bigcup_{m\in\mathbb{N}_0}\mathcal{N}_A^m = \mathcal{N}_A \cup \{\boldsymbol{0}\}$ is the set of all possible count vectors. For any such count vector $\boldsymbol{m} \in \mathcal{N}_A \cup \{\boldsymbol{0}\}$, we consider the (multivariate) *Bernstein basis polynomial* $\mathrm{B}_{A,\boldsymbol{m}}$ of degree m_A on Σ_A, defined by:

$$\mathrm{B}_{A,\boldsymbol{m}}(\boldsymbol{\theta}) := \nu(\boldsymbol{m})\prod_{z\in A}\theta_z^{m_z} = \binom{m_A}{\boldsymbol{m}}\prod_{z\in A}\theta_z^{m_z} \text{ for all } \boldsymbol{\theta} \in \Sigma_A. \tag{2.8}$$

[2] To avoid confusion, we make a (perhaps non-standard) distinction between the multinomial expectation, which is associated with sequences of observations, and the count multinomial expectation, associated with their count vectors.

In particular, of course, $\mathrm{B}_{A,\boldsymbol{\theta}} = 1$. Any linear combination p of Bernstein basis polynomials of degree $n \geq 0$ is a (multivariate) polynomial (gamble) on Σ_A, whose degree $\deg(p)$ is at most n.[3] We denote the linear space of all these polynomials of degree up to n by $\mathcal{V}^n(A)$. Of course, polynomials of degree zero are simply real constants. For any $n \geq 0$, we can introduce a linear isomorphism CoMn_A^n between the linear spaces $\mathcal{G}(\mathcal{N}_A^n)$ and $\mathcal{V}^n(A)$: with any gamble g on $\mathcal{N}_A^n$, there corresponds a polynomial $\mathrm{CoMn}_A^n(g) := \mathrm{CoMn}_A^n(g|\,\cdot\,) = \sum_{\boldsymbol{m}\in\mathcal{N}_A^n} g(\boldsymbol{m})\mathrm{B}_{A,\boldsymbol{m}}$ in $\mathcal{V}^n(A)$, and conversely, for any polynomial $p \in \mathcal{V}^n(A)$ there is a unique gamble b_p^n on $\mathcal{N}_A^n$ such that $p = \mathrm{CoMn}_A^n(b_p^n)$ [8].[4] We denote by $\mathcal{V}(A) := \bigcup_{n\in\mathbb{N}_0} \mathcal{V}^n(A)$ the linear space of all (multivariate) polynomials on Σ_A, of arbitrary degree.

A set $\mathcal{H}_A \subseteq \mathcal{V}(A)$ of polynomials on Σ_A is called *Bernstein coherent* if it satisfies the following properties:

B1. $0 \notin \mathcal{H}_A$;
B2. $\mathcal{V}^+(A) \subseteq \mathcal{H}_A$;
B3. $\mathrm{posi}(\mathcal{H}_A) = \mathcal{H}_A$.

Here, $\mathcal{V}^+(A)$ is the set of *Bernstein positive* polynomials on Σ_A: those polynomials p such that $p(\boldsymbol{\theta}) > 0$ for all $\boldsymbol{\theta}$ in the interior $\mathrm{int}(\Sigma_A) := \{\boldsymbol{\theta} \in \Sigma_A : (\forall x \in A)\theta_x > 0\}$ of Σ_A. As a consequence, for the set $\mathcal{V}_0^-(A) := -\mathcal{V}^+(A) \cup \{0\}$ of *Bernstein non-positive* polynomials:

B4. $\mathcal{V}_0^-(A) \cap \mathcal{H}_A = \emptyset$.

We are now ready to deal with exchangeability. We shall give a definition for coherent sets of desirable gambles that generalises de Finetti's definition [11, 13], and which allows for a generalisation of his representation theorem.

First of all, fix $n \in \mathbb{N}$. Then the subject considers the variables $X_1, \ldots, X_n$ to be exchangeable when he does not distinguish between any gamble f on A^n and its permuted version $\pi^t f$, or in other words, if the gamble $f - \pi^t f$ is equivalent to the zero gamble for—or *indifferent* to—him. This means that he has a so-called *set of indifferent gambles*: $\mathcal{I}_A^n := \{f - \pi^t f : f \in \mathcal{G}(A^n) \text{ and } \pi \in \mathcal{P}^n\}$. If the subject also has a coherent set of desirable gambles $\mathcal{D}_A^n$, then this set must be compatible with the set of indifferent gambles $\mathcal{I}_A^n$, in the sense that it must satisfy the rationality requirement $\mathcal{D}_A^n + \mathcal{I}_A^n = \mathcal{D}_A^n$ [8, 23]. We then say that the sequence $X_1, \ldots, X_n$, and the model $\mathcal{D}_A^n$, are *exchangeable*. Next, the countably infinite sequence of variables $X_1, \ldots, X_n \ldots$ is called exchangeable if all the finite subsequences $X_1, \ldots, X_n$ are, for $n \in \mathbb{N}$. This means that all models $\mathcal{D}_A^n$, $n \in \mathbb{N}$ are exchangeable. They should of course also be temporally consistent.

---

[3] The degree may be smaller than n because the sum of all Bernstein basis polynomials of fixed degree is one. Strictly speaking, these polynomials p are restrictions to Σ_A of multivariate polynomials q on $\mathbb{R}^A$, called *representations* of p. For any p, there are multiple representations, with possibly different degrees. The smallest such degree is then called the degree $\deg(p)$ of p.

[4] Strictly speaking, Eq. 2.7 only defines the count multinomial expectation operator CoMn_A^n for $n > 0$, but it is clear that the definition extends trivially to the case $n = 0$.

Theorem 1 *[Representation Theorem [8]] The sequence of sets $\mathcal{D}_A^n$ of desirable gambles on A^n, $n \in \mathbb{N}$ is coherent, temporally consistent and exchangeable if and only if there is a Bernstein coherent set $\mathcal{H}_A$ of polynomials on Σ_A such that for all $\hat{n} \in \mathbb{N}$, all gambles f on $A^{\hat{n}}$, all $\check{\boldsymbol{m}} \in \mathcal{N}_A$ and all $\check{\boldsymbol{x}} \in [\check{\boldsymbol{m}}]$:*

$$f \in \mathcal{D}_A^{\hat{n}} \Leftrightarrow Mn_A^{\hat{n}}(f) \in \mathcal{H}_A \text{ and } f \in \mathcal{D}_A^{\hat{n}}\rfloor\check{\boldsymbol{x}} \Leftrightarrow Mn_A^{\hat{n}}(f)\mathrm{B}_{A,\check{\boldsymbol{m}}} \in \mathcal{H}_A. \tag{2.9}$$

In that case this representation $\mathcal{H}_A$ *is unique and given by* $\mathcal{H}_A := \bigcup_{n\in\mathbb{N}} Mn_A^n(\mathcal{D}_A^n)$.
The representation $\mathcal{H}_A$ is a set of polynomials that plays the same role as a density, or distribution function, on Σ_A in the precise-probabilistic case. It follows from Eq. 2.9 that $\mathcal{H}_A$ *completely determines* all predictive inferences about the sequence of variables $X_1, \ldots, X_n, \ldots$, as it fixes all prior predictive models $\mathcal{D}_A^{\hat{n}}$ and all posterior predictive models $\mathcal{D}_A^{\hat{n}}\rfloor\check{\boldsymbol{x}}$.

Equation 2.9 also tells us that the posterior predictive models $\mathcal{D}_A^{\hat{n}}\rfloor\check{\boldsymbol{x}}$ only depend on the observed sequence $\check{\boldsymbol{x}}$ through the count vector $\check{\boldsymbol{m}} = \boldsymbol{T}(\check{\boldsymbol{x}})$: Count vectors are *sufficient statistics* under exchangeability. For this reason, we shall from now on denote these posterior predictive models by $\mathcal{D}_A^{\hat{n}}\rfloor\check{\boldsymbol{m}}$ as well as by $\mathcal{D}_A^{\hat{n}}\rfloor\check{\boldsymbol{x}}$. Also, every now and then, we shall use $\mathcal{D}_A^{\hat{n}}\rfloor\boldsymbol{0}$ as an alternative notation for $\mathcal{D}_A^{\hat{n}}$.

An immediate but interesting consequence of Theorem 1 is that updating on observations preserves exchangeability: after observing the values of the first $\check{n}$ variables, with count vector $\check{\boldsymbol{m}}$, the remaining sequence of variables $X_{\check{n}+1}, X_{\check{n}+2}, \ldots$ is still exchangeable, and Eq. 2.9 tells us that its representation is given by the Bernstein coherent set of polynomials $\mathcal{H}_A\rfloor\check{\boldsymbol{m}}$ defined by:

$$\mathcal{H}_A\rfloor\check{\boldsymbol{m}} := \{p \in \mathcal{V}(A) : \mathrm{B}_{A,\check{\boldsymbol{m}}}\, p \in \mathcal{H}_A\}. \tag{2.10}$$

For the special case $\check{\boldsymbol{m}} = \boldsymbol{0}$, we find that $\mathcal{H}_A\rfloor\boldsymbol{0} = \mathcal{H}_A$. Clearly, $\mathcal{H}_A\rfloor\check{\boldsymbol{m}}$ is completely determined by $\mathcal{H}_A$. One can consider $\mathcal{H}_A$ as a prior model on the parameter space Σ_A, and $\mathcal{H}_A\rfloor\check{\boldsymbol{m}}$ plays the role of the posterior that is derived from it. We see from Eqs. 2.9 and 2.10 that—similar to what happens in a precise-probabilistic setting—the multinomial distribution serves as a direct link between on the one hand, the 'prior' $\mathcal{H}_A$ and its prior predictive inference models $\mathcal{D}_A^{\hat{n}}$ and, on the other hand, the 'posterior' $\mathcal{H}_A\rfloor\check{\boldsymbol{m}}$ and its posterior inference models $\mathcal{D}_A^n\rfloor\check{\boldsymbol{m}}$. Recalling our convention for $\check{\boldsymbol{m}} = \boldsymbol{0}$, we can summarise this as follows: for all $\hat{n} \in \mathbb{N}$ and all $\check{\boldsymbol{m}} \in \mathcal{N}_A \cup \{\boldsymbol{0}\}$:

$$\mathcal{D}_A^{\hat{n}}\rfloor\check{\boldsymbol{m}} = \{f \in \mathcal{G}(A^{\hat{n}}) : \mathrm{Mn}_A^{\hat{n}}(f) \in \mathcal{H}_A\rfloor\check{\boldsymbol{m}}\} \tag{2.11}$$

and, as an immediate consequence,

$$\underline{P}_A^{\hat{n}}(f|\check{\boldsymbol{m}}) = \sup\{\mu \in \mathbb{R} : \mathrm{Mn}_A^{\hat{n}}(f) - \mu \in \mathcal{H}_A\rfloor\check{\boldsymbol{m}}\} \text{ for all } f \in \mathcal{G}(A^{\hat{n}}). \tag{2.12}$$

The sets of desirable polynomials $\mathcal{H}_A$ are the fundamental models, as they allow us to determine the $\mathcal{H}_A\rfloor\check{\boldsymbol{m}}$ and all predictive models uniquely.

2.6 Inference Systems

We have seen in the previous section that, once we fix a category set A, predictive inferences about exchangeable sequences assuming values in A are completely determined by a Bernstein coherent set $\mathcal{H}_A$ of polynomials on Σ_A. So, if we had some way of associating a Bernstein coherent set $\mathcal{H}_A$ with every possible set of categories A, this would completely fix all predictive inferences. This leads us to the following definition.

Definition 2 [Inference systems] We denote by $\mathbb{F}$ the collection of all category sets, i.e. finite non-empty sets. An *inference system* is a map Φ that maps any category set $A \in \mathbb{F}$ to some set of polynomials $\Phi(A) = \mathcal{H}_A$ on Σ_A. An inference system Φ is *coherent* if for all category sets $A \in \mathbb{F}$, $\Phi(A)$ is a Bernstein coherent set of polynomials on Σ_A.
So, a coherent inference system is a way to systematically associate coherent predictive inferences with any category set. Since the inference principles in Sect. 2.4 impose connections between predictive inferences for different category sets, we now see that we can interpret these inference principles—or rather, represent them mathematically—as properties of, or restrictions on, coherent inference systems.

2.7 Representation Insensitivity and Specificity Under Exchangeability

Let us now investigate what form the inference principles of representation insensitivity RI2 and specificity SP2 take for predictive inference under exchangeability, when such inference can be completely characterised by Bernstein coherent sets of polynomials. This will allow us to reformulate these principles as constraints on—or properties of—inference systems.

Recalling the notations and assumptions in Sect. 2.4, we start by considering the surjective (onto) map $C_\rho : \mathbb{R}^A \to \mathbb{R}^B$, defined by $C_\rho(\boldsymbol{\alpha})_z := \sum_{x \in A : \rho(x) = z} \alpha_x$ for all $\boldsymbol{\alpha} \in \mathbb{R}^A$ and all $z \in B$. It allows us to give the following elegant characterisation of representation insensitivity.

Theorem 2 *An inference system Φ is representation insensitive if and only if for all category sets A and B such that there is an onto map $\rho : A \to B$, for all $p \in \mathcal{V}(B)$ and all $\boldsymbol{m} \in \mathcal{N}_A \cup \{\boldsymbol{0}\}$: $(p \circ C_\rho)\mathrm{B}_{A,\boldsymbol{m}} \in \Phi(A) \Leftrightarrow p\mathrm{B}_{B,C_\rho(\boldsymbol{m})} \in \Phi(B)$.*

Next, we turn to specificity. Let us define the surjective map $r_B : \mathbb{R}^A \to \mathbb{R}^B$ by: $r_B(\boldsymbol{\alpha})_z := \alpha_z$ for all $\boldsymbol{\alpha} \in \mathbb{R}^A$ and all $z \in B$. So in particular, $r_B(\boldsymbol{m})$ is the count vector on B obtained by restricting to B the (indices of the) components of the count vector $\boldsymbol{m}$ on A. We also define the one-to-one map $i_A : \mathbb{R}^B \to \mathbb{R}^A$ by $i_A(\boldsymbol{\alpha})_x := \alpha_x$ if $x \in B$ and 0 otherwise, for all $\boldsymbol{\alpha} \in \mathbb{R}^B$ and all $x \in A$. This map can be used to define the

following one-to-one maps ${}^{r}\mathrm{I}_{B,A} : \mathcal{V}(B) \to \mathcal{V}(A)$, for any $r \in \mathbb{N}_0$, as follows:

$$ {}^{r}\mathrm{I}_{B,A}(p) := \sum_{\boldsymbol{n} \in \mathcal{N}_B^{deg(p)+r}} b_p^{deg(p)+r}(\boldsymbol{n}) \mathrm{B}_{A,i_A(\boldsymbol{n})} \text{ for all polynomials } p \text{ in } \mathcal{V}(B). \quad (2.13) $$

The maps ${}^{r}\mathrm{I}_{B,A}$ allow us to give the following elegant characterisation of specificity:

Theorem 3 *An inference system Φ is specific if and only if for all category sets A and B such that $B \subseteq A$, for all $p \in \mathcal{V}(B)$, all $\boldsymbol{m} \in \mathcal{N}_A \cup \{\boldsymbol{0}\}$ and all $r \in \mathbb{N}_0$: ${}^{r}\mathrm{I}_{B,A}(p)\mathrm{B}_{A,\boldsymbol{m}} \in \Phi(A) \Leftrightarrow p\mathrm{B}_{B,r_B(\boldsymbol{m})} \in \Phi(B)$.*

2.8 The Vacuous Inference System

In this and the following sections, we provide explicit and interesting examples of representation insensitive and specific inference systems. We begin with the simplest one: the vacuous inference system Φ_{V}, which is the smallest, or most conservative, coherent inference system. It associates with any category set A the smallest Bernstein coherent set $\Phi_{\mathrm{V}}(A) = \mathcal{H}_{V,A} := \mathcal{V}^+(A)$ containing all the Bernstein positive polynomials—the ones that are guaranteed to be there anyway, by Bernstein coherence alone. Since $\mathcal{V}^+(A)$ consists of all the polynomials that are positive on $\mathrm{int}(\Sigma_A)$ we easily derive that, for any $\check{\boldsymbol{m}} \in \mathcal{N}_A \cup \{\boldsymbol{0}\}$, $\mathcal{H}_{V,A}\rfloor\check{\boldsymbol{m}} = \mathcal{H}_{V,A} = \mathcal{V}^+(A)$. The predictive models for this inference system are now straightforward to find, as they follow directly from Eqs. 2.11 and 2.12. For any $\hat{n} \in \mathbb{N}$ and any $\check{\boldsymbol{m}} \in \mathcal{N}_A \cup \{\boldsymbol{0}\}$, we find that

$$ \mathcal{D}^{\hat{n}}_{\mathrm{V},A} = \mathcal{D}^{\hat{n}}_{\mathrm{V},A}\rfloor\check{\boldsymbol{m}} = \{f \in \mathcal{G}(A^{\hat{n}}) : \mathrm{Mn}^{\hat{n}}_A(f) \in \mathcal{V}^+(A)\} \quad (2.14) $$

$$ \underline{P}^{\hat{n}}_{\mathrm{V},A}(f) = \underline{P}^{\hat{n}}_{\mathrm{V},A}(f|\check{\boldsymbol{m}}) = \min_{\boldsymbol{\theta} \in \Sigma_A} \mathrm{Mn}^{\hat{n}}_A(f|\boldsymbol{\theta}) \text{ for all } f \in \mathcal{G}(A^{\hat{n}}). \quad (2.15) $$

In particular, $\mathcal{D}^1_{\mathrm{V},A} = \mathcal{D}^1_{\mathrm{V},A}\rfloor\check{\boldsymbol{m}} = \mathcal{G}_{>0}(A)$, and $\underline{P}^1_{\mathrm{V},A}(f) = \underline{P}^1_{\mathrm{V},A}(f|\check{\boldsymbol{m}}) = \min f$ for all $f \in \mathcal{G}(A)$. These are the most conservative exchangeable predictive models, and they arise from making no other assessments than exchangeability alone. They are not very interesting, because they involve no non-trivial commitments, and they do not allow learning from observations.

Even though it makes no non-trivial inferences, the vacuous inference system satisfies representation insensitivity and specificity.

Theorem 4 *The vacuous inference system Φ_{V} is coherent, representation insensitive and specific.*

We now show that there is, besides Φ_{V}, an infinity of other, more committal, specific and representation insensitive coherent inference systems.

2.9 The IDMM Inference Systems

Imprecise Dirichlet models (or IDMs, for short) are a family of parametric inference models introduced by Walley [26] as conveniently chosen sets of *Dirichlet densities* $\mathrm{di}_A(\cdot|\boldsymbol{\alpha})$ with constant prior weight s:

$$\{\mathrm{di}_A(\cdot|\boldsymbol{\alpha}): \boldsymbol{\alpha} \in \mathrm{K}_A^s\}, \text{ with } \mathrm{K}_A^s := \{\boldsymbol{\alpha} \in \mathbb{R}_{>0}^A: \alpha_A = s\} = \{s\boldsymbol{t}: \boldsymbol{t} \in \mathrm{int}(\Sigma_A)\}, \quad (2.16)$$

for any value of the (so-called) hyperparameter $s \in \mathbb{R}_{>0}$ and any category set A. The Dirichlet densities $\mathrm{di}_A(\cdot|\boldsymbol{\alpha})$ are defined on $\mathrm{int}(\Sigma_A)$.

These IDMs generalise the imprecise beta models introduced earlier by Walley [25]. In a later chapter [29], Walley and Bernard introduced a closely related family of predictive inference models, called the IDMMs. We use the ideas behind Walley's IDM(M)s to construct an interesting family of coherent inference systems. Interestingly, we shall need a slightly modified version of Walley's IDMs to make things work. The reason for this is that Walley's original version, as described by Eq. 2.16, has a number of less desirable properties, that were either unknown to, or ignored by, Walley and Bernard. For our present purposes, it suffices to mention that, contrary to what is often claimed, and in contradistinction with our new version, inferences using the original version of the IDM(M) do not always become more conservative (or less committal) as the hyperparameter s increases.

In our version, rather than using the hyperparameter sets K_A^s, we consider the sets

$$\Delta_A^s := \{\boldsymbol{\alpha} \in \mathbb{R}_{>0}^A : \alpha_A < s\} \text{for any } s \in \mathbb{R}_{>0}.$$

Observe that $\Delta_A^s = \{s'\boldsymbol{t} : s' \in \mathbb{R}_{>0}, s' < s \text{ and } \boldsymbol{t} \in \mathrm{int}(\Sigma_A)\} = \bigcup_{0<s'<s} \mathrm{K}_A^{s'}$. For any $s \in \mathbb{R}_{>0}$, and any category set A, we now consider the following set of desirable polynomials p, with positive Dirichlet expectation $\mathrm{Di}_A(p|\boldsymbol{\alpha})$ for all hyperparameters $\boldsymbol{\alpha} \in \Delta_A^s$:

$$\mathcal{H}_{\mathrm{IDM},A}^s := \{p \in \mathcal{V}(A) : (\forall \boldsymbol{\alpha} \in \Delta_A^s)\mathrm{Di}_A(p|\boldsymbol{\alpha}) > 0\}.$$

We shall see further on in Theorem 5 that this set is Bernstein coherent. We call the inference system Φ_{IDM}^s, defined by $\Phi_{\mathrm{IDM}}^s(A) := \mathcal{H}_{\mathrm{IDM},A}^s$ for all category sets A, the *IDMM inference system* with hyperparameter $s > 0$. The corresponding updated models are, for any $\boldsymbol{m} \in \mathcal{N}_A \cup \{\mathbf{0}\}$, given by:

$$\mathcal{H}_{\mathrm{IDM},A}^s\rfloor\boldsymbol{m} = \{p \in \mathcal{V}(A) : (\forall \boldsymbol{\alpha} \in \Delta_A^s)\mathrm{Di}_A(p|\boldsymbol{m} + \boldsymbol{\alpha}) > 0\} \quad (2.17)$$

Using these expressions, the predictive models for the IDMM inference system are straightforward to find; it suffices to apply Eqs. 2.11 and 2.12. For any $\hat{n} \in \mathbb{N}$ and any $\boldsymbol{m} \in \mathcal{N}_A \cup \{\mathbf{0}\}$:

$$\mathcal{D}_{\mathrm{IDM},A}^{s,\hat{n}}\rfloor\boldsymbol{m} = \{f \in \mathcal{G}(A^{\hat{n}}) : (\forall \boldsymbol{\alpha} \in \Delta_A^s)\mathrm{Di}_A(\mathrm{Mn}_A^{\hat{n}}(f)|\boldsymbol{m} + \boldsymbol{\alpha}) > 0\}, \quad (2.18)$$

$$\underline{P}_{\mathrm{IDM},A}^{s,\hat{n}}(f|\boldsymbol{m}) = \inf_{\boldsymbol{\alpha} \in \Delta_A^s} \mathrm{Di}_A(\mathrm{Mn}_A^{\hat{n}}(f)|\boldsymbol{m} + \boldsymbol{\alpha}) \text{ for all } f \in \mathcal{G}(A^{\hat{n}}), \quad (2.19)$$

where:

$$\mathrm{Di}_A(\mathrm{Mn}_A^{\hat{n}}(f)|\check{\boldsymbol{m}}+\boldsymbol{\alpha}) = \sum_{\hat{\boldsymbol{m}}\in\mathcal{N}_A^{\hat{n}}} \mathrm{Hy}_A^{\hat{n}}(f|\hat{\boldsymbol{m}}) \frac{1}{(\check{\boldsymbol{m}}_A+\alpha_A)^{(\hat{n})}} \binom{\hat{n}}{\hat{\boldsymbol{m}}} \prod_{x\in A} (\check{\boldsymbol{m}}_x+\alpha_x)^{(\hat{\boldsymbol{m}}_x)}.$$

In general, these expressions seem forbidding, but for $\hat{n}=1$, the so-called *immediate prediction models* are manageable enough: for any $\check{\boldsymbol{m}} \in \mathcal{N}_A \cup \{\mathbf{0}\}$

$$\mathcal{D}^{s,1}_{\mathrm{IDM},A}\rfloor\check{\boldsymbol{m}} = \left\{ f \in \mathcal{G}(A) : f > -\frac{1}{s}\sum_{x\in A} f(x)\check{\boldsymbol{m}}_x \right\}, \tag{2.20}$$

$$\underline{P}^{s,1}_{\mathrm{IDM},A}(f|\check{\boldsymbol{m}}) = \frac{1}{\check{\boldsymbol{m}}_A+s}\sum_{x\in A} f(x)\check{\boldsymbol{m}}_x + \frac{s}{\check{\boldsymbol{m}}_A+s}\min f \text{ for all } f \in \mathcal{G}(A) \tag{2.21}$$

Interestingly, the immediate prediction models of our version of the IDMM inference system coincide with those of Walley's original version.

The IDMM inference systems constitute an uncountably infinite family of coherent inference systems, each of which satisfies the representation insensitivity and specificity requirements.

Theorem 5 *For any $s \in \mathbb{R}_{>0}$, the IDMM inference system Φ^s_{IDM} is coherent, representation insensitive and specific.*

2.10 The Haldane Inference System

We can ask ourselves whether there are representation insensitive (and specific) inference systems whose *posterior* predictive lower previsions become precise (linear) previsions. In the present section, we show that this is indeed the case. We use the family of IDMM inference systems Φ^s_{IDM}, $s \in \mathbb{R}_{>0}$, to define an inference system Φ_H that is more committal than each of them:

$$\Phi_H(A) = \mathcal{H}_{\mathrm{H},A} := \bigcup_{s\in\mathbb{R}_{>0}} \mathcal{H}^s_{\mathrm{IDM},A} = \bigcup_{s\in\mathbb{R}_{>0}} \Phi^s_{\mathrm{IDM}}(A) \text{ for all category sets } A.$$

We call this Φ_H the *Haldane inference system*, for reasons that will become clear further in this section.

Theorem 6 *The Haldane inference system Φ_H is coherent, representation insensitive and specific.*

It can be shown that, due to its representation insensitivity, the Haldane system satisfies prior near-ignorance: this means that before making any observation, its immediate prediction model is vacuous, and as far away from a precise probability model as possible. But after making even a single observation, its inferences become

precise-probabilistic: They coincide with the inferences generated by the Haldane (improper) prior. To get there, we first take a look at the models involving sets of desirable gambles. For any $\check{\boldsymbol{m}} \in \mathcal{N}_A \cup \{\mathbf{0}\}$:

$$\mathcal{H}_{\mathrm{H},A}\rfloor\check{\boldsymbol{m}} = \{p \in \mathcal{V}(A) : (\exists s \in \mathbb{R}_{>0})(\forall \boldsymbol{\alpha} \in \Delta_A^s)\mathrm{Di}_A(p|\check{\boldsymbol{m}} + \boldsymbol{\alpha}) > 0\}. \tag{2.22}$$

The corresponding predictive models are easily derived by applying Eq. 2.11. For any $\check{\boldsymbol{m}} \in \mathcal{N}_A \cup \{\mathbf{0}\}$:

$$\mathcal{D}^{\hat{n}}_{\mathrm{H},A}\rfloor\check{\boldsymbol{m}} = \{f \in \mathcal{G}(A^{\hat{n}}) : (\exists s \in \mathbb{R}_{>0})(\forall \boldsymbol{\alpha} \in \Delta_A^s)\mathrm{Di}_A(\mathrm{Mn}^{\hat{n}}_A(f)|\check{\boldsymbol{m}} + \boldsymbol{\alpha}) > 0\}. \tag{2.23}$$

The immediate prediction models are obtained by combining Eqs. 2.23, 2.18 and 2.20. For any $\check{\boldsymbol{m}} \in \mathcal{N}_A$:

$$\mathcal{D}^1_{\mathrm{H},A} = \mathcal{G}_{>0}(A) \text{ and } \mathcal{D}^1_{\mathrm{H},A}\rfloor\check{\boldsymbol{m}} = \left\{f \in \mathcal{G}(A) : \sum_{x\in A} f(x)\check{\boldsymbol{m}}_x > 0\right\} \cup \mathcal{G}_{>0}(A). \tag{2.24}$$

It turns out that the expressions for the corresponding lower previsions are much more manageable. In particular, for $\check{\boldsymbol{m}} = \boldsymbol{0}$:

$$\underline{P}^{\hat{n}}_{\mathrm{H},A}(f) = \min_{x\in A} f(x, x, \ldots, x) \text{ for all } f \in \mathcal{G}(A^{\hat{n}}), \tag{2.25}$$

and for any $\check{\boldsymbol{m}} \in \mathcal{N}_A$:

$$\underline{P}^{\hat{n}}_{\mathrm{H},A}(f|\check{\boldsymbol{m}}) = \overline{P}^{\hat{n}}_{\mathrm{H},A}(f|\check{\boldsymbol{m}}) = P^{\hat{n}}_{\mathrm{H},A}(f|\check{\boldsymbol{m}}) = \sum_{\boldsymbol{n}\in\mathcal{N}^{\hat{n}}_A} \mathrm{Hy}^{\hat{n}}_A(f|\boldsymbol{n}) \binom{\hat{n}}{\boldsymbol{n}} \frac{\prod_{x\in A} \check{m}_x^{(n_x)}}{\check{m}_A^{(\hat{n})}}. \tag{2.26}$$

For the immediate prediction models, we find that for any $\check{\boldsymbol{m}} \in \mathcal{N}_A$:

$$\underline{P}^1_{\mathrm{H},A}(f) = \min f \text{ and } P^1_{\mathrm{H},A}(f|\check{\boldsymbol{m}}) = \sum_{x\in A} f(x)\frac{\check{\boldsymbol{m}}_x}{\check{\boldsymbol{m}}_A} \text{ for all } f \in \mathcal{G}(A), \tag{2.27}$$

The precise posterior predictive previsions in Equation 2.26 are exactly the ones that would be found were we to formally apply Bayes's rule with a multinomial likelihood and *Haldane's improper prior* [14, 15, 17], whose 'density' is a function on $\mathrm{int}(\Sigma_A)$ proportional to $\prod_{x\in A}\theta_x^{-1}$. This, of course, is why we use Haldane's name for the inference system that produces them. Our argumentation shows that there is nothing wrong with these posterior predictive previsions, as they are based on coherent inferences. In fact, our analysis shows that there is an infinity of *precise and proper* priors on the simplex Σ_A that, together with the multinomial likelihood, are coherent with these posterior predictive previsions: every linear prevision on

$\mathcal{V}(A)$ that dominates the coherent lower prevision $\underline{H}_{\mathrm{H},A}$ on $\mathcal{V}(A)$,[5,6] as defined by $\underline{H}_{\mathrm{H},A}(p) := \sup\{\mu \in \mathbb{R} : p - \mu \in \mathcal{H}_{\mathrm{H},A}\}$ for all polynomials p on Σ_A.

2.11 Conclusion

We believe this is the first chapter that tries to deal in a systematic fashion with predictive inference under exchangeability using imprecise probability models. A salient feature of our approach is that we consistently use coherent sets of desirable gambles as our uncertainty models of choice. This allows us, in contradistinction with most other approaches in probability theory, to avoid problems with determining unique conditional models from unconditional ones when conditioning on events with (lower) probability zero. A set of polynomials $\mathcal{H}_A$ completely determines all prior and posterior predictive models $\mathcal{D}_A^{\hat{n}}\rfloor\check{\boldsymbol{m}}$ and $\underline{P}_A^{\hat{n}}(\,\cdot\,|\check{\boldsymbol{m}})$, even when the (lower) prior probability $\underline{P}_A^{\check{n}}([\check{\boldsymbol{m}}]) = \underline{H}_A(\mathrm{B}_{A,\check{\boldsymbol{m}}})$ of observing the count vector $\check{\boldsymbol{m}}$ is zero. An approach using only lower previsions and probabilities would make this much more complicated and involved, if not impossible. Indeed, it can be proved that any inference system that satisfies representation insensitivity has near-vacuous prior predictive models, and that therefore its prior predictive lower previsions must satisfy $\underline{P}_A^{\check{n}}([\check{\boldsymbol{m}}]) = 0$. This simply means that it is *impossible* in a representation insensitive inference system for the prior lower previsions to uniquely determine posteriors. And therefore, any systematic way of dealing with such inference systems must be able to resolve—or deal with—this non-unicity in some way. We believe our approach involving coherent sets of desirable gambles is one of the mathematically most elegant ways of doing this.

We might also wonder whether there are other representation insensitive and specific inference systems. We suggest, as candidates for further consideration, the inference systems that can be derived using Walley's bounded derivative model [27], and inference systems that can be constructed using sets of infinitely divisible distributions, as recently proposed by Mangili and Benavoli [20].

Acknowledgements Gert de Cooman's research was partially funded through project number 3G012512 of the Research Foundation Flanders (FWO). Jasper De Bock is a PhD Fellow of the Research Foundation Flanders and wishes to acknowledge its financial support. Marcio Diniz was supported by FAPESP (São Paulo Research Foundation), under the project 2012/14764-0 and wishes to thank the SYSTeMS Research Group at Ghent University for its hospitality and support during his sabbatical visit there.

[5] Actually, a suitably adapted version of coherence, where the gambles are restricted to the polynomials on Σ_A.

[6] It is an immediate consequence of the F. Riesz extension theorem that each such linear prevision is the restriction to polynomials of the expectation operator of some unique σ-additive probability measure on the Borel sets of Σ_A; see for instance [6].

References

1. Augustin, T., Coolen, F.P.A., de Cooman, G., Troffaes, M.C.M. (eds.): Introduction to Imprecise Probabilities. Wiley (2014)
2. Bernard, J.M.: Bayesian analysis of tree-structured categorized data. Revue Internationale de Systémique **11**, 11–29 (1997)
3. Bernard, J.M.: An introduction to the imprecise Dirichlet model for multinomial data. Int. J. Approx. Reason **39**, 123–150 (2005)
4. Cifarelli, D.M., Regazzini, E.: De Finetti's contributions to probability and statistics. Stat. Sci. **11**, 253–282 (1996)
5. Couso, I., Moral, S.: Sets of desirable gambles: conditioning, representation, and precise probabilities. Int. J. Approx. Reason **52**(7), 1034–1055 (2011)
6. de Cooman, G., Miranda, E.: The F. Riesz representation theorem and finite additivity. In: D. Dubois, M.A. Lubiano, H. Prade, M.A. Gil, P. Grzegorzewski, O. Hryniewicz (eds.) Soft Methods for Handling Variability and Imprecision (Proceedings of SMPS 2008), pp. 243–252. Springer (2008)
7. de Cooman, G., Miranda, E.: Irrelevant and independent natural extension for sets of desirable gambles. J. Artif. Intell. Res. **45**, 601–640 (2012). http://www.jair.org/vol/vol45.html
8. de Cooman, G., Quaeghebeur, E.: Exchangeability and sets of desirable gambles. Int. J. Approx. Reason **53**(3), 363–395 (2012). (Special issue in honour of Henry E. Kyburg, Jr.).
9. de Cooman, G., Miranda, E., Quaeghebeur, E.: Representation insensitivity in immediate prediction under exchangeability. Int. J. Approx. Reason **50**(2), 204–216 (2009). doi:10.1016/j.ijar.2008.03.010
10. de Cooman, G., Quaeghebeur, E., Miranda, E.: Exchangeable lower previsions. Bernoulli **15**(3), 721–735 (2009). doi:10.3150/09-BEJ182. http://hdl.handle.net/1854/LU-498518
11. de Finetti, B.: La prévision: ses lois logiques, ses sources subjectives. Annales de l'Institut Henri Poincaré **7**, 1–68 (1937). (English translation in [18])
12. de Finetti, B.: Teoria delle Probabilità. Einaudi, Turin (1970)
13. de Finetti, B.: Theory of Probability: A Critical Introductory Treatment. Wiley, Chichester (1974–1975). (English translation of [12], two volumes)
14. Haldane, J.B.S.: On a method of estimating frequencies. Biometrika **33**, 222–225 (1945)
15. Jeffreys, H.: Theory of Probability. Oxford Classics series. Oxford University Press (1998). (Reprint of the third edition (1961), with corrections)
16. Johnson, N.L., Kotz, S., Balakrishnan, N.: Discrete Multivariate Distributions. Wiley Series in Probability and Statistics. Wiley, New York (1997)
17. Jaynes, E.T.: Probability Theory: The Logic of Science. Cambridge University Press (2003)
18. Kyburg Jr., H.E., Smokler, H.E. (eds.): Studies in Subjective Probability. Wiley, New York (1964). (Second edition (with new material) 1980)
19. Lad, F.: Operational Subjective Statistical Methods: A Mathematical, Philosophical and Historical Introduction. Wiley (1996)
20. Mangili, F., Benavoli, A.: New prior near-ignorance models on the simplex. In: F. Cozman, T. Denœux, S. Destercke, T. Seidenfeld (eds.) ISIPTA '13 – Proceedings of the Eighth International Symposium on Imprecise Probability: Theories and Applications, pp. 213–222. SIPTA (2013)
21. Moral, S.: Epistemic irrelevance on sets of desirable gambles. Ann. Math. Artif. Intell. **45**, 197–214 (2005). doi:10.1007/s10472-005-9011-0
22. Quaeghebeur, E.: Introduction to Imprecise Probabilities, (Chapter: Desirability). Wiley (2014)
23. Quaeghebeur, E., de Cooman, G., Hermans, F.: Accept & reject statement-based uncertainty models. Int. J. Approx. Reason. (2013). (Submitted for publication)
24. Rouanet, H., Lecoutre, B.: Specific inference in ANOVA: From significance tests to Bayesian procedures. Br. J. Math. Stat. Psychol. **36**(2), 252–268 (1983). doi:10.1111/j.2044-8317.1983.tb01131.x

25. Walley, P.: Statistical Reasoning with Imprecise Probabilities. Chapman and Hall, London (1991)
26. Walley, P.: Inferences from multinomial data: learning about a bag of marbles. J. R. Stat. Soc., Series B **58**, 3–57 (1996). (With discussion)
27. Walley, P.: A bounded derivative model for prior ignorance about a real-valued parameter. Scand. J. Stat. **24**(4), 463–483 (1997). doi:10.1111/1467-9469.00075
28. Walley, P.: Towards a unified theory of imprecise probability. Int. J. Approx. Reason **24**, 125–148 (2000)
29. Walley, P., Bernard, J.M.: Imprecise probabilistic prediction for categorical data. Tech. Rep. CAF-9901, Laboratoire Cognition et Activitées Finalisées, Université de Paris 8 (1999)
30. Williams, P.M.: Indeterminate probabilities. In: M. Przelecki, K. Szaniawski, R. Wojcicki (eds.) Formal Methods in the Methodology of Empirical Sciences, pp. 229–246. Reidel, Dordrecht (1976). (Proceedings of a 1974 conference held in Warsaw)

Chapter 3
Bayesian Learning of Material Density Function by Multiple Sequential Inversions of 2-D Images in Electron Microscopy

Dalia Chakrabarty and Shashi Paul

Abstract We discuss a novel inverse problem in which the data is generated by the sequential contractive projections of the convolution of two unknown functions, both of which we aim to learn. The method is illustrated using an application that relates to the multiple inversions of image data recorded with a scanning electron microscope, with the aim of learning the density of a given material sample and the microscopy correction function. Given the severe logistical difficulties in this application of taking multiple images at different viewing angles, a novel imaging experiment is undertaken, resulting in expansion of information. In lieu of training data, it is noted that the highly discontinuous material density function cannot be modelled using a Gaussian process (GP) as the parametrisation of the required nonstationary covariance function of such a GP cannot be achieved without training data. Consequently, we resort to estimating values of the unknown functions at chosen locations in their domain—locations at which an image data are available. Image data across a range of resolutions lead to multiscale models which we use to estimate material densities from the micrometre to nanometre length scales. We discuss applications of the method in nondestructive learning of material density using simulated metallurgical image data, as well as perform inhomogeneity detection in multicomponent composite on nanometre scales, by inverting real image data of a brick of nanoparticles.

D. Chakrabarty (✉)
Department of Statistics, University of Warwick, Coventry CV4 7AL, UK

Department of Mathematics, University of Leicester, Leicester LE1 7RH, UK
e-mail: d.chakrabarty@warwick.ac.uk

S. Paul
Emerging Technologies Research Centre, De Montfort University, The Gateway, Leicester LE1 9BH, UK
e-mail: spaul@dmu.ac.uk

A. Polpo et al. (eds.), *Interdisciplinary Bayesian Statistics,*
Springer Proceedings in Mathematics & Statistics 118, DOI 10.1007/978-3-319-12454-4_3

3.1 Introduction

In general, an inverse problem entails the estimation of the unknown model parameter vector $\boldsymbol{\rho} \in \mathbb{R}^d$, given available data that comprises measurements of the variable $\boldsymbol{I}$. In some circumstances, the measurable is vector-valued, ($\boldsymbol{I} \in \mathbb{R}^m$), while in other applications it can also be tensor-valued.

The measurements would be noisy, with $\boldsymbol{\epsilon}$ representing the noise in the data. Then, defining the measurable as a function of the unknown model parameter, we write $\boldsymbol{I} = \boldsymbol{\xi}(\boldsymbol{\rho}) + \boldsymbol{\epsilon}$, where this functional relationship between the measurable $\boldsymbol{I}$ and model parameter $\boldsymbol{\rho}$ is $\boldsymbol{\xi}(\cdot)$. Then, estimation of the unknown $\boldsymbol{\rho}$ is given in the classical paradigm by operating the inverse of $\boldsymbol{\xi}(\cdot)$ on the data. However, the fatal flaw in this plan to estimate $\boldsymbol{\rho}$ is that in real-life situations $\boldsymbol{\xi}(\cdot)$ is not necessarily known.

We advance a classification of inverse problems: inverse problems of Type I are those in which $\boldsymbol{\xi}(\cdot)$ is known, distinguished from inverse problems of Type II in which $\boldsymbol{\xi}(\cdot)$ is unknown. As far as inverse problems of Type II are concerned, as mentioned above, these problems are characterised by the ambition of estimating an unknown model parameter vector, the relation of which to the data is also unknown [1, 6, 8, 20, 25–27]. In fact, on first glance, such a problem appears difficult and is perhaps more conventionally treated as a general modelling challenge within the domain of application. However, treating this as an inverse problem suggests requiring to learn $\boldsymbol{\xi}(\cdot)$, in addition to estimating the unknown $\boldsymbol{\rho}$ [4]. In principle, $\boldsymbol{\rho}$ can be estimated following learning $\boldsymbol{\xi}(\cdot)$ by fitting splines or wavelets to training data and then inverting this learnt function to operate it on the measured values of $\boldsymbol{I}$, or by modelling $\boldsymbol{\xi}(\cdot)$ with a Gaussian process (GP). However, in high-dimensions, fitting using splines or wavelets, and particularly of inversion of the fit function, becomes difficult; splines and wavelets also fail to capture the correlations between components of the sought function in high-dimensions. At the same time, it is difficult to achieve parametrisation of the covariance kernels of a high-dimensional GP–especially when the application motivates a nonstationary covariance for this GP, i.e. if the sought function enjoys discontinuous support.

However, the crucial point to realise is that neither fitting of splines/wavelets, nor modelling with a GP is feasible in the absence of training data; it is after all to training data that the fitting can be performed. By "training data", is implied data that comprises computed/measured value $\boldsymbol{i}^\star$ of $\boldsymbol{I}$ at known or chosen value $\boldsymbol{\rho}^\star$ of $\boldsymbol{\rho}$, i.e. the data set $\{(\boldsymbol{\rho}_k^\star, \boldsymbol{i}_k^\star)\}_{k=1}^n$. Therefore, in applications marked by the unavailability of training data, these methods are not useful unless simplifications are included in the modelling. Such can alternatively be addressed via a state space-based modelling strategy in which the likelihood is written in terms of the state space density, into the support of which, the unknown model parameter $\boldsymbol{\rho}$ is embedded [4]. An application of this method to missing data was discussed by [4].

The lack of training data can plague even inverse problems of Type I, for which the known functional relation between data and model parameters is known [2, 3, 11, 13, 22]. An example of such an inverse problem is a known projection of the model parameter vector results in the measurable. Inversion of the data is possible

in principle but the difficulty of such inversion depends on the complexity of the projection operation in question. In the Bayesian framework, the posterior probability of an unknown system property or system behaviour—represented by the function $\rho(\boldsymbol{X})$—given the available measurements of its projection ($\boldsymbol{I}$), is computed; here $\boldsymbol{X} \in \mathbb{R}^p$. We consider the fiduciary case of $\rho(\cdot)$ to be scalar-valued, but the sought system behaviour could be a high-dimensional function as well. One way of learning this sought system function $\rho(\boldsymbol{X})$ could be to model it using a GP of appropriate dimensionality. However, such an attempt stands thwarted when training data is not available. Even when such data exists, parametrising the covariance kernel of the GP when $\rho(\cdot)$ is not compactly supported, is acutely difficult. Basically, then one needs to forgo the very ambition of attempting to learn the correlation structure of the unknown function $\rho(\cdot)$ and has to settle for the learning of values of $\rho(\boldsymbol{X})$ at chosen values of $\boldsymbol{X}$. This can be achieved by discretising the range of values that $\boldsymbol{X}$ takes in the application under consideration and defining $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_k)^T$ where $\rho_i := \rho(\mathbf{x})$ for $\mathbf{x} \in [\mathbf{x}_{i-1}, \mathbf{x}_i)$, $i = 1, \ldots, k$. Thus, it is the values of the system function at chosen points in its support that then defines the sought model parameter vector $\boldsymbol{\rho}$.

In this paper, we present a particularly hard inverse problem of Type I; see [5]. First, we will find that the data results from a sequence of projections of the convolution of two unknown functions. Second, there is no training data available. Third, the unknown that depicts the intrinsic system property, is known to be highly discontinuous, multimodal, sparse in some examples while dense in others, and convex or concave. Finally, on the other unknown function, there is sometimes low and sometimes high levels of information available. We will address this inverse problem in a Bayesian framework, and attempt to address the various nuances of this hard problem by attempting a novel experiment that allows for expansion of information, developing priors on the sparsity of the unknown that represents an intrinsic system property, building less and more informed models of the other unknown using elicitation from the literature and finally, by pragmatically resorting to the learning of values of the unknown functions at discrete points in their respective supports, in place of trying to learn the functions themselves. We allow the discreteness of the data to propel the choice of discreteness in the support of the unknown functions. Thus, variation in the scale of resolution at which data are available gives rise to mutiscale models. We recognise the convolution of the unknown functions to be unique in our framework of expanded information. The particular example that will embody this harder-than-usual inverse problem, pertains to an application in electron microscopy.

3.2 Application at Hand

In the application, we aim to learn the density function of a given (cuboidal) slab of a material, by inverting images of this system taken with electron microscopes. In electron microscopy, a system is imaged using a beam of electrons that is made

incident of the system [7, 14, 23]. A beam of electrons is made incident on the (ith point on the) surface of the material slab; let there be a total of N_I such incidences on the system surface, i.e. $i = 1, \ldots, N_I$. Once the electron beam impinges on the surface, the electrons of the beam, owing to their energy, can penetrate the surface and travel over a distance inside the bulk of the material slab, before coming to a halt. This depth of penetration depends on the energy E of the electrons, as well as on the material constitution of the sample (atomistic parameters of the constituent materials such as its atomic number $\mathcal{Z}$ and weight, and the mass per unit volume of the materials). In fact, models have been fit to the penetration depths realised in simulation studies so that we know the maximal distance that electrons of a given energy ($E = \epsilon$) can go up to is $\propto \epsilon^{1.65}$ [12], where the constant of proportionality is a function of the material parameters.

These studies also indicate that the shape that is carved out in the interior of the 3-D material slab, by the intrusive electrons, resembles a hemi-spherical bowl for materials with high values of $\mathcal{Z}$. For lower $\mathcal{Z}$ materials, the shape is more like a pear. We assume in our modelling that the shape sculpted out by the electrons of the incident beam is hemispherical; the radius of this hemispherical region is known (proportional to $\epsilon^{1.65}$) and its centre is at the point of incidence of the beam. Thus, the more energetic the electrons in the incident beam, the bigger is the volume of this hemispherical region. This region signifies the 3-D volume within which atomistic interactions are taking place between the beam electrons and the material particles; therefore, microscopists refer to this region as the "interaction-volume".

These atomistic interactions inside any interaction-volume give rise to radiations of different types. The radiation generated within an interaction-volume then escapes the bulk of the material slab, by making its way out of the surface. At the surface, the radiation is captured by a detector placed at the point of incidence of the beam, i.e. the centre of the hemispherical interaction-volume. The detector captures this radiation in the form of a pixel on the 2-D image of the system taken with the scanning electron microscope (SEM), of which the detector is a component. In other words, a contractive projection of the emerging radiation, onto the centre of the interaction-volume, is captured as a measurable. Actually, as the radiation generated inside the interaction-volume makes its way out, it gets modulated via further interactions with the material molecules; such modulation is modelled as convolution of the radiation intensity with the "microscopy correction function" or a kernel. It is the projection of the convolution of the radiation density and the kernel that is captured as the pixel on the image taken with an SEM. Thus, the convolution of the unknown material density and the unknown microscopy correction function, onto the centre of the interaction-volume, gives rise to an image datum. Here, the density of the radiation generated at any point inside the interaction-volume is proportional to the material density function at this point.

It is customary in image inversion (aimed at the estimation of a system property) that multiple images be generated by varying an imaging parameter—usually the angle at which the image is taken. This ties in directly with the concept of Radon transform [10, 13] of a function $f(\cdot)$ that enjoys compact support in

$\mathbb{R}^n$; the Radon transform gives the projection of this function onto the hyperplane p that is inclined at given angle ϕ and is at distance $\boldsymbol{\psi}$ from the origin: $\mathcal{R}[f](\phi, \boldsymbol{\psi}) := \int_p f(\boldsymbol{x})\delta(p - \boldsymbol{\psi} \cdot \boldsymbol{x})d\boldsymbol{x}$. The inverse Radon transform is also defined but it involves taking spatial derivative of the projection, i.e. the image data in this application. Therefore, if the image data is discontinuous, the inversion may not be numerically stable. Such is the case in our application which is also characterised by lack of compact support of the sought material density function. Li and Speed [15] suggested that noise in the image data can also render the inversion unstable. Importantly, the implementation of the Radon transform requires the viewing angle (ϕ) as an input; this may not be known or may not be a measurable in certain applications. Panaretos [19] discusses one such situation while in our application, the measurement of the viewing angle is logistically cumbersome and difficult enough to warrant its measurement and variation impossible. First, the sample will have to be seated on an inclined stub for its orientation with respect to the electron beam to be rendered nonzero. This tilt, however, causes the distribution of the electronic paths inside the material slab to become skewed towards one side and no theoretical models of this lopsided interaction-volume is known in the microscopy literature (as the lopsidedness depends on the tilt angle). Moreover, installing a differently inclined stub into the imaging chamber of the SEM implies breaking the vacuum that is required for SEM imaging experiments. This is most cumbersome for the practising micoscopist. Finally, it is nearly an impossible task to reconfigure the electron beam to target the exact same area on the newly inclined sample, as the targeted area on the originally inclined sample. Markoe and Qunito and Rullgård [16, 24] discussed possible numerical instability of the inverse Radon transform given limited angle images. Thus, we realise that implementation of the inverse Radon transform of the image data is not acceptable in our application. However, it is imperative that multiple images be taken of our material sample, by varying some imaging parameter, which we realise cannot be the angle of inclination of the electron beam to the sample surface (as established above). We suggest a novel way of acquiring multiple images: by imaging the same system at N_E number of different beam energies. The image taken by impinging an identified part of the sample at N_I number of points, with a beam of electrons with energy $E = \epsilon_j$ forms the jth image of this system—comprising N_I pixels. Then, the image of the same system taken at $E = \epsilon_{j+1}$ is the $j + 1$th image with N_I pixels in it, but this time, the ith pixel manifests information that comes from an interaction-volumes of radius $R0^{(j+1)}$, as distinguished from the ith pixel in the jth image, which manifests information coming from a concentric interaction-volume of radius $R0^{(j)}$, with $R0^{(j)} < R0^{(j+1)}$. In principle, intercomparison of the intensity of the ith pixel in the jth and $j + 1$th images allows us a method of performing piecewise construction of the sub-surface density function. Here, $i = 1, \ldots, N_I$ and $j = 1, \ldots, N_E$.

3.3 Modelling

Above we discussed the generation of the $I_i^{(j)}$th image datum in the ith pixel of the image taken with the jth value of beam energy; $I_i^{(j)}$ results from a projection onto the centre of the interaction-volume, of the convolution of the unknown material density $\rho(\boldsymbol{X})$ and the microscopy correction function $\eta(\boldsymbol{X})$ where $\boldsymbol{X} = (X_1, X_2, X_3)^T$. Here, the cuboidal system is spanned by the 3-D Cartesian coordinate system with axes X_1, X_2, X_3 where the X_3-axis is along the sub-surface depth direction that is orthogonal to the system surface on the $X_1 - X_2$ plane. Now, the centre of the hemispherical interaction-volume is an incidence point of the electron beam. Projection onto this point is a composition of three sequential, contractive, orthogonal and commutable projections. For example, the projection $\mathcal{C}$ onto a beam incidence point could be due to projection $\mathcal{C}_1$ onto the $X_1 - X_2$ plane, followed by projection $\mathcal{C}_2$ onto the X_1 axis, followed by projection $\mathcal{C}_3$ onto the centre of the hemispherical interaction-volume. Thus, $\mathcal{C} = \mathcal{C}_1 \circ \mathcal{C}_2 \circ \mathcal{C}_3$.

The ijth interaction-volume is the one generated by the beam of energy $E = \epsilon_j$, incident on the ith incidence point; thus its radius is $R0^{(j)}$. We define the projection of the material density and kernel inside the ijth interaction-volume to be $(\rho * \eta)_i^{(j)}$,

$\forall\ i = 1, \ldots, N_I,\ \ j = 1, \ldots, N_E$. Then, the classical representation of this inverse problem is $I_i^{(j)} = \mathcal{C}[(\rho * \eta)_i^{(j)}] + \epsilon_I^{(j)}$, where $\epsilon_I^{(j)}$ represents the error in the measurement of the ijth image datum. The classical picture then motivates three sequential inversions of the image data to have to be performed for each i and j in order for the convolution of $(\rho * \eta)_i^{(j)}$ to be estimated. Thereafter, we will require deconvolution of the estimated $\{(\rho * \eta)_i^{(j)}\}_{i,j}$ so that we have individual estimates of the two unknown functions. This is what renders this a harder-than-usual inverse problem. However, we attempt Bayesian inference of the unknown functions as we discuss below. The projection is clarified as:

$$\mathcal{C}(\rho * \eta)_i^{(j)} = \frac{\int_{R=0}^{R0^{(k)}} \int_{\psi=0}^{\phi_{max}} R dR d\psi \int_{x_3=0}^{x3_{max}(R,\psi)} \rho(\mathbf{s}) * \eta(\mathbf{s}) dx_3}{\int_{R=0}^{R_{max}} \int_{\psi=0}^{\psi_{max}} R dR d\psi} \tag{3.1}$$

for $i = 1, \ldots, N_I$. Here, the vector $\mathbf{s}$ represents value of displacement from the point of incidence $(x1_i, x2_i, 0)$, to a general point (x_1, x_2, x_3), inside the interaction-volume, i.e.

$$\begin{aligned} \mathbf{s} &:= (x_1 - x1_i, x_2 - x2_i, x_3)^T = (R\cos\psi, R\sin\psi, z)^T \\ R &= \sqrt{(x_1 - x1_i)^2 + (x_2 - x2_i)^2} \quad \tan\psi = \frac{x_2 - x2_i}{x_1 - x1_i}. \end{aligned} \tag{3.2}$$

Here, the ith point of incidence is $(x1_i, x2_i, 0)$, $x1_i := i$ modulo $(\sqrt{N_I})$, $x2_i := \text{int}(i/\sqrt{N_I}) + 1$ for $i = 1, \ldots, N_I$ and $j = 1, \ldots, N_E$.

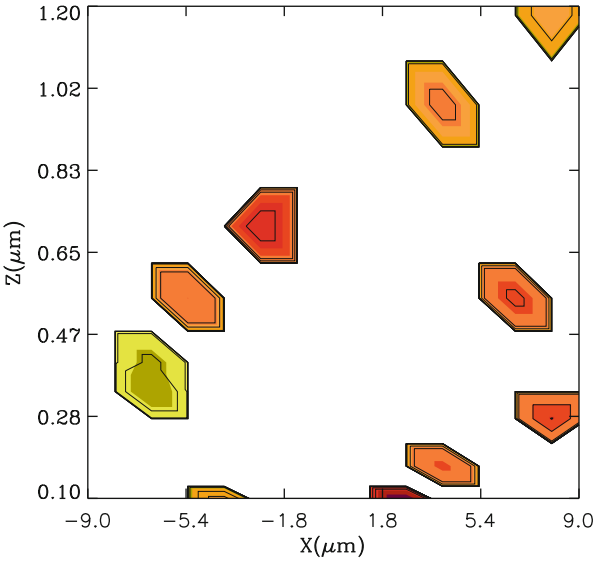

Fig. 3.1 Unfilled contour plot (in *black solid lines*) of material density parameters estimated using synthetic image data that was generated by projecting the convolution of a chosen sparse density and correction function in a simulation of SEM imaging carried out at a coarse resolution (chosen to simulate imaging by X-rays). The estimates on the $X_1 = 0$ plane are displayed. The true material density function is depicted in the filled contours

3.3.1 Why We Discretise the Domain of the Unknown Functions

Typical SEM images reveal that the image data is highly discontinuous; see Fig. 3.4 for the real SEM image taken by us of a brick of nanoparticles that was deposited in the laboratory. This indicates that the underlying material density function and/or the microscopy correction function are/is highly discontinuous. Microscopy literature however includes simulation models that are available for the correction function for certain, known material samples are imaged [7, 17]. Such shapes do not indicate discontinuity. This motivates us to consider the discontinuity observed in image data to be generally due to the discontinuous nature of the material density function. Then in principle, we could model the discontinuous $\rho(X_1, X_2, X_3)$ with a GP, which will have to be ascribed a nonstationary covariance function. That too, the level of discontinuity in $\rho(\boldsymbol{X})$ can be high—typically characterised by few isolated islands of over-density (see Figs. 3.1 and 3.4). The sought GP required to model such a function then needs to have covariance kernels that suffer discontinuous spatial variation. Modelling a nonstationary covariance function with kernels that exhibit nonsmooth variation is hard, especially when no training data is available—for a new material sample, $\rho(\boldsymbol{X})$ is not known a priori at any value of $\boldsymbol{X}$. When time replicated data are available, Paciorek and Schervish [18] offer a prescription for a globally nonstationary (but locally stationary) covariance function marked by smooth spatial variation of the kernels. Suggesting a more general nonstationary covariance structure in the presence of training data is a topical challenge in statistics that is under consideration.

In lieu of an available solution, we resort to discretising the range of values of $\boldsymbol{X}$ and learn the material density at these values. Thus, with no training data, it is not possible to model the abruptly varying spatial correlation structure that underlies

the highly discontinuous, trivariate density function of real-life material samples, so that prediction at other values of $\boldsymbol{X}$ is not possible either. We allow the resolution (ω) available in the data to dictate the discretisation of $\rho(\boldsymbol{X})$. Here, ω is the length over which structure is resolved in the image data. Thus, ω cannot be less than the diameter of the beam of electrons that is used to image with an SEM. In practice, SEM images can have very fine resolution (when imaged in the radiation of back scattered electrons, $\omega \sim$ nanometres, or nm) or can have coarse resolution (when imaged in X-rays, $\omega \sim$ micrometres, or μm).

3.3.2 A General Voxel and a General Interaction-Volume

Thus, we consider $\rho(x_1, x_2, x_3)$ to be a constant within a voxel where the ijth voxel is defined by $x_1 \in [x1_i, x1_i + \omega)$, $x_2 \in [x2_i, x2_i + \omega)$, $x_3 \in [R0^{(j-1)}, R0^{(j)})$, where $R0^{(0)}$=0, the ith point of incidence of the beam is $(x1_i, x2_i, 0)$, $x1_i :=$ i modulo $(\sqrt{N_I})$, $x2_i :=$ int$(i/\sqrt{N_I}) + 1$. Here $i = 1, \ldots, N_I$ and $j = 1, \ldots, N_E$.

We refer to the material density in the ijth voxel as ρ_{ij}. In order to achieve identifiability of solutions for the two unknown functions, we set the microscopy correction function to be a function of the depth coordinate X_3 alone, i.e. the correction function is modelled as independent of beam incidence location. Then, the range of X_3 values is discretised and for $x_3 \in [R0^{(j-1)}, R0^{(j)})$, $\eta_j := \eta(x_3)$. Thus, there are $N_I \times N_E$ number of material density parameters and N_E number of correction function parameters that we estimate.

The ijth interaction-volume on the other hand is the hemispherical interaction-volume with centre at $(x1_i, x2_i, 0)$, generated by the electron beam of energy $E = \epsilon_j$, i.e. has a radius of $R0^{(j)} \propto \epsilon_j^{1.65}$ where the constant of proportionality is material dependent. Thus, the size of the lateral cross-section (parallel to the $X_3 = 0$ plane) of the interaction-volume is given by the energy of the electrons in the beam and the material at hand while the size of a voxel is determined by ω, the resolution of the data.

3.3.3 Multi-Scale Models

Depending on the resolution of the image data, ω, there can be multiple or even less than a whole voxel, the lateral cross-section of which fits inside that of the ijth interaction-volume. When ω is large enough to warrant the latter for all j, the material density inside the ijth interaction-volume, for a given x_3, is the constant material density inside the enveloping voxel. In that case, integration of $\rho * \eta$ over the plane at this x_3 inside this interaction volume is integration of a constant. This is easy compared to the other case (of fine resolution) when multiple voxels fit inside the interaction-volume; in that case, this integration in not over a constant integrand. Thus, we can ease the difficulty of computation of the projection integral (Eq. 3.1) in

the case of coarse resolution but not when the resolution is fine. This differentiation corresponds to multiscale models differentiated by how the computation of $\mathcal{C}(rho * \eta)_{ij}$ is carried out (Eq. 3.1). Our three models are:

1. Model 1: $\omega \sim \mu$m and high $\mathcal{Z}$ to ensure that $\pi(R0^{(k)})^2 \leq \omega^2$, $\forall\, k = 1, \ldots, N_{\text{eng}}$. Integration over the angular coordinate (ψ in Eq. 3.1) is not required.
2. Model 2: $\pi(R0^{(k)})^2 \leq \omega^2$, $\forall\, k = 1, \ldots, k_{in}$ and $\pi(R0^{(k)})^2 > \omega^2$, $\forall\, k = k_{in} + 1, \ldots, N_{\text{eng}}$. To avoid performing the integration over ψ in Eq. 3.1, we compute the nearest neighbour averaging over density.
3. Model 3: $\pi(R0^{(k)})^2 > \omega^2$, $\forall\, k = 1, \ldots, N_{\text{eng}}$. Integration over ψ cannot be avoided.

3.4 Priors on Sparsity of Material Density Function

We develop priors on the material density function, such that the priors adapt to the density. One way to check for voxels that contain zero density is to check if $\mathcal{C}(\rho*\eta)_i^{(j)} = \mathcal{C}(\rho*\eta)_{-im}^{(j)}$ (where $\mathcal{C}(\rho*\eta)_{-im}^{(j)}$ is the projection onto the centre of the ijth interaction-volume performed without considering density in the mjth voxel). If it holds, then $\rho_{mj} = 0$. But computationally speaking, such checking is very expensive as it entails as many checks as voxels inside this interaction-volume. Rather, we check if a voxel has zero density at a chosen probability. Therefore, we check if the following statement is true with some probability:

$$\mathcal{C}(\rho * \eta)_i^{(j)} \leq \mathcal{C}(\rho * \eta)_i^{(j-1)} \Longrightarrow \rho_{ij} = 0 \tag{3.3}$$

$\forall\, i, j$. Define the random variable τ_{ij}, with $0 < \tau_{ij} \leq 1$, such that

$$\tau_{ij} := \frac{\mathcal{C}(\rho * \eta)_i^{(j)}}{\mathcal{C}(\,\rho * \eta)_i^{(j-1)}}, \quad \text{if} \quad \mathcal{C}(\rho * \eta)_i^{(j)} \leq \mathcal{C}(\rho * \eta)_i^{(j-1)}, \mathcal{C}(\rho * \eta)_i^{(j-1)} \neq 0$$

$$\tau_{ij} := 1 \quad \text{otherwise.}$$

Then, Statement 3.3 is recast as: "smaller is τ_{ij}, higher is the probability $\nu(\tau_{ij})$ that ρ_{ij}=0", where $\nu(\tau_{ij})$ is

$$\nu(\tau_{ij}) = p^{\tau_{ij}}(1-p)^{1-\tau_{ij}}, \tag{3.4}$$

with the hyper-parameter p controlling the nonlinearity of response of $\nu(\tau_{ij})$ to increase in τ_{ij}. The hyper-parameter $p \sim \mathcal{U}[0.6, 0.99]$.

The prior on the density parameter ρ_{ij} is then defined as

$$\pi_0(\rho_{ij}) = \exp\Big[-\big(\rho_{ij}\,\nu(\tau_{ij})\big)^2\Big]. \tag{3.5}$$

Thus, $\pi_0(\rho_{ij}) \in$[0,1] $\forall\, i, j$. Only for i, j for which τ_{ij} is unity, is $\nu(\tau_{ij})$ a constant, $(= p)$ and only for such i, j, the prior on the material density parameter is proportional to a half-normal density distribution (with nonnegative support). For all other

i, j, the prior is not of any recognisable parametric form; nonparametric priors on sparsity are also reported, including the nonparametric prior without mixing [9]. Our prior also does not include mixing, but like other priors on sparsity, it adapts well to the sparsity in the material density function

3.5 Priors on Microscopy Correction Function

In the literature on electron microscopy, models that approximate the shape of the microscopy correction function have been advanced [17, 21] as akin to a folded-normal distribution. However, the scale of this function and details of the shape are not known a priori but need to be estimated using system-specific Monte Carlo simulations of the system or approximations based on curve-fitting techniques; see [7]. Such simulations or model-based calculations are material-specific and are based on the assumption of homogeneous material samples, amongst other assumptions. Given this situation, it is meaningful to seek to learn the correction function from the image data.

We recall that we model the correction function as dependent on X_3 alone and instead of learning the function $\eta(X_3)$, we estimate the discrete values of it such that for $x_3 \in [R0^{(j-1)}, R0^{(j)}) \implies \eta_j := \eta(x_3)$, $j = 1, \ldots, N_E$, $R^{(0)} = 0$. We develop the more-well informed model for the correction function in which it is approximated by a scaled, two-parameter folded normal density. In the less-well developed model, folded-normal priors are slapped on η_j, $\forall\, j = 1, \ldots, N_E$.

3.6 Inference

The likelihood is defined in terms of the distance between the image datum $\tilde{I}_i^{(j)}$ and the projection of the unknowns onto the centre of the ijth interaction volume. A Gaussian likelihood is chosen:

$$\mathcal{L}\left(\rho_{11}, \ldots, \rho_{1N_E}, \ldots, \rho_{N_I1}, \ldots, \rho_{N_IN_E}, \eta_1, \ldots, \eta_{N_E} | \tilde{I}_1^{(1)}, \ldots, \tilde{I}_1^{(N_E)}, \ldots, \tilde{I}_{N_I}^{(N_E)}\right) =$$

$$\prod_{j=1}^{N_E}\prod_{i=1}^{N_I} \frac{1}{\sqrt{2\pi}\sigma_i^{(j)}} \exp\left[-\frac{\left(\mathcal{C}(\rho * \eta)_i^{(j)} - \tilde{I}_i^{(j)}\right)^2}{2\left(\sigma_i^{(j)}\right)^2}\right], \tag{3.6}$$

where the noise in the image datum $\tilde{I}_i^{(j)}$ is $\sigma_i^{(j)}$.

The priors on the density and correction function parameters are invoked allowing the formulation of the joint posterior probability density of the unknown parameters, given the image data. We sample from the posterior probability density of the unknown parameters $\{\rho_{ij}, \eta_j\}$ given the image data $\{\tilde{I}_i^{(j)}\}$ using (adaptive) Metropolis within Gibbs.

Had the inference been classical, we could state that in the small noise limit, $(\rho * \eta)_i^{(j)}$ is the unique solution to the least squares problem $\tilde{I}_i^{(j)} = \mathcal{C}(\rho * \eta)_i^{(j)}$. In the presence of noise in the data, $\rho * \eta$ is no longer unique;the bound on the lack of uniqueness is given by condition number $\kappa(\mathbf{C})$ of the matrix version $\mathbf{C}$ of the operator $\mathcal{C}$. As $\mathbf{C}$ is product of orthogonal projection matrices $\kappa(\mathbf{C}) = 1$. Therefore, fractional deviation of uniqueness in learnt value of $\rho * \eta$ is the same as the noise in the image data, which is at most 5 %.

Of course, we perform Bayesian inference on our parameters and in this framework, the uncertainty on learnt parameters governed by strength of priors for a given data set. We have conducted a number of simulation studies in which we vary the different model parameters that can perturb the likelihood. Such parameters include the noise in the data and the level of sparsity in the material density function; we also vary the level of information in the priors of the correction function parameters

3.7 Results Using Simulated Data

We carry out inversion of synthetic image data generated by projecting the convolution of the material density and correction functions for simulated samples of an alloy of two chosen metals (Nickel and Tungsten). The projection is performed to simulate SEM imaging under a chosen resolution. The synthetic image data are then inverted to estimate the material density and correction function parameters of this alloy, which are then compared to the known material density and correction function of the simulated model.

For a simulated model, the true density of which is sparse, the learnt material density parameters on the $X_2 - X_3$ plane, at $X_1 = 0$, is displayed in Fig. 3.1 in unfilled contours that are superimposed upon the true density function (shown in filled contours). For another simulated model, the density of which is not sparse, the variation in the estimated density parameters is superimposed in red over the true values that are shown in black in Fig. 3.2. In both cases the imaging resolution chosen to simulate the SEM imaging is chosen to be coarse (μm), so as to render Model 1 relevant. Figure 3.3 represents the comparison of the correction function parameters estimated using a synthetic image data simulated at a slightly more fine resolution (making Model 2 relevant), and the true values. A less-informed model of the priors on the correction function parameters was used in this estimation.

3.8 Results Using Real SEM Image Data

We produced a brick of Silver and Aluminium nanoparticles in the laboratory and imaged it at 11 different beam energies. The image data generated at one of these energies is shown in the left panel of Fig. 3.4. A fraction of this image data (contained in a rectangular area spanning 101×101 pixels) is inverted to help learn the unknown parameters. The estimated density parameters are shown on a contour plot in the right panel of Fig. 3.4.

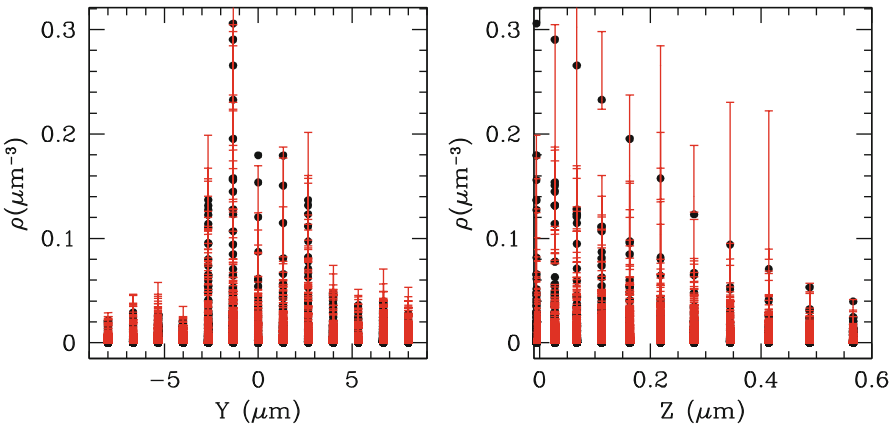

Fig. 3.2 The left and right panels respectively show variation with X_2 and X_3 of material density parameters estimated using synthetic image data that was generated by projecting the convolution of a chosen nonsparse density and correction function, in a simulation of SEM imaging carried out at a coarse resolution (chosen to simulate imaging by X-rays). The true material density is in *black* while the estimates are shown in *red*, accompanied by the 95 % HPD regions

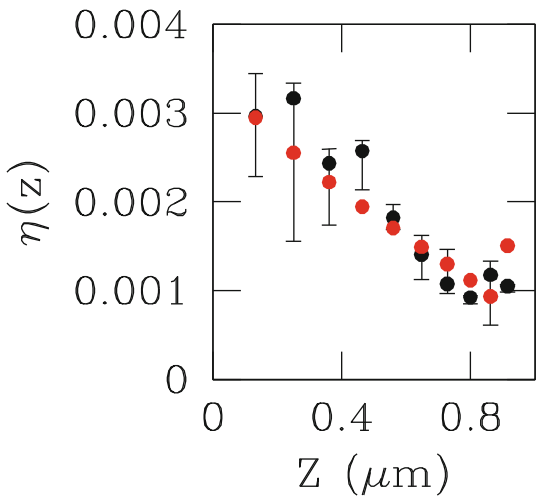

Fig. 3.3 Variation with X_3 of correction function parameters estimated using synthetic image data that was generated by projecting the convolution of a chosen density and correction function, in a simulation of SEM imaging carried out at a slightly less coarse resolution than that used to generate results in Figs. 1 and 2 (rendering our Model 2 relevant). The true parameter values are shown in *black* while the estimates are shown in *red*, accompanied by the 95 % HPD regions.

3.9 Conclusions

This inverse problem stands out in its complexity owing to the multiple inversions of the image data that are required to estimate the convolution of the two sought unknown functions; further, deconvolution of this estimate leads to solutions for the unknowns. There is no training data available for us to train a model for the density function on. Also, the material density function is anticipated as being not compactly supported in $\mathbb{R}^3$. Besides, it can be sparse in some material samples but not so in

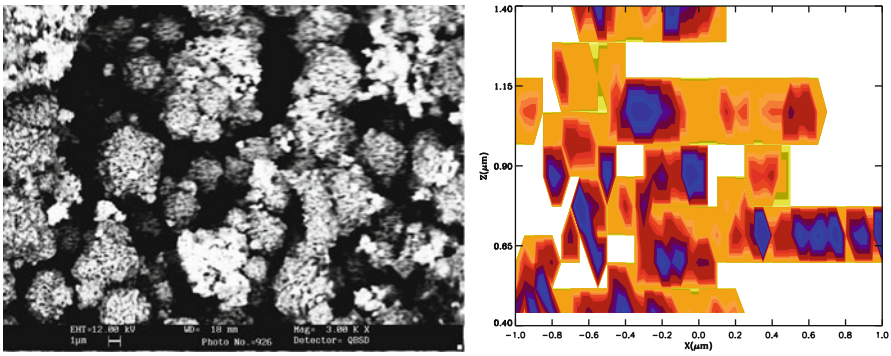

Fig. 3.4 *Left:* Image data of a brick of Silver and Aluminium nanoparticles, taken with an SEM, at beam energy of 12 keV. *Right:* Contour plot (at $X_1 = 0$) representing material density parameters estimated by inverting real image data taken at energies of 10, 11,. . . 20eV

others. We address the problem within a Bayesian framework. The salient features of the methodology advanced to tackle is inverse problem include the following.

- Novel imaging experiment: multiple images taken of same point on the sample, at distinct values of an easy-to-manipulate parameter (beam energy E) that allow for information on density structure to arrive from different sub-surface depths.
- Set correction function dependence only on sub-surface depth: $\eta(X_3)$. Known value of $\eta(0)$ helps achieve identifiability.
- Develop priors on sparsity of material density function.
- Priors on kernel are elicited from the microscopy literature.
- Discretise the unknown functions: $\rho_{ij} = \rho(x1_i, x2_i, x_3)$ if $X_1 = x1_i$, $X_2 = x2_i$ define the ith beam incidence location and $z \in [R0^{(j-1)}, R0^{(j)})$; similarly, $\eta_j = \eta(z)$ if $z \in [R0^{(j-1)}, R0^{(j)})$. Here, $i = 1, \ldots, N_I$, $j = 1, \ldots, N_E$. We perform Bayesian inference on the unknown parameters ρ_{ij} and η_j.
- Formulate multiscale models to address inversion of image data recorded at different length scales, i.e. different imaging resolutions.

Applications to real image data of a system of nanoparticles as well as to simulated metallurgical image data demonstrate the efficiency of the method across a range of length scales.

Acknowledgments The authors would like to thank Dr Nare Gabrielyan, Emerging Technologies Research Centre, De Montfort University, Leicester, UK, for performing the SEM imaging used in the work.

References

1. Bennett, A.F., McIntosh, P.C.: Open ocean modeling as an inverse problem: tidal theory. J Phys Oceanogr **12**, 1004–1018 (1982)
2. Bertero, M., Boccacci, P.: Introduction to Inverse Problems in Imaging. Taylor and Francis, London (1998)

3. Bishop, T.E., Babacan, S.D., Amizik, B., Katsaggelos, A.K., Chan, T., Molina, R.: Blind image deconvolution: problem formulation and existing approaches. In: P. Campisi, K. Egiazarian (eds.) Blind Image Deconvolution: Theory and Applications, pp. 1–41. CRC Press, Taylor and Francis, London (2007)
4. Chakrabarty, D., Saha, P.: Inverse Bayesian estimation of gravitational mass density in galaxies from missing kinematic data. Am. J. Comput. Math. **4**(1), 6–29 (2014)
5. Chakrabarty, D., Rigat, F., Gabrielyan, N., Beanland, R., Paul, S.: Bayesian density estimation via multiple sequential inversions of 2-D images with application in electron microscopy. Technometrics (2014). doi:10.1080/00401706.2014.923789
6. Draper, D., Mendes, B.: Bayesian environmetrics: uncertainty and sensitivity analysis and inverse problems. (2008). http://users.soe.ucsc.edu/draper/draper-brisbane-2008.pdf.
7. Goldstein, J., Newbury, D.E., Joy, D.C., Lyman, C.E., Echlin, P., Lifshin, E., Sawyer, L., Michael, J.: Scanning Electron Microscopy and X-ray Microanalysis. Springer, New York (2003)
8. Gouveia, W.P., Scales, J.A.: Bayesian seismic waveform inversion: parameter estimation and uncertainty analysis. J. Geophys. Res. **130**(B2), 2759 (1998)
9. Greenshtein, E., Park, J.: Application of non parametric empirical Bayes estimation to high dimensional classification. J. Mach. Learn. Res. **10**, 1687–1704 (2009)
10. Helgason, S.: The Radon Transform. Progress in Mathematics. Birkhauser, Boston (1999)
11. Jugnon, V., Demanet, L.: Interferometric inversion: a robust approach to linear inverse problems. In: J.M. Bernardo, J.O. Berger, A.P. Dawid, A.F.M. Smith (eds.) Proceedings of SEG Annual Meeting, Houston, (Sept. 2013)
12. Kanaya, K., Okamaya, S.: Penetration and energy-loss theory of electrons in solid targets. J. Phys. D., Appl. Phys. **5**(1), 43 (1972)
13. Kutchment, P.: Generalised transforms of the radon type and their applications. In: G. Olafsson, E.T. Quinto (eds.) The Radon Transform, Inverse Problems, and Tomography, vol. 63, p. 67. American Mathematical society, Providence (2006)
14. Lee, R.E.: Scanning Electron Microscopy and X-Ray Microanalysis. Prentice-Hall, New Jersey (1993)
15. Li, L., Speed, T.: Parametric deconvolution of positive spike trains. Ann. Stat. **28**, 1270 (2000)
16. Markoe, A., Quinto, E.T.: An elementary proof of local invertibility for generalized and attenuated radon transforms. SIAM J. Math. Anal. **16**, 1114 (1985)
17. Merlet, C.: An accurate computer correction program for quantitative electron probe microanalysis. Mikrochim. Acta **114/115**, 363 (1994)
18. Paciorek, C.J., Schervish, M.: Spatial modelling using a new class of nonstationary covariance functions. Environmetrics **17**, 483–506 (2006)
19. Panaretos, V.M.: On random tomography with unobservable projection angles. Ann. Stat. **37**(6), 3272 (2009)
20. Parker, R.L.: Geophysical Inverse Theory (Princeton Series in Geophysics). Princeton University Press, Princeton (1994)
21. Pouchou, J.L., Pichoir, F.: PAP (ρZ) procedure for improved quantitative microanalysis. In: J.T. Armstrong (ed.) Microbeam Analysis. San Francisco Press, San Francisco, (1984)
22. Qiu, P.: A nonparametric procedure for blind image deblurring. Comput. Stat. Data Anal. **52**, 4828–4842 (2008)
23. Reed, S.J.B.: Electron Microprobe Analysis and Scanning Electron Microscopy in Geology. Cambridge University Press, Cambridge (2005)
24. Rullgård, H.: Stability of the inverse problem for the attenuated radon transform with 180 data. Inverse Probl. **20**, 781 (2004)
25. Stuart, A.: Inverse problems: a Bayesian perspective. Acta Numerica. **19**, 451–559 (2010) (Cambridge University Press)
26. Stuart, A.: Bayesian approach to inverse problems. Provide an introduction to the forthcoming book Bayesian Inverse Problems in Differential Equations by M. Dashti, M. Hairer and A.M. Stuart; available at arXiv:math/1302.6989 (2013)
27. Tarantola, A.: Inverse Problem Theory and Methods for Model Parameter Estimation. SIAM, Philadelphia (2005)

Chapter 4
Problems with Constructing Tests to Accept the Null Hypothesis

André Rogatko and Steven Piantadosi

Abstract Futility designs have been proposed and used by constructing classical (non-Bayesian) hypothesis tests such that the decision of therapeutic interest is to accept the null hypothesis. A consequence is that the probability of accepting (failing to reject) the null when the null is false is unknown. Reversal of the conventional null and alternative hypotheses is not required either to demonstrate futility/nonsuperiority, or to align type I and II errors with their consequences. Conventional methods to test whether the response to the investigational agent is superior to a comparative control (superiority trial) are preferable and in the case that the null hypothesis is rejected, the associated type I error is known.

4.1 Introduction

In recent years, a class of formal futility designs to eliminate unpromising therapies in middle development has been proposed [1–4]. These designs are used especially for trials in neurologic disease where their properties appear to resonate with the needs of therapeutic development, and their use seems to have increased in recent years [5]. This design carries significant flaws that investigators should know before implementing.

The essence of a futility design is an effect threshold or threshold improvement below which we lose interest in further developing a treatment. The effect threshold is based on clinical considerations. Performance below the threshold stops the trial or development. If we defined a threshold for improvement, the treatment would be discarded as nonsuperior. In futility designs it is typical for the null hypothesis to be that the treatment exceeds some specified tolerance for superiority, while the alternative hypothesis is that the treatment does not meet the criterion for superiority.

A. Rogatko (✉) · S. Piantadosi
Biostatistics and Bioinformatics Research Center, Cedars-Sinai Medical Center,
Los Angeles, CA 90048, USA
e-mail: andre.rogatko@cshs.org

S. Piantadosi
e-mail: steven.piantadosi@cshs.org

A. Polpo et al. (eds.), *Interdisciplinary Bayesian Statistics,*
Springer Proceedings in Mathematics & Statistics 118, DOI 10.1007/978-3-319-12454-4_4

This is an inversion of the classical setup for the null and alternative hypothesis, and gives the design its name.

4.2 Statistical Background

The classical approach to construct a null hypothesis is through a statement of equivalence or lack of improvement. This is more than a simple convention because of the asymmetry in the way type I and II errors are managed. A hypothesis test provides certain and quantifiable control over the type I error. When we reject the null (classically when we declare that there is a significant difference or a treatment improvement) we know exactly the probability of a type I error. In other words, the chance of advancing a truly ineffective therapy is controlled at the α-level of the test. Provided we are wise about the ways that a type I error can be inflated (e.g., multiplicity), the probability of this error is always under the control of the experimenter by the choice of the α-level and critical value for the test. When we reject the null, the type II error is irrelevant.

R. A. Fisher, in his classical work published in 1935 wrote: "... the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis."

Since then, as well summarized by S. Wellek in the introduction of his book "Testing Statistical Hypotheses of Equivalence and Noninferiority" [6]: "It is a basic fact well known to every statistician that in any hypotheses testing problem there is an inherent logical asymmetry concerning the roles played by the two statements (traditionally termed null and alternative hypotheses) between which a decision shall be taken on the basis of the data collected in a suitable trial or experiment: Any valid testing procedure guarantees that the risk of deciding erroneously in favor of the alternative does not exceed some prespecified bound whereas the risk of taking a wrong decision in favor of the null hypothesis can typically be as high as 1 minus the significance level (i.e., 95 % in the majority of practical applications)."

Therefore, the null hypothesis is usually constructed by reversing the scientist's belief regarding what he/she would like to prove. The Neyman-Pearson lemma (NPL) [7], which provides the foundation of classical (non-Bayesian) hypothesis testing, assures that for a fixed type I error probability (α), the power of the test is maximized.

In the development of new therapies, the concept of designing trials to screen out ineffective therapeutic regimens (phase II clinical trials) has been established since the early work of Gehan [8]. Many single arm designs have been described and extensively used [9]. They are all constructed such that the rejection of the null hypothesis leads to the "interesting decision," e.g., H_0: the new treatment is no better than the standard treatment versus H_A: the new treatment is better than the standard treatment. As any valid frequentist test procedure, one wishes to reject the null that the new treatment is no better than the standard treatment and accept the alternative. Rejecting the null at the preset significance value alpha will assure that the probability

of rejecting the null when the null is true is bounded by alpha. NPL guarantees that the power of this test is maximal.

In summary, as Aberson [10] elaborated: "Null hypothesis significance testing do not allow for conclusions about the likelihood that the null hypothesis is true, only whether it is unlikely that null is true. ... Rejecting H_0 tells us something meaningful. Specifically, that the parameter specified in H_0 is not likely to be the parameter actually observed in the population. Failing to reject H_0 tells us we cannot rule out the value specified in H_0 as a likely value for the parameter. Keeping in mind that scientific reasoning centers on principles of falsification, it becomes clear that rejecting H_0 provides falsification whereas failing to reject H_0 does not. Using this reasoning, rejecting H_0 is valid scientific evidence whereas failing to reject H_0 is not." More succinctly, as insightfully worded by Altman and Bland [11]: "Absence of evidence is not evidence of absence."

4.3 Example

In a series of articles, Palesch, Tilley and colleagues [1, 4] introduced the concept of designing trials to screen out ineffective therapeutic regimens in the field of neurology denoting this class of trials by "futility studies." Levin [3] highlights their importance and futility trials have been conducted since in the search for better treatments for neurologic diseases, e.g., [2, 12]. The hypotheses and decision process are defined as follows [4]: "H_0: $p_{tx} - p^* > \Delta$ versus H_A: $p_{tx} - p^* < \Delta$, where p_{tx} is the hypothesized proportion of treated subjects with a favorable outcome, p^* is the reference proportion of favorable outcomes for the single-arm phase II futility study, and Δ is the "minimally worthwhile improvement" in the proportion of favorable outcomes. If we reject the null hypothesis that a "minimally worthwhile" improvement exists, we conclude the benefit of the new treatment is less than what we would want, and it is futile to proceed to further testing in a phase III trial. If we fail to reject the null hypothesis that a minimally worthwhile improvement exists, we conclude there is insufficient evidence of futility, and the treatment deserves further testing in a phase III trial to determine its efficacy."

That is, the hypotheses were reversed. The consequence of constructing the hypotheses by reversing H_0 and H_A is that by accepting (failing to reject) the null, the probability of accepting the null when the null is false is unknown (NPL tells us only that it is minimal). This explains why statistical tests are constructed so that the rejection of the null is the interesting result. Note that one can always accept the null by decreasing the sample size. Although adequate power can be sought by estimating an appropriate sample size before a trial is designed, many other assumptions and guesses are made before the trial is realized about unknown parameters (e.g., p^*).

The notion of "proving the null hypothesis" in the context of NPL has been studied for a long time [13, 14] and nowadays, the concept of equivalence and noninferiority trials are well established [6].

4.4 Discussion

The type II error given for any experiment was a hypothetical, conditioned on a specific effect size. There were many type II error probabilities, some high and some low, depending on the hypothetical effect size (i.e., a power curve). Failing to reject the null hypothesis does not inform us which specific alternative is the relevant one, and therefore does not tell us the type II error. We can only say that we have a low probability of discarding a large effect size and a higher chance of discarding a lower effect size. Thus, the hypothesis testing framework codifies an asymmetry: the type I error is knowable unconditionally and the type II error is not.

We can now see a pitfall of reversing the null and alternative hypotheses. Suppose the null hypothesis states "the effect of treatment X exceeds standard therapy by 30 % or more" and the alternative states "the effect of treatment X is less than a 30 % improvement." If results cause us to reject the null, treatment X will not be developed further and we know exactly the chance that we have discarded a useful therapy: the chance that treatment X exceeded our threshold but was discarded is exactly the α-level of our hypothesis test. In contrast, if we do not reject the null hypothesis, we continue to develop treatment X and are unsure of the true probability that it actually performs below our intended threshold.

These properties may not always be the wisest choice. Recall that strong evidence against the null is required to reject, and strong evidence is unlikely to mislead. Weak evidence is more likely to mislead, and unlikely to reject the null [15]. When the null hypothesis is one of efficacy/improvement, weak evidence can contribute to further development and does not illuminate the chance of carrying forward a loser. Strong evidence (rejection of the null), which typically is desirable, actually supports nonefficacy or nonsuperiority. Why do we want to generate strong evidence that a new therapy is inferior to our benchmark? Broadly speaking, it may be inferentially and ethically more appropriate for us to generate strong evidence that a new treatment is superior rather than inferior.

Gauging type I and II errors to account for the cost of false positives or false negatives can be easily implemented in conventional designs [16]. When comparing treatments in phase III studies, one judges whether a new treatment is superior to a standard control. In this case, generally the control is of known efficacy and should not be superseded without strong evidence of the superiority of its competitor. Accordingly, the chance of a false positive is often set to 5 % ($\alpha = 0.05$ = type I error) and the chance of a false negative to a larger number, say 20 % or less ($\beta = 0.2 = 1$ - power = type II error), However, in phase II (or in "futility") trials, the relative costs of false-positive and false-negative conclusions may be reversed. On the one hand, we do not want to inflict an ineffective treatment on more patients than is absolutely necessary, but we do not want to reject a potentially effective treatment. The cost of a false positive in phase II trials is that of repeated studies, which will eventually demonstrate the lack of efficacy. However, the cost of a false negative is that a useful treatment is completely discarded. Accordingly, if we were to set the chance of a false positive to, say, $\alpha = 0.15$ and the chance of a false negative to $\beta = 0.05$, we would

be protecting the new treatment from an untimely demise. For example, suppose we test the null hypothesis that a treatment has a 15 % response frequency versus the alternative of a 35 % response frequency. After evaluating 35 patients, if ten or more responses are observed, we may reject the null hypothesis. The trial has an actual 2.9 % type I error and an 83 % power (17 % type II error). If we change the cut point from ten to eight responses, using the same sample size, the trial has a 14 % type I error and a 95.8 % power (4.2 % type II error). The sum of error probabilities is about the same for both arrangements, but the second design protects an effective treatment from premature rejection. Moreover, by increasing power from 83 to 95.8 %, we have decreased the false-negative error probability by a factor of four.

Reversal of the null and alternative hypotheses is not required either to construct an optimistic pipeline, demonstrate futility/nonsuperiority, or to align type I and II errors with their consequences. For example, we could conventionally construct the null hypothesis to be "the effect of treatment X is less than a 30 % improvement over standard," and the alternative hypothesis to be "the effect of treatment X exceeds standard therapy by 30 % or more." This is a clean nonsuperiority hypothesis. If we want an optimistic pipeline that favors moving therapies forward, we might set the type I error at 10 % or even 20 %. It could be a very serious error to miss a treatment that actually represented a 50 % improvement, in which case we might want the type II error for that alternative to be as small as say 5 %. Then weak evidence makes it difficult to advance a treatment, and strong evidence is likely to advance a treatment when it is actually good. Thus the properties that we admire in a "futility" design can be achieved using conventional ideas.

Acknowledgements This paper is supported in part by the National Center for Research Resources, Grant UL1RR033176, and is now at the National Center for Advancing Translational Sciences, Grant UL1TR000124 (AR and SP), Grant 5P01CA098912-05 (AR), Grant P01 DK046763 (AR) and Grant 2R01CA108646-07A1 (AR) . The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

1. Palesch, Y.Y., Tilley B.C.: An efficient multi-stage, single-arm Phase II futility design for ALS. Amyotroph. Later. Scler. Other Mot. Neuron Disord. **5**, 55–56 (2004).
2. Elm, J.J., Goetz, C.G., Ravina, B., Shannon, K., Wooten, G.F., Tanner, C.M., et al.: A responsive outcome for Parkinson's disease neuroprotection futility studies. Ann. Neurol. **57**, 197–203 (2005).
3. Levin, B.: The utility of futility. Stroke **36**, 2331 (2005).
4. Palesch, Y.Y., Tilley, B.C., Sackett, D.L., Johnston, K.C., Woolson, R.: Applying a phase II futility study design to therapeutic stroke trials. Stroke **36**, 2410 (2005).
5. Levin, B.: Selection and futility designs. In: Ravina, B., Cummings, J., McDermott, M., Poole, R.M. (eds.) Clinical Trials in Neurology: Design, Conduct, Analysis, pp. 78–90. Cambridge University Press, Cambridge (2012).
6. Wellek, S.: Testing Statistical Hypotheses of Equivalence and Noninferiority, 2nd edn. Chapman & Hall/CRC, Boca Raton. (2010).

7. Neyman, J.: On the problem of the most efficient tests of statistical hypotheses. Philos. Trans. Royal Soc. Lond. Ser. A Contain. Pap. Math. Phys. Charact. **231**, 289–337 (1933).
8. Gehan, E.A.: Determination of number of patients required in a preliminary and a follow-up trial of a new chemotherapeutic agent. J. Chron. Dis. **13**, 346 (1961).
9. Simon, R.: Optimal 2-stage designs for phase-II clinical-trials. Controll. Clin. Tr. **10**, 1–10 (1989).
10. Aberson, C.: Interpreting null results: improving presentation and conclusions with confidence intervals. J. Artic. Support Null Hypothesis **1**, 36 (2002).
11. Altman, D.G., Bland, J.M.: Statistics notes: absence of evidence is not evidence of absence. BMJ **311**, 485 (1995).
12. Kieburtz, K., Tilley, B., Ravina, B., Galpern, W., Shannon, K., Tanner, C., et al.: A pilot clinical trial of creatine and minocycline in early Parkinson disease: 18-month results. Clin. Neuropharmacol. **31**, 141 (2008).
13. Dunnett, C.W., Gent, M.: Significance testing to establish equivalence between treatments, with special reference to data in form of 2 x 2 tables. Biometrics **33**, 593–602 (1977).
14. Blackwelder, W.C.: Proving the null hypothesis in clinical-trials. Controll. Clin. Tr. **3**, 345 (1982).
15. Royall, R.M.: Statistical Evidence : A Likelihood Paradigm. Chapman & Hall, Boca Raton (1997).
16. Rogatko, A., Litwin, S.: Phase II studies: which is worse, false positive or false negative? J. Natl. Cancer Inst. **88**, 461 (1996).

Chapter 5
Cognitive-Constructivism, Quine, Dogmas of Empiricism, and Münchhausen's Trilemma

Julio Michael Stern

Abstract The Bayesian research group at University of São Paulo has been exploring a specific version of cognitive constructivism (Cog-Con) that has, among its most salient features, a distinctive objective character. Cog-Con is supported by a specially designed measure of statistical significance, namely, $\mathrm{ev}(H \mid X)$—the Bayesian epistemic value of sharp hypotheses H, given the observed data X. This article explores possible parallels or contrasts between Cog-Con and the epistemological framework developed by the philosopher Willard van Orman Quine.

A gente vive repetido, o repetido... Digo: O real não está na saída nem na chegada: Ele se dispõem para a gente é no meio da travessia. ...Existe é homem humano. Travessia.
João Guimarães Rosa (1908–1967); Grande Sertão, Veredas.

We live repeating the repeated... I say: What's real can be found neither at departure nor upon arrival: It only becomes available during the journey. ...What exists is the living man; Crossing-over.

Warum ist Wahrheit fern und weit? Birgt sich hinab in tiefste Gründe? Niemand versteht zur rechten Zeit! Wenn man zur rechten Zeit verstünde, So wäre Wahrheit nah und breit, Und wäre lieblich und gelinde.
Goethe (1749–1832); Westöstlicher Diwan. As quoted in Rosenzweig (1921).

Why is truth so remote and far away? Down at the deepest bottom hold astray? Nobody understands its proper time! If, however, in its due time understood; Then truth would be close and sublime; And would be graceful and good.

...the idea of circularity: That's where everything begins.
Heinz von Foerster (1911–2002); Understanding Systems.

J. M. Stern (✉)
Institute of Mathematics and Statistics (IME-USP), University of São Paulo, São Paulo, Brazil
e-mail: jstern@ime.usp.br

A. Polpo et al. (eds.), *Interdisciplinary Bayesian Statistics*,
Springer Proceedings in Mathematics & Statistics 118, DOI 10.1007/978-3-319-12454-4_5

5.1 Introduction

For the last 18 years, the Bayesian research group of the Institute of mathematics and Statistics of the University of São Paulo (IME-USP) has been exploring a specific version of cognitive constructivism (Cog-Con) that has among its most salient features a distinctive objective character and the support of specially designed tools of Bayesian Statistics.

In previous presentations about the Cog-Con epistemological framework, for audiences with interests spanning Logic and Epistemology, foundations of Bayesian Statistics, and foundations of science, we were asked several questions concerning possible parallels or contrasts between Cog-Con and the epistemological framework developed by the philosopher Willard van Orman Quine. This chapter begins to explore this topic.

Section 5.2 gives a succinct overview of Cog-Con—the cognitive constructivist epistemological framework used in this chapter. The following sections explore similarities and differences between Cog-Con and Quine's epistemological frameworks. Section 5.3 analyzes the *Two Dogmas of Empiricism* denounced by Quine, as well as the *Third Dogma* proposed by Davidson, making some parallels between Cog-Con and Quine's epistemological frameworks. Section 5.4 contrasts the two frameworks on their respective strategies to anchor a scientific theory to reality. Section 5.5 uses the Münchhausen Trilemma to continue the comparative analysis of the two frameworks, while Sect. 5.6 compares them on the role played by Ontology and Metaphysics. Section 5.7 brings our final remarks.

5.2 Objects as Tokens for Eigen-Solutions

Eigen-solution is the key concept of the objective version of Cog-Con used in this chapter. Eigen-solutions emerge as invariant entities (i.e., as operational eigen-equilibrium-invariant-fixed-...-solutions-states-behaviors-points) for an autonomous system interacting with its environment. The fundamental insight of this epistemological framework is summarized by the celebrated aphorism of Heinz von Foerster—*objects are tokens for eigen-solutions*. In other words, objects, and the names we use to call them, stand for and point at such invariant entities (see [44] and [30], pp. 127, 145).

In Cog-Con, the concept of autonomy can be further specified using the idea of concrete or abstract *autopoietic system*—a system organized as a network of processes of production of components that, through their interactions and transformations, recursively regenerate the same production network and its constituent components (see [17], p. 78).

As possible examples of an autonomous interacting system in its respective environment, we may consider a bacteria in its culture medium, a human individual living in his or her social environment, or a scientific discipline and its field of study,

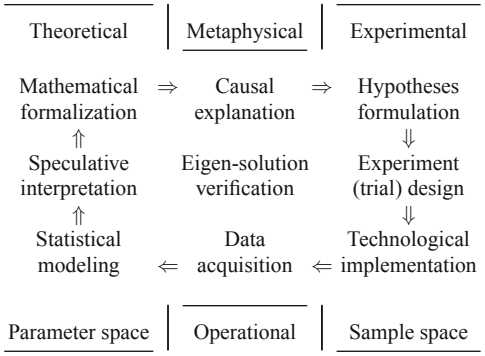

Fig. 5.1 Scientific production diagram

i.e., the discipline's standard language, theories, empirical means and methods, experimental tools and equipment, etc., that have been developed for the continuous research of its area of expertise.

Figure 5.1 depicts a generic diagram of systemic interaction for a scientific discipline. The right side of the square depicts activities related to empirical trial or experiment design and implementation. The lower side depicts operations related to data observation, generation, acquisition, and analysis. The left side relates to theoretical modeling and formalization. Finally, the upper side relates to metaphysics; the conception and application of causal mechanisms, explaining why things behave the way they do. In statistical models, observable quantities of interest are represented by (stochastic) variables in the sample space, while latent (unobservable) quantities of interest are represented by variables in the parameter space.

Eigen-solutions can be tagged or labeled by words, and these words can be articulated in language. Of course, the articulation rules defined for a given language, its grammar and semantics, only make the language useful if they somehow correspond to the composition rules for the objects the words stand for. Ontologies are carefully controlled languages used in the practice of science. They are developed as tools for procedure specification, thinking, and communication. According to the constructivist perspective, keywords in scientific ontologies are labels for eigen-solutions. This is the constructivist approach to the classic problem of external symbol grounding and alignment of scientific ontologies, as discussed in [38].

Eigen-solutions are characterized by four essential attributes, namely, *precision, stability, separability, and composability*. Depending on context, precision is better described as sharpness, lower dimensionality, discreetness, singularity, etc. In several well-known examples in exact sciences, these four essential attributes lead to the concept of basis, for example: basis of a finite vector space, like in linear algebra; basis of a Hilbert space, like in Fourier or Wavelet analysis, etc. Nevertheless, the concept of eigen-solution and its four essential attributes is so important in the CogCon framework, that it is used as a fundamental metaphor in far more general, and not necessarily formalized, contexts.

In the context of statistics, the essential attributes of eigen-solutions makes them amenable for representation as statements of a very special form, namely, sharp or precise hypotheses. A hypothesis $H = \{\theta \in \Theta \mid g(\theta) \leq 0 \wedge h(\theta) = 0\}$, states that the (continuous vector) parameter θ of a statistical model lays in a region of the parameter space specified by (vector) inequalities and equality constraints. Sharp hypotheses must have at least one equality constraint. The treatment of sharp statistical hypotheses, in turn, is far from trivial, demanding the employment of suitable mathematical techniques and requiring the use of coherent methods of inference.

For the past 18 years, the Bayesian Statistics research group of IME-USP, at the University of São Paulo, has been developing a significance measure known as the e-value—the epistemic value of a sharp or precise statistical hypotheses H, given the observational data X or, in short, $\text{ev}(H \mid X)$; see [16, 19, 20]. At the same time, we have been exploring the strong algebraic properties for composition of e-values and the characterization of these rules for combination and propagation of truth functions and truth values as abstract belief calculi or logical inferential systems, see [6, 31, 40]. These algebraic rules or logical properties further reflect essential compositional properties of the underlying (or represented) eigen-solutions. Moreover, we have been developing Cog-Con as a suitable epistemological counterpart for this novel statistical theory, and vice-versa, see Stern [32–38].

5.3 Parallels Between Cog-Con and Quine's Epistemological Frameworks

In the opening paragraph of his famous paper of 1951, Willard van Orman Quine denounces *Two Dogmas of Empiricism*, as stated in the first of the following quotations. In 1974, Donald Davidson denounced what he saw as a third dogma, as stated in the second quotation. However, as stated in the third quotation after carefully constraining the context and scope of his argument, Quine accepts the premise of scheme-content dualism as necessary, see further comments on Sect. 5.6.

> Modern empiricism has been conditioned in large part by two dogmas. One is a belief in some fundamental cleavage between truths which are *analytic*, or grounded in meanings independently of matters of fact and truths which are *synthetic*, or grounded in fact. The other dogma is *reductionism*: the belief that each meaningful statement is equivalent to some logical construct upon terms which refer to immediate experience. Both dogmas, I shall argue, are ill founded. One effect of abandoning them is, as we shall see, a blurring of the supposed boundary between speculative metaphysics and natural science. Another effect is a shift toward pragmatism. [22, p. 20]
>
> I want to urge that this... dualism of scheme and content, of organizing system and something waiting to be organized, cannot be made intelligible and defensible. It is itself a dogma of empiricism, the third dogma. The third, and perhaps the last, for if we give it up it is not clear that there is anything distinctive left to call empiricism. [9, p. 11]
>
> If empiricism is construed as a theory of truth, then what Davidson imputes to it as a third dogma is rightly imputed and rightly renounced... As a theory of evidence, however, empiricism remains with us, minus indeed the two old dogmas. The third purported dogma, understood now in relation not to truth but to warranted belief, remains intact. It has both

> a descriptive and a normative aspect, and in neither do I think of it as a dogma. It is what makes scientific method partly empirical rather than solely a quest for internal coherence. [25, p. 39]

Closely related to rejecting the first dogma, analytic–synthetic dualism, is Quine's idea of *naturalism*, namely,

> ... the recognition that it is within science itself, and not in some prior philosophy, that reality is properly to be identified and described. [25, p. 21]

Closely related to rejecting the second dogma, reductionism, is Quine's idea of *confirmational holism*, stated as follows:

> The totality of our so-called knowledge or beliefs... is a man-made fabric which impinges on experience only along the edges. Or, to change the figure, total science is like a field of force whose boundary conditions are experience. A conflict with experience at the periphery occasions readjustments in the interior of the field. But the total field is so undetermined by its boundary conditions, experience, that there is much latitude of choice as to what statements to re-evaluate in the light of any single contrary experience. No particular experiences are linked with any particular statements in the interior of the field, except indirectly through considerations of equilibrium affecting the field as a whole. [22, pp. 42–43]

In summary, Quine's epistemological framework involves the following five interconnected premises.

1. Rejection of *analytic–synthetic dualism*
2. Rejection of *reductionism*
3. Acceptance of *scheme–content dualism*
4. Acceptance of *naturalism*
5. Acceptance of *confirmational holism*

We believe that these five premises account for most of the similarities often perceived between Cog-Con and Quine's epistemological frameworks, and with good reason indeed.

Let us start with a comparative examination of Quine's third premise in relation to the Cog-Con framework. Cog-Con departing point is the existence of a concrete or abstract autonomous system interacting with its environment. Of course, a clear distinction between the system itself, including its characteristic features or internal organization, versus an external environment and its inherent form, i.e., the structural constraints of the system's domain, range or scope of interaction, is a necessary condition for analyzing such an interaction process. Hence, we immediately correlate Cog-Con's *system–environment* distinction with *scheme–content* dualism.

Let us proceed asserting that, in the Cog-Con framework, Mathematics is regarded as a *quasi-empirical science*, using a terminology developed by the philosopher Imre Lakatos, see [14, 15, 42]. For a detailed analysis of this concept, in the context of the Cog-Con framework, see [36]. For the purposes of this chapter, it suffices to say that mathematics as a quasi-empirical science is an idea consonant with Quine's premises of rejecting analytic–synthetic dualism and accepting naturalism.

Finally, let us turn our attention to Fig. 5.1, depicting the *Scientific production diagram*. From this diagram, we see that the production of eigen-solutions in the

practice of science depends not only on the pertinent scientific theory as a whole, but also on the necessary technological means, on appropriate methods of experiment design and data analysis, etc. Clearly, these ideas are in accord with Quine's premises of rejecting reductionism and accepting confirmational holism. Hence, concerning the five premises examined in this section, we find Cog-Con and Quine's epistemological frameworks in comprehensive agreement.

Before ending this section, we should emphasize that, in both of the epistemological frameworks at hand, rejection of reductionism and acceptance of confirmational holism does not conflict with the concept of having testable statements somehow inserted in the fabric of scientific knowledge. However, there are important differences on the form, role, and position of such testable statements according to each of the two epistemological frameworks, as examined in the following sections.

5.4 Contrasting Strategies Against Skepticism

The holistic perspective, preeminent either in Quine's or in the the Cog-Con frameworks, brings the danger of vicious circularity that, in turn, easily leads to skepticism. Hence, in both of these frameworks, it is important to defeat skepticism. Specifically, it is important to carefully explain how to anchor the theory and methods of a scientific discipline into *reality*. Contrasting the two epistemological frameworks' strategies for defeating skepticism, we find the main differences between the two approaches. Let us start analyzing the short but very dense *Reply to Stroud* in [26].

Stroud [26, p. 457] raises some questions concerning *the possibility that the world is completely different in general from the way our sensory impacts and our internal makeup lead us to think of it.*

Quine [26, pp. 473–474] gives his reply in three basic steps, namely:

- First, he assumes *an inclusive theory of the world, regimented in the framework of predicate logic*;
- Next, using the standard tools of predicate logic, he *view[s] this [Stroud's] possibility in the perspective of proxy functions and displaced ontologies*;
- Finally, after appropriate manipulation and reinterpretation of predicative statements, he comes to the conclusion that:

> The structure of our theory of the world will remain unchanged. Even its links to observational evidence will remain undisturbed, for the observation sentences are conditioned holophrastically to stimulations, irrespective of any reshuffling of objective reference.
> *Once we take observation sentences holophrastically, however, reference and objects generally go theoretical. The indifference or inscrutability of ontology comes to apply across the board.*
> *The relations of language to stimulations are unaffected by any shift of our ontology through proxy functions. Evidently then such shifts are indifferent to use, to meaning. The platitude that meaning determines reference goes by the board, along with the absoluteness of reference. The objects, or values of variables, serve only as nodes in a verbal network whose termini a quis et ad quos are observations, stimulatory occasions.*

Quine's arguments in his answer to Stroud depend critically on two additional (and explicit) premises:

6-Quine. Acceptance of a *predicate logic structure*, i.e., assuming that predicate logic (or similar belief calculi) gives the symbolic structure of choice for the *regimentation* of science.

7-Quine. Acceptance of the *finite regress* premise, that is, to assume the availability of terminal nodes when parsing down well-formulated truth-bearing statements within the scope a well-regimented theory. In Quine's framework, these terminal or bottom nodes come in the form of observation sentences, conditioned holophrastically to stimulations (In logic and computer science, the standard representation of this parsing process is a tree structure having its starting node or root at the top and its terminal nodes or leaves at the bottom).

In contrast, in the Cog-Con framework, these two additional premises are replaced by alternatives of a very different nature, namely,

6-Cog-Con. Acceptance of a *statistical model structure*, i.e., assuming that probability and mathematical statistics (or alternative models based on similar belief calculi), gives a symbolic structure of choice for testing operations (verification / falsification) in empirical science.

7-Cog-Con. Acceptance of *objects as tokens for eigen-solutions*—invariants or fixed-points in essentially cyclic and recursively defined processes.

Moreover, the premises of *objects as tokens for eigen-solutions* and *statistical model structure* are linked by assuming the availability of *sharp statistical hypotheses* as check points in well-posed scientific disciplines. Successful (statistical) testing of such sharp hypotheses implies an *evaluation of objectivity* (or verification of *existence*), for a corresponding set of objects in the pertinent scientific ontology.

Statistical models make use of real-vector variables for representing observed quantities or latent parameters. Moreover, in such models, learning (ex. Bayesian updates) and (stochastic) convergence properties are appropriately described by continuous mathematics, in sharp contrast with the discrete character of propositional logic.

The contrasting assumptions examined so far in Cog-Con and Quine's epistemological frameworks are also related to the perceived position, from a central to peripheral range, of the testable statements in scientific knowledge. For Quine, these contact-points with reality have a peripheral position, as clearly stated in his analogy of scientific knowledge to *a man-made fabric which impinges on experience only along the edges* [22, p. 42]. In contrast, the Cog-Con framework relates its testable hypotheses to valid of eigen-solutions, entities that, by their very nature, are perceived as deeply (and recursively) embedded in systemic activity. We believe that this topic can be further elucidated studying the Münchhausen Trilemma.

5.5 Münchhausen Trilemma

The difference between adopting the premise of *finite regress*, in Quine's approach to epistemology, and adopting the premise of *objects as tokens for eigen-solutions*, in the Cog-Con epistemological framework, may be seen as a consequence of choosing different horns of the celebrated Münchhausen trilemma, as it was formulated by [2, p. 18]:

> If one demands a justification for everything, one must also demand a justification for the knowledge to which one has referred back the views initially requiring foundation. This leads to a situation with three alternatives, all of which appear unacceptable: in other words, to a trilemma which, in view of the analogy existing between our problem and the one which the celebrated and mendacious baron once had to solve, I should like to call the Münchhausen trilemma. For, obviously, one must choose here between
> (a) an infinite regress, which seems to arise from the necessity to go further and further back in the search for foundations, and which, since it is in practice impossible, affords no secure basis;
> (b) a logical circle in the deduction, which arises because, in the process of justification, statements are used which were characterized before as in need of foundations, so that they can provide no secure basis; and, finally,
> (c) the breaking-off of the process at a particular point, which, admittedly, can always be done in principle, but involves an arbitrary suspension of the principle of sufficient justification.

Of course, each epistemological framework must deal with the perils of its favorite (or least feared) horn of Münchhausen trilemma:

- Quine must show how to achieve a finite break-off when parsing down well-posed truth-bearing statements formulated in the scope of a theory regimented according to his chosen symbolic structure, namely, predicate logic.
- In our opinion, Quine should also accomplish the far more difficult task of convincing his potential clientele that his choices 6 and 7, namely, his premises of predicate logic structure and finite regress, are the most appropriate for the (naturalized) epistemological task of verifying and evaluating scientific theories. Moreover, from the same naturalistic perspective, it would be very helpful if the same choices were also the most suitable for use in active duty by scientists trying to test or validate their empirical models and working hypothesis.

Of course, the proponents of the Cog-Con framework must also justify their choices 6 and 7. In particular it is vital to:

- Justify abandoning predicate logic as the single belief calculus used for epistemological reasoning, i.e., the only set of rules used for truth analysis and propagation in the evaluation of scientific models, and the introduction of continuous logics or belief calculi such as probability or possibility theory. The introduction of new belief calculi requires careful argumentation, since predicate logic and its variants have a long-standing tradition of playing all alone the aforementioned role in many frameworks for epistemology and philosophy of science.

- Show how to tame the circularity looming at horn (b) of Münchhausen trilemma into a virtuous form, for vicious forms of circularity threaten to poison the health of any epistemological framework, and hand over victory to skepticism.

Quine himself, and many of his followers, have worked extensively in the aforementioned tasks. Although Cog-Con in general, and the objective version of it used in this chapter in particular, are much younger—presumably only at the beginning of their development; several already published theoretical results and practical applications show how these tasks can be successfully accomplished.

Ultimately, it will be up to the user to choose the framework that he or she considers the most natural, intuitive, and well-adapted to its field of interest. In the case of Cog-Con, the available or required mathematical tools may have an influence in this choice. On one hand, statistics is (or should be) a well-known tool of the trade for the working scientist. On the other hand, the abstract representation of virtuous forms of circularity is a far less common task.

Nonetheless, some fields, most specially computer science, have developed powerful formalisms for the abstract representation of essentially circular processes that produce, nevertheless, well-defined objects. Some nonstandard logics and set theories are, in fact, mathematical tools tailor-made for the logical representation of such recursively defined objects. As interesting examples, we could mention Hypersets or Non-Well-Founded Set Theory and its derivatives, see for example ([1], [3], [4, Chap. 4], [5], [10, Chap. 8]).

5.6 Ontology and Metaphysics

Once again, we start noticing striking similarities between Cog-Con and Quines's frameworks at important departing points, followed by significant differences in their evolution. The common departing point we have in mind is the relation between ontology (understood as studies on what there is) and language, as stated in [22, p. 16]:

> Our acceptance of an ontology is, I think, similar in principle to our acceptance of a scientific theory, say a system of physics: we adopt, at least insofar as we are reasonable, the simplest conceptual scheme into which the disordered fragments of raw experience can be fitted and arranged... To whatever extent the adoption of any system of scientific theory may be said to be a matter of language, the same-but no more-may be said of the adoption of an ontology.

As seen in Sect. 5.2, from the Cog-Con perspective, ontologies are carefully controlled languages used in the practice of science, developed as tools for procedural description and specification, theoretical reasoning, and communication. A scientific ontology includes a (formal) definition of the collection of objects of knowledge of a given scientific discipline and its organization, i.e., the (semantic) relations that exist between these objects. Hence, so far, Cog-Con and Quine's frameworks seem to be in good agreement concerning issues of ontology and language.

However, as previously seen, Quine's epistemological framework leads him to ascertain the *indifference or inscrutability of ontology*. This indifference, in turn, takes him to a strict observation/ prediction/ verificationalist position, that gives

little room for more important roles to be played by ontology or metaphysics. In [27, p. 31], see also [11, pp. 137–138], he states:

> Reference and ontology recede thus to the status of mere auxiliaries. True sentences, observational and theoretical, are the alpha and omega of the scientific enterprise.

In [24, p. 80], he goes to the extreme of saying of the Vienna Circle that they - *espoused a verification theory of meaning but did not take it seriously enough*, see also [11, p. 226]. Ironically, this takes Quine to embrace an extreme form of positivism concerning the role of metaphysics, in perfect syntony with the ideas of Auguste Comte, founder of the Positivist movement, as seen in the following quotes.

> Now it is an ironical but familiar fact that though the business of science is describable in unscientific language as the discovery of cause, the notion of cause itself has no firm place in science. The disappearance of causal terminology from the jargon of one branch of science and another has seemed to mark the progress in understanding of the branches concerned. [23, p. 242], as quoted in [12, p. 358].
>
> The first characteristic of the Positive Philosophy is that it regards all phenomena as subjected to invariable natural Laws. Our business is - seeing how vain is any research into what are called Causes, whether first or final, - to pursue an accurate discovery of these Laws, with a view to reducing them to the smallest possible number. By speculating upon causes, we could solve no difficulty about origin and purpose. Our real business is to analyze accurately the circumstances of phenomena, and to connect them by the natural relations of succession and resemblance. [8, p. 27], quoted in [11, p. 220].

In contrast, from the Cog-Con epistemological framework's perspective, ontology and metaphysics are at the center stage of the scientific scenario, and play a preeminent role in good epistemological analysis.

From the Cog-Con perspective, Metaphysics concerns causal explanations telling why things are the way they do. These are the narratives and metaphors, often intertwined with abstracts symbolic statements, we use to build our understanding, to gain insight or intuition about objects in our world and the way they work. Hence, good scientific ontologies, even if informally defined, are an indispensable tool for metaphysical and theoretical reasoning, and both ontology and metaphysics are necessary supports for the conception, development, implementation, and consistent application of experimental procedures and operational means and methods. As such, ontology and metaphysics are of paramount importance in scientific life, taking part in all steps of the production diagram depicted at Fig. 5.1.

From the Cog-Con perspective, the clarity of understanding, insight, or intuition provided to its end users by a good ontology and its associated metaphysical (causal) explanations must be carefully and thoroughly considered. Such considerations may even justify the nontrivial concomitant use of alternative but formally equivalent ontologies and theoretical frameworks. Indeed, let us consider the case of two or more of such formally equivalent frameworks like, for example, Newtonian, Lagrangean, and Hamiltonian classical mechanics, or the standard and Feynman's path integral formulation of quantum mechanics. From the Cog-Con perspective, even if two such frameworks are formally equivalent, their different ontological commitments and characteristic causal relations are still very relevant in the active practice of science and, therefore, from a Cog-Con perspective, equally important for epistemology.

For a detailed discussion of the Cog-Con perspective, including similar ideas of Max Born, and further contrasts with Positivism's strict observation/ prediction/ verificationalist position, see ([32, Chap. 5], [7, 36, 37, 40]. Max Planck [21, pp. 171–172] presents a comparable position:

> Positivism lacks the driving force for serving as a leader on this road. True, it is able to eliminate obstacles, but it cannot turn them into a productive factors. For its activity is essentially critical, its glace is directed backward. But progress, advancement requires new associations of ideas and new queries, not based on the results of measurements alone, but going beyond them, and toward such things the fundamental attitude of Positivism is one of aloofness.

5.7 Future Research Final Remarks

Self-Identities at Neurath's Ship

As a wordy epistemological framework, Cog-Con should be useful for statical analysis of a scientific discipline, i.e., it should be a helpful device for making sense of the daily activities of the system as it operates in a given status quo. However, the Cog-Con framework also aims at dynamical analyses, aspiring to provide useful tools for understanding science in its development and evolution. We believe that Cog-Con can offer valuable insights at the never-ending processes of reconstruction while staying afloat that characterizes Neurath's Ship. In particular, we believe that the Cog-Con approach can provide intuitive guidelines for engineering and using computational tools for *ontology alignment*, see [38]. Synchronic and diachronic ontology alignments, in turn, can help us to understand the underlying continuities of systems in their development and evolution, i.e., can help us to understand the safe-guarding of self-identities during the brave voyages of Neurath's Ship.

Handling Sticky Things

> Truth, meaning and belief are sticky concepts. They stick together.... Meaning and belief ... can be separated, like Siamese twins, only by artificial means... But it is the remaining pair, truth and belief, that seems to me to have got unobservedly stuck.

These are among the opening statements of *On the very Idea of a Third Dogma*, Quine [25, p. 38]. The objective of the present article was to make a high contrast and minimalist comparison between Cog-Con and Quine's epistemological frameworks. Future research should enrich this bare-bone and schematic comparison with finer points and higher resolution details. Contrasting appropriate concepts of truth, meaning, and belief in Cog-Con, Quine, and also Davidson's epistemological frameworks should be part of this effort. Sticky as they are, a comparative study of these three concepts will eventually involve a related triad of sticky things, namely, the concepts of objectivity, subjectivity and inter-subjectivity. Moreover, this line of research

requires discussion of the roles played in naturalized epistemologies by Ontology and Metaphysics in greater depth than we could afford at the present chapter.

Handling Uncertainties in Circular Ontologies

A self-defined mission of CODATA, the International Committee on Data for Science and Technology, is *to periodically provide the scientific and technological communities with a self-consistent set of internationally recommended values of the basic constants and conversion factors of physics and chemistry based on all of the relevant data available at a given point in time*. For more on the importance and meaning of universal constants, these *immutable building blocks blocks of the edifice of theoretical physics*, see [21, pp. 170–172].

Many international laboratories are involved in this on-going effort to obtain ever more precise values for universal constants, using a variety of alternative experimental methods. However, each value obtained for one of these fundamental constants is a (nonlinear) function of the value of several others constants. Such a project creates a large database where entities are circularly defined, a situation that can be handled using methods based on Aczel's hypersets and its derivatives. Moreover, the reliability of each one of these experiments, the quality and conditions of its implementation, and a variety of other influencing factors must also be accounted for. This creates a second-order effect of circular propagation of uncertainties.

A simple approach to quantify uncertainties about universal constants could be based on the (convergent, circular) propagation of variances and covariances, the second-order statistical moments. Borges and Stern [6] and Stern [31] suggest possible ways to handle empirical hypotheses and to treat more complex relations.

We are interested in extensions of already existing computational tools, like Pakkan and Akman [18] and Iordanov [13], aiming at the easy handling of sets of circular uncertainties. Such computational tools could, in turn, be used in the effort of evaluating scientific ontologies and constructing or justifying synchronic and diachronic ontology alignments, see [38].

Acknowledgments The author is grateful for the support received from IME-USP, the Institute of Mathematics and Statistics of the University of São Paulo; FAPESP, the São Paulo Research Foundation (grant CEPID 2013/07375-0); and CNPq, the Brazilian National Council for Scientific and Technological Development (grant PQ-306318-2008-3). The author is also grateful for helpful discussions with several of his professional colleagues and for advice received from anonymous referees. Finally, the author is grateful for the organizers and the constructive criticism received from the following discussion fora: EBEB-2014, the XII Brazilian Meeting on Bayesian Statistics, held on March 10-14 at Atibaia, Brazil; EBL-2014, the XVII Brazilian Logic Conference, held on April 7-11 at LNCC, the National Laboratory for Scientific Computing, Petrópolis, Brazil; ALFAn-2014, the III Latin American Analytic Philosophy Conference, held on May 27–31 at Fortaleza, Brazil; and MaxEnt-2014, the XXXIV International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, held on September 21–26, Amboise, France.

References

1. Aczel, P.: Non-Well-Founded Sets. CSLI. Stanford University Press, Stanford (1988)
2. Albert, H.: Treatise on Critical Reason. Princeton University Press, Princeton (1985) (Latest German ed: *Traktat über Kritische Vernunft*, 1991. Tübingen: Mohr)
3. Akman, V., Pakkan, M.: Nonstandard set theories and information management. J. Intell. Inf Syst. **6**(1), 5–31 (1996)
4. Barwise, K.J., Etchemendy, J.: The Liar: An Essay on Truth and Circularity. Oxford University Press, Oxford (1987)
5. Barwise, K.J., Moss, L.: Vicious Circles. On the Mathematics of Non-Wellfounded Phenomena. CSLI, Stanford (1996)
6. Borges, W., Stern, J.M.: The rules of logic composition for the Bayesian epistemic e-values. Log. J. IGPL **15**(5–6), 401–420 (2007)
7. Born, M.: Physics in My Generation. Pergamon, London (1956)
8. Comte, A.: Positive Philosophy. Calvin Blanchard, New York (1858)
9. Davidson, D.: On the very idea of a conceptual scheme. Proc. Addresses Am. Philos. Assoc. **47**, 5–20 (1974)
10. Devlin, K. The Joy of Sets. Springer, New York (1994)
11. Gibson, R. Jr.: The Cambridge Companion to Quine. Cambridge University Press, Cambridge. (2004) (Includes: Chap. 5, by P. Hylton - Quine on Reference and Ontology, pp. 115–150, and Chap. 9, by D. Isaacson - Quine and Logical Positivism, pp. 214–169)
12. Hylton, P.: Quine: Arguments of the Philosophers. Routledge, New York (2010)
13. Iordanov, B.: HyperGraphDB: a generalized graph database. Lect. Notes Comput. Sci. **6185**, 25–36 (2010)
14. Lakatos, I.: Proofs and Refutations: The Logic of Mathematical Discovery. J. Worall, E. Zahar (eds). Cambridge University Press, Cambridge (1976)
15. Lakatos, I.: Philosophical Papers, V.1 - The Methodology of Scientific Research Programmes. V.2. - Mathematics, Science and Epistemology. Cambridge University Press, Cambridge (1978)
16. Madruga, M.R., Esteves, L., Wechsler, S.: On the Bayesianity of Pereira–Stern Tests. Test **10**, 291–299 (2001)
17. Maturana, H.R., Varela, F.J.: Autopoiesis and Cognition. The Realization of the Living. Reidel, Dordrecht (1980)
18. Pakkan, M., Akman, V.: Hypersolver: a graphical tool for commonsense set theory. Inf. Sci. **85**(1–3), 43–61 (1995)
19. Pereira, C.A.B., Stern, J.M.: Evidence and credibility: full Bayesian significance test for precise hypotheses. Entropy J. **1**, 69–80 (1999)
20. Pereira, C.A.B., Wechsler, S., Stern, J.M.: Can a significance test be genuinely Bayesian? Bayesian Anal. **3**(1), 79–100 (2008)
21. Planck, M.: Scientific Autobiography and Other Papers. Williams and Norgate, London (1950)
22. Quine, W.v.O.: From a Logical Point of View. Harvard University Press, Cambridge (1953) (Includes: Chap. 1 - On What There Is, pp. 1–19, and Chap. 2 - Two Dogmas of Empiricism, pp. 20–46)
23. Quine, W.v.O.: Ways of Paradox. Harvard University Press, Cambridge (1966)
24. Quine, W.v.O.: Ontological Relativity and Other Essays. Columbia University Press, New York (1969) (Includes: Epistemology Naturalized, pp. 69–90)
25. Quine, W.v.O.: Theories and Things. Harvard University Press, Cambridge (1981a) (Includes: Chap. 4 - On the Very Idea of a Third Dogma, pp. 38–42)
26. Quine, W.v.O.: Reply to stroud. Midwest Stud. Philos. **6**(1), 473–476 (1981b)
27. Quine, W.v.O.: Pursuit of Truth. Harvard University Press, Cambridge (1992)
28. Rosa, J.G. : Grande Sertão: Veredas. José Olympio, Rio de Janeiro (1956) (Translated as: The Devil to Play in the Backlands - The Devil in the Street, in the Middle of the Whirlwind. Alfred Knopf, NY (1963))

29. Rosenzweig, F.: *Das Büchlein vom gesunden und kranken Menschenverstand*. Düsseldorf: Joseph Melzer. The same poem is also quoted in F. Rosenzweig (1925). Das Neue Denken - Einige nachträgliche Bemerkungen zum Stern der Erlösung. Der Morgen. **1**(4), 426–451 (1921)
30. Segal, L.: The Dream of Reality. Heinz von Foerster's Constructivism. Springer, New York. (2001)
31. Stern, J.M.: Paraconsistent sensitivity analysis for Bayesian significance tests. Lect. Notes Artif. Intell. **3171**, 134–143 (2004)
32. Stern, J.M.: Cognitive constructivism, Eigen-solutions, and sharp statistical hypotheses. Cybern.& Hum. Knowing **14**(1), 9–36 (2007a)
33. Stern, J.M.: Language and the self-reference paradox. Cybern.& Hum. Knowing **14**(4), 71–92 (2007b)
34. Stern, J.M.: Decoupling, sparsity, randomization, and objective Bayesian inference. Cybern. Hum. Knowing **15**(2), 49–68 (2008a)
35. Stern, J.M.: Cognitive Constructivism and the Epistemic Significance of Sharp Statistical Hypotheses. Tutorial book for MaxEnt 2008 - The 28th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering. July 6–11, Boracéia, Brazil (2008b)
36. Stern, J.M.: Constructive verification, empirical induction, and Falibilist deduction: a threefold contrast. Information **2**, 635–650 (2011a)
37. Stern, J.M.: Symmetry, invariance and ontology in Physics and Statistics. Symmetry **3**(3), 611–635 (2011b)
38. Stern, J.M.: Jacob's Ladder and Scientific Ontologies. Cybern. Hum. Knowing 21(3), 9–43 (2014)
39. Stern, J.M., Nakano, F.: Optimization models for reaction networks: information divergence, quadratic programming and Kirchhoff's laws. Axioms **3**, 109–118 (2014)
40. Stern,[1] J.M., Pereira, C.A.B.: Bayesian epistemic values: focus on surprise, measure probability! Log. J. IGPL **22**(2), 236–254 (2014)
41. Stroud, B.: The significance of naturalized epistemology. Midwest Stud. Philos. **6**(1), 455–472 (1981)
42. Szabó, A.: The Beginnings of Greek Mathematics. Budapest, Akadémiai Kiadó (1978)
43. von Foerster, H.: Understanding Systems. Kluwer, Dordrecht (2002)
44. von Foerster, H.: Understanding Understanding: Essays on Cybernetics and Cognition. Springer, New York (2003)

[1] text

Chapter 6
A Maximum Entropy Approach to Learn Bayesian Networks from Incomplete Data

Giorgio Corani and Cassio P. de Campos

Abstract This chapter addresses the problem of estimating the parameters of a Bayesian network from incomplete data. This is a hard problem, which for computational reasons cannot be effectively tackled by a full Bayesian approach. The work around is to search for the estimate with maximum posterior probability. This is usually done by selecting the highest posterior probability estimate among those found by multiple runs of Expectation-Maximization with distinct starting points. However, many local maxima characterize the posterior probability function, and several of them have similar high probability. We argue that high probability is necessary but not sufficient in order to obtain good estimates. We present an approach based on maximum entropy to address this problem and describe a simple and effective way to implement it. Experiments show that our approach produces significantly better estimates than the most commonly used method.

6.1 Introduction

Bayesian networks (BN) are well-established probabilistic graphical models that can represent joint probability distributions over a large number of random variables in a compact and efficient manner by exploiting their conditional independences, encoded through a directed acyclic graph. Inferring BNs from data sets with missing values is a very challenging problem even if the graph is given [10]. This chapter focuses; on inferring the parameters of a BN with *known graph* from incomplete data samples, under the assumption that missingness satisfies MAR (missing-at-random). The missing data make the log-likelihood function nonconcave and multimodal; the

G. Corani (✉)
Istituto Dalle Molle di studi sull'Intelligenza Artificiale (IDSIA), Scuola universitaria professionale della Svizzera italiana (SUPSI), Università della Svizzera italiana (USI), Manno, Switzerland
e-mail: giorgio@idsia.ch

C. P. de Campos
Dalle Molle Institute for Artificial Intelligence, Manno, Switzerland

Queen's University, Belfast, UK
e-mail: c.decampos@qub.ac.uk; cassio@idsia.ch

A. Polpo et al. (eds.), *Interdisciplinary Bayesian Statistics,*
Springer Proceedings in Mathematics & Statistics 118, DOI 10.1007/978-3-319-12454-4_6

most common approach to estimate the parameters is based on the Expectation-Maximization (EM) algorithm [8, 14]. In this case, EM can be used to search for estimates that maximize the posterior probability of the data (rather than the likelihood [17, Sect. 1.6], as it was originally designed [8]). Maximizing the posterior probability rather than the likelihood is recommended, as it generates BN parameter estimates which are less prone to overfitting [13, 14].

With abuse of notation, in the following, we refer to the posterior probability of the data as the MAP score. Maximizing the posterior probability of the data is by far the most used idea to infer BN parameters, even if it does not offer the same advantages of a full Bayesian estimation. For instance, because it does not integrate over the posterior probability, it cannot average over the uncertainty about the parameter estimates. On the other hand, estimation can be performed by fast algorithms, such as EM, while the computational cost of the full Bayesian approach to infer BN parameters is simply prohibitive, especially in domains with many variables. EM almost always converges to a local maximum of the MAP score, so multiple starts from different initialization points are adopted with the aim of avoiding bad local maxima, and eventually the estimate corresponding to the highest MAP score is selected.

One could expect an improvement in the estimation of the parameters by using an algorithm that always obtains the *global* maximum solution of the MAP score instead of a local one, something that cannot be guaranteed with EM. To check this conjecture, we implement an optimization framework which is ensured to find, at least in small-sized problems, the global maximum score. For large domains, such task is computationally intractable, as the problem is known to be NP-hard. However, we show empirically that the global solver produces worse parameter estimates than EM itself does, despite finding estimates with higher MAP scores. The global maximum of the MAP score seems, thus to be subject to some type of overfitting, highlighting severe limitations in the correlation between MAP score and the quality of the parameter estimates. In turn, this opens a question about whether selecting the estimate with highest MAP score is the best approach. Different EM runs typically achieve very close values of the MAP score, and yet return largely different parameter estimates [13, Chap. 20]. Selecting the parameter estimate which maximizes the MAP score is not a robust choice, since the difference in score among competing estimates can be very thin. We note that approaches such as the Bayesian Information Criterion (BIC) do not constitute a solution to this problem: since all the competing estimates refer to the same graph, the BIC (and other similar approaches) would simply select the estimate with highest MAP score.

In view of such considerations, we propose the following idea to estimate BN parameters: One should select the *least informative estimate, namely the maximum entropy one, among those which have a high MAP score*. The maximum entropy criterion can be stated as: "when we make inferences on incomplete information, we should draw them from that probability distribution that has the maximum entropy permitted by the information which we do have"[12]. Thus, our criterion is applied in two steps: (i) computation of the highest MAP score and (ii) selection of the maximum entropy estimate, among those with high MAP score. We implement our

criterion on top of both, our new global solver and a multi-start EM procedure. The idea of using entropy to estimate parameters of BNs from incomplete samples has been previously advocated [4–6, 11, 22], yet our approach and those considerably differ: either they work with continuous variables and very few parameters, or they employ other inference approaches, such as the imprecise Dirichlet model [21], to later use entropy as criterion. We deal with discrete variables with BNs of a great numbers of parameters, and interpret the entropy criterion in a softer manner (as explained later on). Entropy methods have also been applied before for dealing with the uncertainty about the missingness mechanism, where the nature of the censoring data is unknown [2, 19] (we instead assume MAR).

This chapter is divided as follows. Section 6.2 presents the estimation problem and the methods to tackle it. Expectation-Maximization (EM) (Sect. 6.2.1) and nonlinear formulation (Sect. 6.2.2) are described, which are then compared in Sect. 6.2.3. The entropy-based idea is presented in Sect. 6.2.4. Section 6.3 presents experiments comparing the methods. Finally, Sect. 6.4 presents our concluding remarks.

6.2 Methods

We adopt Bayesian networks as framework for our study. Therefore, we assume that the reader is familiar with their basic concepts [13]. A Bayesian network (BN) is a triple $(\mathcal{G}, \mathcal{X}, \mathcal{P})$, where $\mathcal{G}$ is a directed acyclic graph with nodes associated to random variables $\mathcal{X} = \{X_1, \ldots, X_n\}$ over discrete domains $\{\Omega_{X_1}, \ldots, \Omega_{X_n}\}$ and $\mathcal{P}$ is a collection of probability values $p(x_j|\pi_j)$ with $\sum_{x_j \in \Omega_{X_j}} p(x_j|\pi_j) = 1$, where $x_j \in \Omega_{X_j}$ is a category or state of X_j and $\pi_j \in \times_{X \in \Pi_j} \Omega_X$ a (joint) state for the parents Π_j of X_j in $\mathcal{G}$. In a BN, every variable is conditionally independent of its nondescendants given its parents, according to $\mathcal{G}$. Given its independence assumptions, the joint probability distribution represented by a BN is obtained by $p(\mathbf{x}) = \prod_j p(x_j|\pi_j)$, where $\mathbf{x} \in \Omega_{\mathcal{X}}$ and all x_j, π_j (for every j) agree with $\mathbf{x}$.

Nodes of the graph and their associated random variables are interchanged.

The graph $\mathcal{G}$ and the variables $\mathcal{X}$ (and their domains) are assumed to be known; $\theta_{\mathbf{v}|\mathbf{w}}$ is used to denote an estimate for $p(\mathbf{v}|\mathbf{w})$ (with $\mathbf{v} \in \Omega_{\mathbf{V}}$, $\mathbf{w} \in \Omega_{\mathbf{W}}$, $\mathbf{V},\mathbf{W} \subseteq \mathcal{X}$).

We denote as $\mathbf{y}^i$ the i-th *incomplete* instance and by $\mathbf{Y}^i \subseteq \mathcal{X}$ the set of observed variables of the i.i.d. sampled instance i. Given the *incomplete* training data $\mathbf{y} = (\mathbf{y}^1, \ldots, \mathbf{y}^N)$ with N instances such that each $\mathbf{y}^i \in \Omega_{\mathbf{Y}^i}$, we denote by $N_{\mathbf{u}}$ the number of instances of $\mathbf{y}$ that are consistent with the state configuration $\mathbf{u} \in \Omega_{\mathbf{U}}$, where $\mathbf{U} \subseteq \mathcal{X}$. Parameters are estimated by maximizing the posterior probability given $\mathbf{y}$:

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta}} S_{\boldsymbol{\theta}}(\mathbf{y}) = \operatorname*{argmax}_{\boldsymbol{\theta}} \left(\sum_{i=1}^{N} \log \theta_{\mathbf{y}^i} + \alpha(\boldsymbol{\theta}) \right), \tag{6.1}$$

where α represents the prior:

$$\alpha(\boldsymbol{\theta}) = \log \prod_{j=1}^{n} \prod_{x_j} \prod_{\pi_j} \theta_{x_j|\pi_j}^{\alpha_{x_j,\pi_j}}, \text{ and } \alpha_{x_j,\pi_j} = \frac{\text{ESS}}{|\Omega_{X_j}| \cdot |\Omega_{\Pi_j}|},$$

where ESS stands for *equivalent sample size*, which we set to one, as usually done in the literature [13]. The argument $\mathbf{y}$ of $S_{\boldsymbol{\theta}}$ is omitted from now on (S means *score*).

In the experiments, in order to evaluate the quality of estimates, we measure the Kullback–Leibler (KL) divergence between the joint distribution represented by the true BN and the estimated BN (which we name *joint metric*); moreover, we also use the joint marginal distribution of all leaf nodes (named *reasoning metric*). The latter measures how close a reasoning about those leaf variables with the estimated model is to that of the true model

$$\mathrm{KL}_{\mathcal{P}(\mathbf{Z})}(\boldsymbol{\theta}) = \sum_{\mathbf{z}\in\Omega_{\mathbf{Z}}} p(\mathbf{z}) \log\left(\frac{p(\mathbf{z})}{\theta_{\mathbf{z}}}\right),$$

where $\mathbf{Z}$ are the leaves and $p(\mathbf{z}) = \sum_{\mathbf{x}\in\Omega_{\mathcal{X}\setminus\mathbf{Z}}} p(\mathbf{x},\mathbf{z})$ (and respectively for $\theta_{\mathbf{z}}$). This metric requires marginalizing out all nonleaf variables, so it involves all variables in the computation. Because of that, local errors in the estimates can compensate each other, and tend to smooth the differences among methods.

6.2.1 Expectation-Maximization

For a complete data set (that is, $\mathbf{Y}^i = \mathcal{X}$ for all i), we have a concave function on $\boldsymbol{\theta}$:

$$S_{\boldsymbol{\theta}} = \sum_{j=1}^{n} \sum_{x_j} \sum_{\pi_j} N'_{x_j,\pi_j} \log \theta_{x_j|\pi_j},$$

where $N'_{x_j,\pi_j} = N_{x_j,\pi_j} + \alpha_{x_j,\pi_j}$, and the estimate $\hat{\theta}_{x_j|\pi_j} = N'_{x_j,\pi_j}/(\sum_{x_j} N'_{x_j\pi_j})$ achieves highest MAP score. In the case of incomplete data, we have

$$S_{\boldsymbol{\theta}} = \sum_{i=1}^{N} \log \sum_{\mathbf{z}^i} \prod_{j=1}^{n} \theta_{x_j^i|\pi_j^i} + \alpha(\boldsymbol{\theta}), \tag{6.2}$$

where $\mathbf{x} = (\mathbf{y}^i, \mathbf{z}^i) = (x_1^i, \ldots, x_n^i)$ represents a joint state for all the variables in instance i. No closed-form solution is known, and one has to directly optimize $\max_{\boldsymbol{\theta}} S_{\boldsymbol{\theta}}$, subject to

$$\forall_j \forall_{\pi_j} : \sum_{x_j} \theta_{x_j|\pi_j} = 1, \quad \forall_j \forall_{x_j} \forall_{\pi_j} : \theta_{x_j|\pi_j} \geq 0. \tag{6.3}$$

The most common approach to optimize this function is to use the EM method, which completes the data with the expected counts for each missing variable given the observed variables, that is, variables Z_j^i are completed by "weights" $\hat{\theta}^k_{Z_j|\mathbf{y}^i}$ for each i, j of a missing value, where $\hat{\boldsymbol{\theta}}^{\boldsymbol{k}}$ represents the current estimate at iteration k. This idea is equivalent to weighting the chance of having $Z_j^i = z_j$ by the (current) distribution

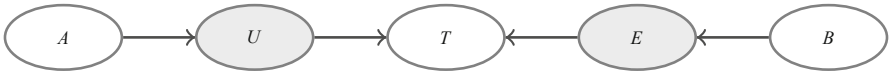

Fig. 6.1 Network BN_1, used in the description of the algorithm and later in the experiments; nodes affected by the missingness process have a grey background

of Z_j given $\mathbf{y}^i$ (this is known as the *E-step*, and requires computations over the BN instantiated with $\mathcal{P} = \hat{\boldsymbol{\theta}}^{\boldsymbol{k}}$ to obtain the estimated probability of missing values). Using these weights together with the actual counts from the data, the sufficient statistics values $N^k_{x_j,\pi_j}$ are computed for every x_j,π_j, and the next (updated) estimate $\hat{\boldsymbol{\theta}}^{\boldsymbol{k+1}}$ is obtained as if the data were complete: $\hat{\theta}^{k+1}_{x_j|\pi_j} = N'^k_{x_j,\pi_j}/(\sum_{x_j} N'^k_{x_j\pi_j})$, where $N'^k_{x_j,\pi_j} = N^k_{x_j,\pi_j} + \alpha_{x_j,\pi_j}$ (this is the *M-step*). As in the first step there is no current estimate $\hat{\boldsymbol{\theta}}^0$, an initial guess has to be used. Using the score itself to test convergence, this procedure achieves a saddle point of Eq. 6.1, which is usually a local optimum of the problem, and may vary according to the initial guess $\hat{\boldsymbol{\theta}}^0$. Hence, it is common to execute multiple runs of EM with distinct initial guesses and then to take the estimate with highest score among them.

6.2.2 Nonlinear Solver

In order to understand whether the good/bad quality of estimates is not simply a product of EM pitfalls to properly optimize Eq. 6.1, we build a systematic way to translate the parameter estimation into a compact nonlinear optimization problem, which is later (globally) solved with an optimization suite. The idea of directly optimizing the score function is not new (see e.g., [17]). Nevertheless, we are not aware of a method that translates the original score function into a simple formulation using symbolic variable elimination.

The main issue regards the internal summations of Eq. (6.2), because there is an exponential number of terms. We process them using a symbolic version of a variable elimination procedure as in [3], but the elimination method is run with target $\theta_{\mathbf{y}^i}$. Instead of numerical computations, it generates the polynomial constraints that precisely describe $\theta_{\mathbf{y}^i}$ in terms of (the still unknown) local conditional probability values of the specification of the BN. Because these values are to be found, they become the variables to be optimized in the polynomials. To clarify the method, we take the example of Fig. 6.1, where E, U might be missing, while the others are always observed. In this example, we need to write the constraints that describe θ_{a^i,b^i,t^i}, because these are instances in the data with missing u^i and e^i. The score function is: $\hat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta}} \max_s s$, subject to Eqs. 6.3 and

$$s \leq \alpha(\boldsymbol{\theta}) + \sum_{i \in N^M} \log \theta_{a^i,b^i,t^i} + \sum_{i \in N^{\neg M}} \log \theta_{a^i,b^i,t^i,e^i,u^i}\,, \tag{6.4}$$

where N^M, $N^{\neg M}$ are index sets of the instances with and without missing values, respectively. Note that an extra optimization variable s was introduced to make the objective function become a constraint (for ease of expose). All $\theta_{\mathbf{v}|\mathbf{w}}$ (for each possible argument **v,w**) and s are *unknowns* to be optimized by the solver. The summation in Eq. 6.4 can be shortened by grouping together elements related to the same states, and variables with no missing value can still be factorized out, obtaining,

$$s \leq \alpha(\boldsymbol{\theta}) + \sum_a N_a \log \theta_a + \sum_b N_b \log \theta_b + \sum_{i \in N^M} \log \theta_{t^i|a^i,b^i} + \sum_{a,u} N_{a,u}^{\neg M} \log \theta_{u|a} + \sum_{b,e} N_{b,e}^{\neg M} \log \theta_{e|b} + \sum_{e,t,u} N_{e,t,u}^{\neg M} \log \theta_{t|e,u}. \tag{6.5}$$

This equation is automatically built by the symbolic variable elimination procedure. As one can see, in this particular example the marginal distributions $p(A)$ and $p(B)$ can be estimated by the standard closed-form solution, as they are roots of the network and (in this example) their corresponding data are always complete. However, this is not true for every term in the equation. For instance, the summation $\sum_{i \in N^M} \log \theta_{t^i|a^i,b^i}$, where the sum runs over the categories a^i, b^i, t^i, comes in Eq. 6.5 and involves elements that are not direct parts of the network specification. It is exactly the job of the symbolic variable elimination to obtain the extra constraints

$$\theta_{t^i|a^i,e} = \sum_u \theta_{u|a^i} \cdot \theta_{t^i|u,e}\,, \qquad \theta_{t^i|a^i,b^i} = \sum_e \theta_{e|b^i} \cdot \theta_{t^i|a^i,e}. \tag{6.6}$$

These equations tie together the *auxiliary* optimization unknowns (such as $\theta_{t^i|a^i,e}$ and $\theta_{t^i|a^i,b^i}$) and the actual parameter estimates of interest, which are a part of the specification of the network (such as $\theta_{u|a^i}$, $\theta_{t^i|u,e}$, $\theta_{e|b^i}$). We emphasize that these derivations are not done by hand (with the user interaction), but instead they are *automatically* processed by the symbolic variable elimination procedure. The left-hand side of Eq. 6.6 comes from the symbolic (variable) elimination of u, while the right-hand side comes from the symbolic elimination of e. Together, they create a mathematical correspondence between $\theta_{t^i|a^i,b^i}$ and actual network parameters. After the symbolic preprocessing, it is up to the polynomial programming solver to optimize the nonlinear problem. We have implemented an adapted version of the reformulation-linearization technique [20], which is a global solver for the problem. The idea of their method is to relax the optimization problem into a linear programming that (somewhat tightly) outer-approximates the original problem. This is integrated into a branch-and-bound search that cuts the parameter space into smaller pieces until the relaxed problems are globally feasible. The reason for choosing such method is the simplicity of the relaxations that are solved at each step.

To make a parallel, the EM algorithm would have to compute $p(E, U|a^i, b^i, t^i)$ (with $\mathcal{P} = \hat{\boldsymbol{\theta}}^k$) for each instance i in the data set, in order to obtain the sufficient statistics of iteration k. Each such computation is in fact a procedure of similar complexity to the one we just did. The main difference between the methods is that we do not work with numbers but with a symbolic version of the computation. This

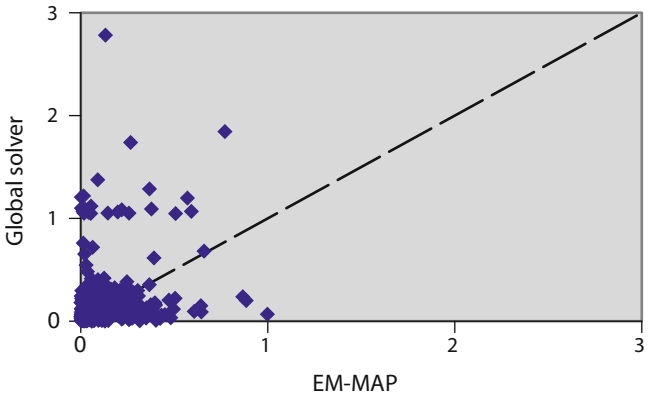

Fig. 6.2 Scatter plot of KL-divergences in the joint metric over data sets produced by BN_1 (detailed in the experiments section). Points *above* the *diagonal* show a worse estimate for the global solver, compared to EM-MAP

might incur some overhead, however, we need to perform the symbolic computation only once, while EM has to perform the (numerical) computations in every of its iterations. The translation method that creates all the constraints has shown to be slower than one but faster than just a few iterations of EM. As EM usually has to run many iterations until convergence, the translation does not introduce great complexity to the overall computation.

6.2.3 Global Solver vs. EM-MAP

We define as an EM-MAP, the approach which performs multiple runs of EM using different initialization points, eventually selecting the estimate corresponding to the highest MAP score. As explained in Sect. 6.2.2, we implemented a global solver based on nonlinear programming. In our experiments, the global solver achieved slightly higher MAP scores than EM-MAP, yielding however worse parameter estimates than EM-MAP. This phenomenon can be seen from Fig. 6.2, where points above the diagonal indicate better estimate for EM-MAP (using the *joint metric* as criterion) and points below the diagonal indicate better estimate for the global solver. Similar results were found in many experiments (not shown) comparing the global solver vs. EM-MAP, which suggests that selecting the estimate with the highest MAP score has drawbacks, being for instance subject to overfitting.

6.2.4 Discriminating High-Score Estimates by Entropy

It is often the case that many distinct estimates have very similar score, and simply selecting the one with the highest one might be an over-simplified decision. There are many estimates that lie within a tiny distance from the global maximum and can be as good as or better than the global one. In order to overcome the drawback just described, we propose the following criterion: to pick the parameter estimate with *maximum entropy, among those which have a high MAP score*. To identify the estimates with high MAP score we adopt a criterion similar to the Bayes factor. When discriminating among two competing models m_1 and m_2 on the basis of the data $\mathbf{y}$, the evidence in favor of m_1 can be considered substantial only if the Bayes factor $P(\mathbf{y}|m_1)/P(\mathbf{y}|m_2)$ is at least some threshold, for instance 2 or 3 [9], where $P(\mathbf{y}|m)$ represents the *marginal* likelihood given m: $P(\mathbf{y}|m) = \int P(\mathbf{y}|m,\theta)p(\theta)d\theta$. Because of the challenges that come with the missing data, we adopt a ratio of MAP scores (a full Bayesian approach would integrate over the parameters, but such computation would be intractable). We assume that if the ratio of the MAP scores among two competing parameter estimates is less than 2, there is no substantial evidence for preferring one over the other. In fact, we found our result to be robust by varying the threshold on the Bayes factor around 2. To choose among the competing estimates with high MAP score (whose MAP score is at least a half of the maximum MAP score known for the data set under consideration), we use the maximum entropy, thus choosing the least informative estimate given the available information [12]. This approach differs from standard maximum entropy inferences previously reported [11, 22] since we first check for high score estimates, and then maximize entropy among them. It can be formally written as:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{j=1}^{n} \sum_{\pi_j} \sum_{x_j} \theta_{x_j|\pi_j} \log \theta_{x_j|\pi_j} \quad \text{subject to} \quad S_{\boldsymbol{\theta}} + c \geq s^*, \tag{6.7}$$

where the function to be maximized is the *local entropy* of a Bayesian network [16], $s^* = \max_{\boldsymbol{\theta}} S_{\boldsymbol{\theta}}$ is the highest MAP score for the problem (which we compute before running this optimization), and c is the logarithm of the ratio of the MAP scores.

The optimization of Eq. 6.7 maximizes entropy, being however constrained to ensure that the MAP score $S_{\boldsymbol{\theta}}$ is high. In fact, we found our result to be robust on the threshold c when letting it vary between 2 and 3.

If it is to use the previously mentioned global solver, the optimization is tackled in two steps: first by globally optimizing the MAP score as already described, and then by solving Eq. 6.7. The quality of the estimates obtained in this case dramatically improves, as can be seen from Fig. 6.3. As the problem is NP-hard, we cannot expect this solver to obtain a global optimal solution in all problem instances. Hence, we adapted our idea to also work within the EM. In this case, we select, among the various estimates generated by the multi-start EM, the maximum entropy estimate among those which have a high MAP score. The high MAP score is checked by computing the ratio of the MAP score with respect to the highest MAP score obtained in the

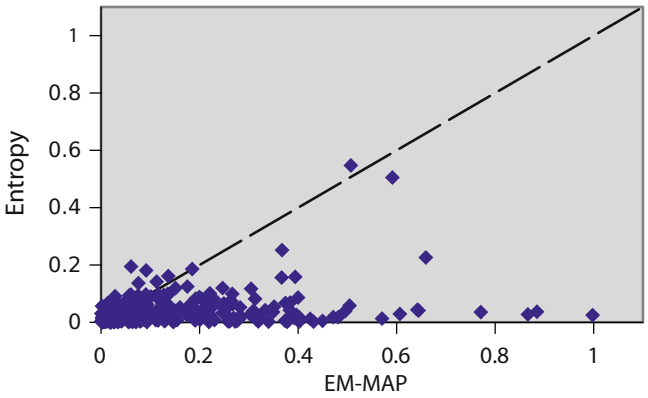

Fig. 6.3 Scatter plot of KL-divergences in the joint metric over data sets produced by BN_1 (detailed in the experiments section). Points *below* the *diagonal* show a better estimate for the global solver coupled with the entropy criterion compared to EM-MAP

different EM runs (thus, the computation of the highest score is only done in an approximate fashion). We call the resulting approach *EM-entropy*. This differs from the maximum entropy approach described before, because we focus only on the many estimates generated by the EM runs. The great benefit is that the implementation becomes straightforward: only a few changes on top of an already running EM suffice. The drawback of EM-entropy is that the true maximum entropy estimate might well be a nonoptimum estimate in terms of score. Thus, even if we increase the number of EM runs, the empirical entropy is still confined to saddle points of the score function, and the resulting estimate may differ. Nevertheless, the experiments will later show that the EM-entropy also produces significantly better estimates than MAP. An insight of the reason for which EM-entropy outperforms EM-MAP is given by Fig. 6.4, which shows an experiment where higher MAP scores do not necessarily imply a better estimate; instead, when comparing estimates that already have high MAP score (the right-most points in Fig. 6.4), entropy is more discriminative than the MAP score itself and has also a stronger correlation with the Kullback-Leibler (KL) divergence.

In any BN with more than a couple of variables, the number of parameters to estimate becomes quickly large and there is only a very small (or no) region of the parameter space with estimates that achieve the very same global maximum value. However, a feasibility region defined by a small percentage away from the maximum score is enough to produce a whole region of estimates, indicating that the region of high score estimates is almost (but not exactly) flat. This is expected in a high-dimensional parameter space of BNs.

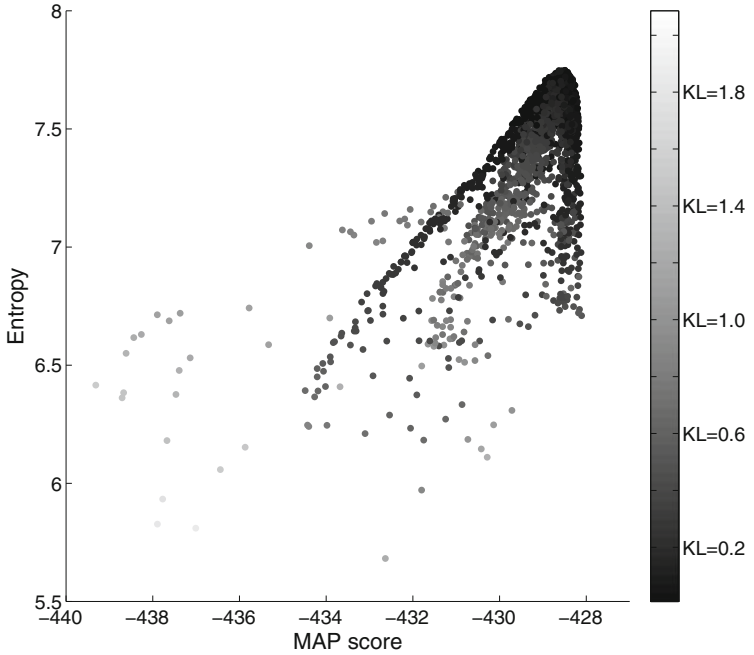

Fig. 6.4 Relation between KL divergence, entropy and score; *darker points* represent lower KL divergence between true and estimated joint distributions. The figure refers to one thousand EM runs performed on an incomplete training set of 200 samples

6.3 Experiments

We perform an empirical study using different BN graphs, sample sizes N and missingness process mp. The experiments were run using the open-source Bayesian Network Toolbox (BNT) [18] for MATLAB.

A triple $\langle \text{BN graph}, N, mp \rangle$ identifies a *setting*; for each setting, we perform 300 *experiments*, each defined as follows: (a) instantiation of the *reference BN*; (b) sampling of N complete instances from the reference BN; (c) application of the missingness process; (d) execution of EM from 30 different initializations; (e) execution of our solvers and estimation procedure using the different methods. We evaluate the quality of the estimates through the *joint metric* and the *reasoning metric* already introduced. We analyze the significance of the differences through the nonparametric Friedman test with significance level of 1 %. By a posthoc procedure applied on the statistic of the test, we generate a rank of methods for each setting and each metric.

The first set of experiments regards the BN graph of Fig. 6.1 (named BN_1), which has been used in previous sections to illustrate the methods. Variables A (binary) and B (ternary) have uniform distributions and are always observed; variables U, E, and T are binary (assuming states true and false); the value of T is defined by the logical

relation $T = E \wedge U$. Variable T is always observed, while U and E are affected by the missingness process: in particular, both U and E are observed if and only if T is true. Therefore, E and U are either both observed and positive, or nonobserved. The missingness process is MAR [13, Sect. 20.1.2] because given T (always observed) the probability of U and E to be missing does not depend on their actual values; E and U are missing in about 85 % of the sampled instances. We assume the conditional probabilities of T to be known, thus focusing on the difficulty of estimating the probabilities related to variables U and E. For both $\langle \mathrm{BN}_1, 100, \mathrm{MAR} \rangle$ and $\langle \mathrm{BN}_1, 200, \mathrm{MAR} \rangle$ and for both the joint and the reasoning metric, the Friedman test returned the following rank: 1st—entropy, 2nd)EM-entropy, 3rd—EM-MAP. The boxplots in the first row of Fig. 6.3 show that the entropy-based methods largely improve over EM-MAP; interestingly, the simple EM-entropy already delivers much of the gain achieved by the more sophisticated entropy method which relies on globally optimal solvers.

In a second set of experiments we use the graph $A \rightarrow B \rightarrow C$, which we call BN_2. We consider two different configurations of number of states for each node, 5-3-5 (meaning A,C with 5 states and B with 3) and 8-4-8 (A,C with 8 states and B with 4). In both cases, we make B randomly missing in 85 % of the instances. Each experiment now includes an additional step, namely the generation of random parameters of the reference BN. From the viewpoint of how realistic is this experiment, one may see BN_2 as a subnetwork (possibly repeated many times) within a much larger BN. For instance, if we see A as the joint parent set of B, and C as the joint children of B, this experiment regards the very same challenges of estimating a node's parameters (in this case B) with missing values in a BN of irrespective number of variables. This graph also captures the BN that could be used for clustering with EM [7]. Despite the simple graph of this BN, the estimation task requires to estimate from incomplete samples a nonnegligible number of parameters, referring to nodes B and C, respectively, $2 \cdot 5 + 4 \cdot 3 = 22$ and $8 \cdot 3 + 7 \cdot 4 = 52$, for each used configuration. To these numbers, one should add the marginals of A, which are however inferred from complete samples and whose estimate is thus identical for all methods. We adopted $N = 300$ for the 5-3-5 configuration and $N = 500$ for the 8-4-8 configuration. In both settings and the two metrics, we obtained the same rank: 1st—entropy, 2nd—EM-entropy, 3rd—EM-MAP. It is worth noting again that the simple EM-entropy improves over EM-MAP. The boxplots are shown in the second row of Fig. 6.3.

To further compare the behavior of EM-entropy and EM-MAP, we run experiments using well-known BNs: i) the Asia network (8 binary variables, 2 leaves) [15], ii) the Alarm network (37 variables with 2 to 4 states each, and 8 leaves) [1] and iii) BNs with randomly generated graphs with 20 variables. In each experiment, we randomly regenerated the parameters of the reference networks. In the case of randomly generated BN graphs, the experimental procedure also includes the generation of the random graph, which is accomplished before drawing the parameters. Given two variables X_i and X_j, an arc from X_i to X_j is randomly included with probability 1/3 if $i < j$ (no arc is included if $j \geq i$, which ensures that the graph is acyclic and has no loops). Furthermore, the maximum number of parents of each variable is set to 4 and the number of states per variable is randomly selected from 2 to 4. After

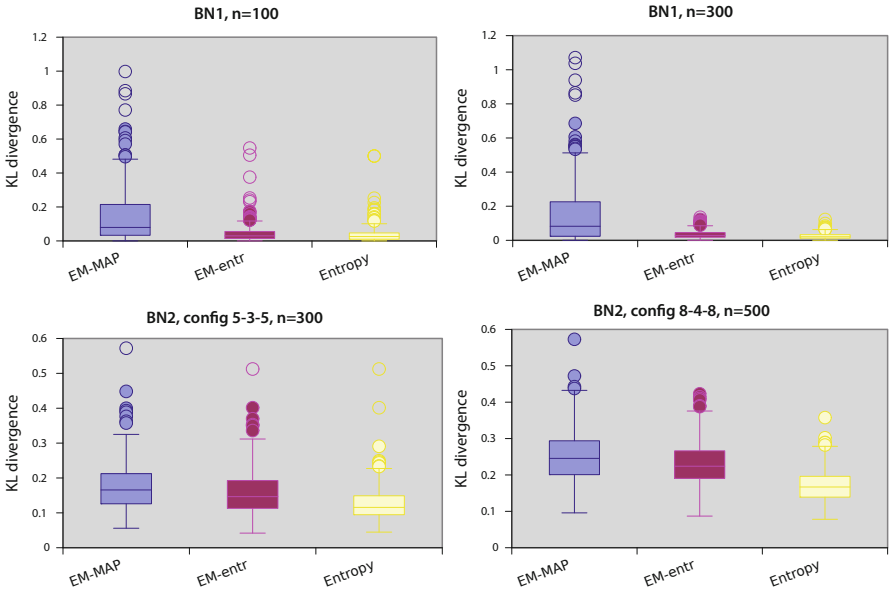

Fig. 6.5 Boxplot of KL-divergences for the joint metric over 300 runs of the experiment with BN_1 and BN_2 (the scale changes between top and bottom graphs)

Table 6.1 Relative medians of KL divergence, i.e., medians of entropy are presented (experiment-wise) divided by the median of MAP. Smaller numbers indicate better performance; in particular, values smaller than 1 indicate a smaller median than MAP

		$n = 100$		$n = 200$	
Net		q = 30 %	60 %	30 %	60 %
Asia	*Joint*	0.96	0.90	0.96	0.91
Asia	*Reasoning*	0.92	0.86	0.99	0.89
Alarm	*Joint*	0.93	0.88	0.93	0.89
Alarm	*Reasoning*	0.95	0.94	0.97	0.96
Random20	*Joint*	0.94	0.89	0.92	0.89
Random20	*Reasoning*	0.92	0.88	0.97	0.92

that, the experiments follow the same workflow as before. We consider a MCAR[1] process, which makes each single value missing with probability q; we use q equals to 30 and 60 %; we moreover consider sample sizes N of 100 and 200.

In all these experiments, EM-entropy performs significantly better than EM-MAP, with respect to both the joint and the reasoning metric. The quantitative difference of

[1] MCAR (or *missing completely at random*) indicates that the probability of each value being missing does not depend on the value itself, neither on the value of other variables.

performance can be seen in Table 6.1, which reports the *relative medians* of metrics, namely the medians of EM-entropy in a certain task, divided by the median of EM-MAP in the same task. The improvement of the median over EM-MAP ranges from 1 to 14 %; most importantly, it is consistent, occurring in all settings. As a final remark, the difference in performance increases when the estimation task is more challenging, typically when the percentage of missing data increases.

6.4 Conclusions

The most common approach to estimate the parameters of a Bayesian network in presence of incomplete data is to search for estimates with maximum posterior probability (MAP). MAP estimation is no harder than maximum likelihood estimation, over which it should be preferred because it yields estimates that are more resilient to overfitting. MAP estimation is much faster than full Bayesian estimation, but does not offer the same advantages of the latter. Many local maxima are usually present and several of them present high posterior probability. Selecting the one which maximizes it is not robust, since the difference among these competing estimates is generally very thin.

We presented an approach to select the least informative estimate, namely the maximum entropy one, among those which have a high posterior probability; our empirical analyses indicate that this approach consistently improves the quality of results. The approach has been implemented with a global solver developed by us and within EM, obtaining in both cases a significant improvement when compared to MAP. In particular, the EM-entropy method for inferring Bayesian networks can be promptly implemented on top of any existing EM implementation for that task. As a future work, we plan to apply these ideas in more general settings of parameter estimation problems from incomplete samples, not just restricted to Bayesian networks.

Acknowledgements The research in this paper has been partially supported by the Swiss NSF grant no. 200021_146606/1.

References

1. Beinlich, I.A., Suermondt, H.J., Chavez, R.M., Cooper, G.F.: The alarm monitoring system: a case study with two probabilistic inference techniques for belief networks. In: Proceedings of the 2nd European Conference on Artificial Intelligence. Medicine, vol. 38, pp. 247–256 (1989)
2. Cowell, R.G.: Parameter learning from incomplete data for Bayesian networks. In: Proceedings of the 7th International Workshop on Artificial Intelligence and Statistics. Morgan Kaufmann (1999)
3. de Campos, C.P., Cozman, F.G.: Inference in credal networks using multilinear programming. In: Proceedings of the 2nd Starting AI Researcher Symposium, pp. 50–61. IOS Press, Valencia (2004)

4. de Campos, C.P., Ji, Q.: Improving Bayesian network parameter learning using constraints. In: Proceedings of the 19th International Conference on Pattern Recognition, pp. 1–4. IEEE (2008)
5. de Campos, C.P., Zhang, L., Tong, Y., Ji, Q.: Semi-qualitative probabilistic networks in computer vision problems. J. Stat. Theory Pract. **3**(1), 197–210 (2009)
6. de Campos, C.P., Ji, Q.: Bayesian networks and the imprecise Dirichlet model applied to recognition problems. In: W. Liu (ed.) Symbolic and Quantitative Approaches to Reasoning with Uncertainty, Lecture Notes in Computer Science, vol. 6717, pp. 158–169. Springer, Berlin (2011)
7. de Campos, C.P., Rancoita, P.M.V., Kwee, I., Zucca, E., Zaffalon, M., Bertoni, F.: Discovering subgroups of patients from DNA copy number data using NMF on compacted matrices. PLoS ONE **8**(11), e79,720 (2013)
8. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. Royal Stat. Soc. Series B **39**(1), 1–38 (1977)
9. Good, I.J.: Studies in the history of probability and statistics. XXXVII A. M. Turing's statistical work in World War II. Biometrika **66**, 393–396 (1979)
10. Heckerman, D.: A tutorial on learning with Bayesian networks. In: Jordan, M. Learning in Graphical Models vol. 89, pp. 301–354. MIT, Cambridge (1998)
11. Huang, B., Salleb-Aouissi, A.: Maximum entropy density estimation with incomplete presence-only data. In: Proceedings of the 12th International Conference on Artificial Intelligence and Statistics: JMLR W&CP 5, pp. 240–247 (2009)
12. Jaynes, E.T.: On the rationale of maximum-entropy methods. Proc. IEEE **70**(9), 939–952 (1982)
13. Koller, D., Friedman, N.: Probabilistic Graphical Models. MIT, Cambridge (2009)
14. Lauritzen, S.L.: The EM algorithm for graphical association models with missing data. Comput. Stat. Data Anal. **19**(2), 191–201 (1995)
15. Lauritzen, S.L., Spiegelhalter, D.J.: Local computations with probabilities on graphical structures and their application to expert systems. J. Royal Stat. Soc. Series B **50**(2), 157–224 (1988)
16. Lukasiewicz, T.: Credal Networks under Maximum Entropy. In: Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence, pp. 363–370. Morgan Kaufmann Publishers Inc. (2000)
17. McLachlan, G.M., Krishnan, T.: The EM Algorithm and Extensions. Wiley, New York (1997)
18. Murphy, K.P.: The Bayes Net Toolbox for MATLAB. In: Comput. Sci. Stat. **33**, 331–350 (2001)
19. Ramoni, M., Sebastiani, P.: Robust learning with missing data. Mach. Learn. **45**(2), 147–170 (2001)
20. Sherali, H.D., Tuncbilek, C.H.: A global optimization algorithm for polynomial programming problems using a reformulation-linearization technique. J. Global Optim. **2**, 101–112 (1992)
21. Walley, P.: Statistical Reasoning with Imprecise Probabilities. Chapman and Hall, New York (1991)
22. Wang, S., Schuurmans, D., Peng, F., Zhao, Y.: Combining statistical language models via the latent maximum entropy principle. Mach. Learn. **60(1–3)**, 229–250 (2005)

Chapter 7
Bayesian Inference in Cumulative Distribution Fields

Ricardo Silva

Abstract One approach for constructing copula functions is by multiplication. Given that products of cumulative distribution functions (CDFs) are also CDFs, an adjustment to this multiplication will result in a copula model, as discussed by Liebscher (J Mult Analysis, 2008). Parameterizing models via products of CDFs has some advantages, both from the copula perspective (e.g. it is well-defined for any dimensionality) and from general multivariate analysis (e.g. it provides models where small dimensional marginal distributions can be easily read-off from the parameters). Independently, Huang and Frey (J Mach Learn Res, 2011) showed the connection between certain sparse graphical models and products of CDFs, as well as message-passing (dynamic programming) schemes for computing the likelihood function of such models. Such schemes allow models to be estimated with likelihood-based methods. We discuss and demonstrate MCMC approaches for estimating such models in a Bayesian context, their application in copula modeling, and how message-passing can be strongly simplified. Importantly, our view of message-passing opens up possibilities to scaling up such methods, given that even dynamic programming is not a scalable solution for calculating likelihood functions in many models.

7.1 Introduction

Copula functions are cumulative distribution functions (CDFs) in the unit cube $[0, 1]^p$ with uniform marginals. Copulas allow for the construction of multivariate distributions with arbitrary marginals—a result directly related to the fact that $F(X)$ is uniformly distributed in $[0, 1]$, if X is a continuous random variable with CDF $F(.)$. The space of models includes semiparametric models, where infinite-dimensional objects are used to represent the univariate marginals of the joint distribution, while a convenient parametric family provides a way to represent the dependence structure. Copulas also facilitate the study of measures of dependence that are invariant with respect to large classes of transformations of the variables, and the design of joint

R. Silva (✉)
Department of Statistical Science and Centre for Computational Statistics and Machine Learning, University College London, London, UK
e-mail: ricardo@stats.ucl.ac.uk

A. Polpo et al. (eds.), *Interdisciplinary Bayesian Statistics,*
Springer Proceedings in Mathematics & Statistics 118, DOI 10.1007/978-3-319-12454-4_7

distributions where the degree of dependence among variables changes at extreme values of the sample space. For a more detailed overview of copulas and its uses, please refer to [6, 11, 19].

A multivariate copula can in theory be derived from any joint distribution with continuous marginals: if $F(X_1, \ldots, X_p)$ is a joint CDF and $F_i(.)$ is the respective marginal CDF of X_i, then $F(F_1^{-1}(.), \ldots, F_p^{-1}(.))$ is a copula. A well-known result from copula theory, Sklar's theorem [19], provides the general relationship. In practice, this requires being able to compute $F_i^{-1}(.)$, which in many cases is not a tractable problem. Specialized constructions exist, particularly for recipes which use small dimensional copulas as building blocks. See [2, 12] for examples.

In this chapter, we provide algorithms for performing Bayesian inference using the product of copulas framework of Liebscher [14]. Constructing copulas by multiplying functions of small dimensional copulas is a conceptually simple construction, and does not require the definition of a hierarchy among observed variables as in [2] nor restricts the possible structure of the multiplication operation, as done by [12] for the space of copula densities that must obey the combinatorial structure of a tree. Our contribution is computational: since a product of copulas is also a CDF, we need to be able to calculate the likelihood function if Bayesian inference is to take place[1]. The structure of our contribution is as follows: (i) we simplify the results of [10], by reducing them to standard message passing algorithms as found in the literature of graphical models [3] (Sect. 7.3); (ii) for intractable likelihood problems, an alternative latent variable representation for the likelihood function is introduced, following in spirit the approach of [25] for solving doubly-intractable Bayesian inference problems by auxiliary variable sampling (Sect. 7.4).

We start with Sect. 7.2, where we discuss with some more detail the product of copulas representation. Some illustrative experiments are described in Sect.7.5. We emphasize that our focus in this short chapter is computational, and we will not provide detailed applications of such models. Some applications can be found in [9].

7.2 Cumulative Distribution Fields

Consider a set of random variables $\{U_1, \ldots, U_p\}$, each having a marginal density in $[0, 1]$. Realizations of this distribution are represented as $\{u_1, \ldots, u_p\}$. Consider the problem of defining a copula function for this set. The product of two or more CDFs is a CDF, but the product of two or more copulas is in general not a copula—marginals are not necessarily uniform after multiplication. In [14], different constructions based on products of copulas are defined so that the final result is also a copula. In particular,

[1] Pseudo-marginal approaches [1], which use estimates of the likelihood function, are discussed briefly in the last section.

for the rest of this chapter we will adopt the construction

$$C(u_1, \ldots, u_p) \equiv \prod_{j=1}^{K} C_j(u_1^{a_{1j}}, \ldots, u_p^{a_{pj}}) \tag{7.1}$$

where $a_{i1} + \cdots + a_{iK} = 1$, $a_{ij} \geq 0$ for all $1 \leq i \leq p$, $1 \leq j \leq K$, with each $C_j(\cdot, \ldots, \cdot)$ being a copula function.

Independently, Huang and Frey [8, 9] derived a product of CDFs model from the point of view of graphical models, where independence constraints arise due to the absence of some arguments in the factors (corresponding in (7.1) to setting some exponents a_{ij} to zero). Independence constraints from such models include those arising from models of marginal independence [4, 5].

Example 7.1 We first adopt the graphical notation of [4] to describe the factor structure of the cumulative distribution network (CDN) models of Huang and Frey, where a bidirected edge $U_m \leftrightarrow U_n$ is included if U_m and U_n appear together as arguments to any factor in the joint CDF product representation. For instance, for the model $C(u_1, u_2, u_3) \equiv C_1(u_1, u_2^{1/2}) C_2(u_2^{1/2}, u_3)$ we have the corresponding network

$$U_1 \leftrightarrow U_2 \leftrightarrow U_3$$

First, we can verify this is a copula function by calculating the univariate marginals. Marginalization is a computationally trivial operation in CDFs: since $C(u_1, u_2, u_3)$ means the probability $P(U_1 \leq u_1, U_2, \leq u_2, U_3 \leq u_3)$, one can find the marginal CDF of U_1 by evaluating $C(u_1, \infty, \infty)$. One can then verify that $P(U_i \leq u_i) = u_i$, $i = \{1, 2, 3\}$, which is the CDF of an uniform random variable given that $u_i \in [0, 1]$. One can also verify that U_1 and U_3 are marginally independent (by evaluating $C(u_1, \infty, u_3)$ and checking it factorizes), but that in general U_1 and U_3 are *not* conditionally independent given U_2. □

See [4, 5, 9] for an in-depth discussion of the independence properties of such models, and [14] for a discussion of the copula dependence properties. Such copula models can also be defined conditionally. For a (non-Gaussian) multiple regression model of outcome vector $\mathbf{Y}$ on covariate vector $\mathbf{X}$, a possible parameterization is to define the density of $p(y_i \mid \mathbf{x})$ and the joint copula $C(U_1, \ldots, U_p)$ where $U_i \equiv P(Y_i \leq y_i \mid \mathbf{x})$. Copula parameters can also be functions of $\mathbf{X}$.

Bayesian inference can be performed to jointly infer the posterior distribution of marginal and copula parameters for a given dataset. For simplicity of exposition, from now on we will assume our data is continuous and follows univariate marginal distributions in the unit cube. We then proceed to infer posteriors over copula parameters only[2]. We will also assume that for regression models the copula parameters do not depend on the covariate vector $\mathbf{x}$. The terms "CDN" and "cumulative distribution

[2] In practice, this could be achieved by fitting marginal models $\hat{F}_i(\cdot)$ separately, and transforming the data using plug-in estimates as if they were the true marginals. This framework is not uncommon in frequentist estimation of copulas for continuous data, popularized as "inference function for margins", IFM [11].

fields" will be used interchangeably, with the former emphasizing the independence properties that arise from the factorization of the CDF.

7.3 A Dynamic Programming Approach for Aiding Markov Chain Monte Carlo (MCMC)

Given the parameter vector θ of a copula function and data $\mathcal{D} \equiv \{\mathbf{U}^{(1)}, \ldots, \mathbf{U}^{(N)}\}$, we will describe Metropolis–Hastings approaches for generating samples from the posterior distribution $p(\theta \mid \mathcal{D})$. The immediate difficulty here is calculating the likelihood function, since (7.1) is a CDF function. Without further information about the structure of a CDF, the computation of the corresponding probability density function (PDF) has a cost that is exponential in the dimensionality p of the problem. The idea of a CDN is to be able to provide a computationally efficient way of performing this operation if the factorization of the CDF has a special structure.

Example 7.2 Consider a "chain-structured" copula function given by $C(u_1, \ldots, u_p) \equiv C_1(u_1, u_2^{1/2}) C_2(u_2^{1/2}, u_3^{1/2}) \ldots C_{p-1}(u_{p-1}^{1/2}, u_p)$. We can obtain the density function $c(u_1, \ldots, u_p)$ as

$$\begin{aligned} c(u_1, \ldots, u_p) &= \left[\frac{\partial^2 C_1(u_1, u_2^{1/2})}{\partial u_1 \partial u_2}\right]\left[\frac{\partial^{p-2} C_2(u_2^{1/2}, u_3^{1/2}) \ldots C_{p-1}(u_{p-1}^{1/2}, u_p)}{\partial u_3 \ldots \partial u_p}\right] + \\ &\quad \left[\frac{\partial C_1(u_1, u_2^{1/2})}{\partial u_1}\right]\left[\frac{\partial^{p-1} C_2(u_2^{1/2}, u_3^{1/2}) \ldots C_{p-1}(u_{p-1}^{1/2}, u_p)}{\partial u_2 \ldots \partial u_p}\right] \\ &\equiv \frac{\partial^2 C_1(u_1, u_2^{1/2})}{\partial u_1 \partial u_2} \times m_{2\to 1}(u_2) + \frac{\partial C_1(u_1, u_2^{1/2})}{\partial u_1} \times m_{2\to 1}(\bar{u}_2) \end{aligned}$$

Here, $m_{2\to 1} \equiv [m_{2\to 1}(u_2)\, m_{2\to 1}(\bar{u}_2)]^{\mathsf{T}}$ is a two-dimensional vector corresponding to the factors in the above derivation, known in the graphical modelling literature as a *message* [3]. Due to the chain structure of the factorization, computing this vector is a recursive procedure. For instance,

$$\begin{aligned} m_{2\to 1}(u_2) &= \left[\frac{\partial C_2(u_2^{1/2}, u_3^{1/2})}{\partial u_3}\right]\left[\frac{\partial^{p-3} C_3(u_3^{1/2}, u_4^{1/2}) \ldots C_{p-1}(u_{p-1}^{1/2}, u_p)}{\partial u_4 \ldots \partial u_p}\right] + \\ &\quad \left[C_2(u_2^{1/2}, u_3^{1/2})\right]\left[\frac{\partial^{p-2} C_3(u_3^{1/2}, u_4^{1/2}) \ldots C_{p-1}(u_{p-1}^{1/2}, u_p)}{\partial u_3 \ldots \partial u_p}\right] \\ &\equiv \frac{\partial C_2(u_2^{1/2}, u_3^{1/2})}{\partial u_3} \times m_{3\to 2}(u_3) + C_2(u_2^{1/2}, u_3^{1/2}) \times m_{3\to 2}(\bar{u}_3) \end{aligned}$$

implying that computing the two-dimensional vector $m_{2\to 1}$ corresponds to a summation of two terms, once we have precomputed $m_{3\to 2}$. This recurrence relationship corresponds to a $\mathcal{O}(p)$ dynamic programming algorithm. □

The idea illustrated by the above example generalizes to trees and junction trees. The generalization is implemented as a message passing algorithm by [8, 10] named the *derivative-sum-product* algorithm. Although [8] represents CDNs using *factor graphs* [13], neither the usual independence model associate with factor graphs holds in this case (instead the model is equivalent to other already existing notations, as the bidirected graphs used in [4]), nor the derivative-sum-product algorithm corresponds to the standard sum-product algorithms used to perform marginalization operations in factor graph models. Hence, as stated, the derivative-sum-product algorithm requires new software, and new ways of understanding approximations when the graph corresponding to the factorization has a high treewidth, making junction tree inference intractable [3]. In particular, in the latter case Bayesian inference is doubly-intractable (following the terminology introduced by [17]) since the likelihood function cannot be computed.

Neither the task of writing new software nor deriving new approximations are easy, with the full junction tree algorithm of [10] being considerably complex[3]. In the rest of this Section, we show a simple recipe on how to reduce the problem of calculating the PDF of a CDN to the standard sum-product problem.

Let (7.1) be our model. Let $\mathbf{z}$ be a p-dimensional vector of integers, each $z_i \in \{1, 2, ..., K\}$. Let $\mathcal{Z}$ be the p^K space of all possible assigments of $\mathbf{z}$. Finally, let $I(.)$ be the indicator function, where $I(x) = 1$ if x is a true statement, and zero otherwise.

The product rule states that

$$\frac{\partial^p C(u_1, \ldots, u_p)}{\partial u_1 \ldots \partial u_p} = \sum_{\mathbf{z} \in \mathcal{Z}} \prod_{j=1}^{K} \phi_j(\mathbf{u}, \mathbf{z}) \tag{7.2}$$

where

$$\phi_j(\mathbf{u}, \mathbf{z}) \equiv \frac{\partial^{\sum_i^p I(z_i = j)} C_j(u_1^{a_{1j}}, \ldots, u_p^{a_{pj}})}{\prod_{i \ s.t. \ z_i = j} \partial u_i}$$

To clarify, the set $i \ s.t. \ z_i = j$ are the indices of the set of variables $\mathbf{z}$ which are assigned the value of j within the particular term in the summation.

From this, we interpret the function

$$p_c(\mathbf{u}, \mathbf{z}) \equiv \prod_{j=1}^{K} \phi_j(\mathbf{u}, \mathbf{z}) \tag{7.3}$$

as a joint density/mass function over the space $[0, 1]^p \times \{1, 2, \ldots, K\}^p$ for a set of random variables $\mathbf{U} \cup \mathbf{Z}$. This interpretation is warranted by the fact that $p_c(.)$ is nonnegative and integrates to 1. For the structured case, where only a subset of

[3] Please notice that [10] also presents a way of calculating the gradient of the likelihood function within the message passing algorithm, and as such has also its own advantages for tasks such as maximum likelihood estimation or gradient-based sampling. We do not cover gradient computation in this chapter.

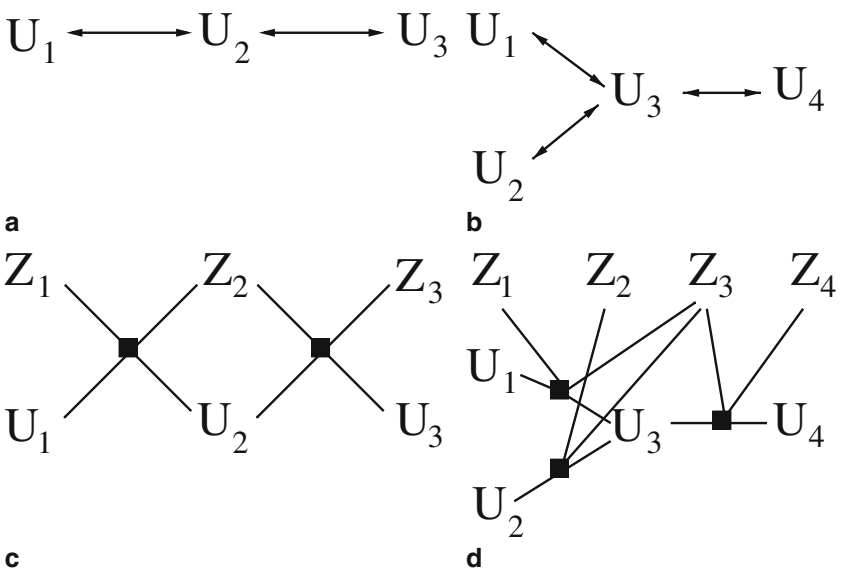

Fig. 7.1 In **a** and **b**, a simple chain and tree models represented both as bidirected graphs. In **c** and **d**, our corresponding extended factor graph representations with auxiliary variables **Z**

$\{U_1, \ldots, U_p\}$ are arguments to any particular copula factor $C_j(.)$, the corresponding sampling space of z_i is $\mathcal{Z}_i \subseteq \{1, 2, \ldots, K\}$, the indices of the factors which are functions of U_i. This follows from the fact that for a variable y unrelated to $\mathbf{x}$ we have $\partial f(\mathbf{x})/\partial y = 0$, and as such for $z_i = j$ we have $\phi_j(\mathbf{u}, \mathbf{z}) = p_c(\mathbf{u}, \mathbf{z}) = 0$ if $C_j(.)$ does not vary with u_i. From this, we also generalize the definition of $\mathcal{Z}$ to $\mathcal{Z}_1 \times \cdots \times \mathcal{Z}_p$.

The formulation (7.3) has direct implications to the simplification of the derivative-sum-product algorithm. We can now cast (7.2) as the marginalization of (7.3) with respect to **Z**, and *use standard message-passing algorithms*. The independence structure now follows the semantics of an undirected Markov network [3] rather than the bidirected graphical model of [4, 5]. In Figure 7.1 we show some examples using both representations, where the Markov network independence model is represented as a factor graph. The likelihood function can then be computed by this formulation of the problem using black-box message passing software for junction trees.

Now that we have the tools to compute the likelihood function, Bayesian inference can be carried. Assume we have for each $\phi_j(.)$ a set of parameters $\{\theta_j, \mathbf{a}_j\}$, of which we want to compute the posterior distribution given some data $\mathcal{D}$ using a MCMC method of choice. Notice that, after marginalizing **Z** and assuming the corresponding graph is connected, all parameters are mutually dependent in the posterior since (7.2) does not factorize in general. This mirrors the behaviour of MCMC algorithms for the Gaussian model of marginal independence as described by [24]. Unlike the Gaussian model, there are no hard constraints on the parameters across different factors. Unlike the Gaussian model, however, factorizations with high treewidth cannot be tractably treated.

7.4 Auxiliary Variable Approaches for Bayesian Inference

For problems with intractable likelihoods, one possibility is to represent it as the marginal of a latent variable model, and then sample jointly latent variables and the parameters of interest. Such auxiliary variables may in some contexts help with the mixing of MCMC algorithms, although we do not expect this to happen in our context, where conditional distributions will prove to be quite complex. In [24], we showed that even for small dimensional Gaussian models, the introduction of latent variables makes mixing much worse. It may nevertheless be an idea that helps to reduce the complexity of the likelihood calculation up to a practical point.

One straightforward exploration of the auxiliary variable approach is given by (7.3): just include in our procedure the sampling of the discrete latent vector $\mathbf{Z}^{(d)}$ for each data point d. The data-augmented likelihood is tractable and, moreover, a Gibbs sampler that samples each Z_i conditioned on the remaining indicators only needs to recompute the factors where variable U_i is present. The idea is straightforward to implement, but practitioners should be warned that Gibbs sampling in discrete graphical models also has mixing issues, sometime severely. A possibility to mitigate this problem is to "break" only a few of the factors by analytically summing over some, but not all, of the auxiliary $\mathbf{Z}$ variables in a way that the resulting summation is equivalent to dynamic programming in a tractable subgraph of the original graph. Only a subset will be sampled. This can be done in a way analogous to the classic cutset conditioning approach for inference in Markov random fields [20]. In effect, any machinery used to sample from discrete Markov random fields can be imported to the task of sampling $\mathbf{Z}$. Since the method in Section 7.3 is basically the result of marginalizing $\mathbf{Z}$ analytically, we describe the previous method as a "collapsed" sampler, and the method where $\mathbf{Z}$ is sampled as a "discrete latent variable" formulation of an auxiliary variable sampler.

This nomenclature also helps to distinguish those two methods for yet another third approach. This third approach is inspired by an interpretation of the independence structure of bidirected graph models as given via a directed acyclic graph (DAG) model with latent variables. In particular, consider the following DAG $\mathcal{G}'$ constructed from a bidirected graph $\mathcal{G}$: (i) add all variables of $\mathcal{G}$ as observed variables to $\mathcal{G}'$; (ii) for each clique S_i in $\mathcal{G}$, add at least on hidden variable to $\mathcal{G}'$ and make these variables a parent of all variables in S_i. If hidden variables assigned to different cliques are independent, it follows that the independence constraints among the observed variables of $\mathcal{G}$ and $\mathcal{G}'$ [21] are the same, as defined by standard graphical separation criteria[4]. See Figure 7.2 for examples.

The same idea can be carried over to CDNs. Assume for now that each CDF factor has a known representation given by

$$P_j(U_1 \leq u_1^{a_{1j}}, \ldots, U_p \leq u_p^{a_{pj}}) = \int \left\{ \prod_{i=1}^{p} P_{ij}(U_i \leq u_i^{a_{ij}} \mid \mathbf{h}_j) \right\} p_{\mathbf{h}_j}(\mathbf{h}_j)\, d\mathbf{h}_j$$

[4] Known as Global Markov conditions, as described by e.g. [21].

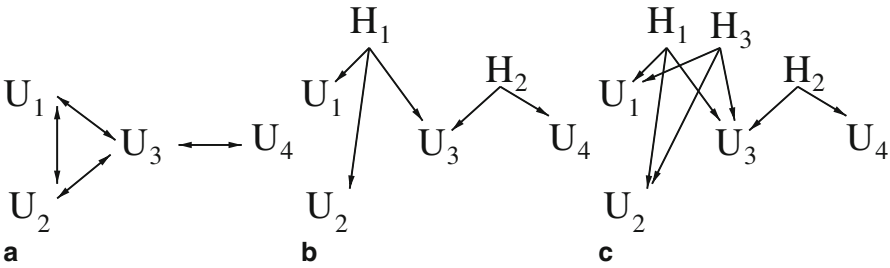

Fig. 7.2 The independence constraints implied by **a** among variables U_1, U_2 and U_3 are also implied by **b**) and **c** according to standard graphical separation criteria (the Global Markov properties described in, e.g. [21])

and that P_{ij} is not included in the product if U_i is not in factor j. Assume further that the joint distribution of $\mathbf{H} \equiv \cup_j \mathbf{H}_j$ factorizes as

$$p_{\mathbf{H}}(\mathbf{h}) \equiv \prod_{j=1}^{K} p_{\mathbf{h}_j}(\mathbf{h}_j)$$

It follows that the resulting PDF implied by the product of CDFs $\{C_j(.)\}$ will have a distribution Markov with respect to a (latent) DAG model over $\{\mathbf{U}, \mathbf{H}\}$, since

$$\begin{aligned} \frac{\partial^p P(\mathbf{U} \leq \mathbf{u} \mid \mathbf{h}) p_{\mathbf{H}}(\mathbf{h})}{\partial u_1 \ldots \partial u_p} &= p_{\mathbf{H}}(\mathbf{h}) \prod_{i=1}^{p} \frac{\partial \{\prod_{j \in Par(i)} P_{ij}(U_i \leq u_i^{a_{ij}} \mid \mathbf{h}_j)\}}{\partial u_i} \\ &\equiv p_{\mathbf{H}}(\mathbf{h}) \prod_{i=1}^{p} p_i(u_i \mid \mathbf{h}_{Par(i)}) \end{aligned} \tag{7.4}$$

where $Par(i)$ are the "parents" of U_i: the subset of $\{1, 2, ..., K\}$ corresponding to the factors where U_i appears. The interpretation of $p_i(.)$ as a density function follows from the fact that again $\prod_{j \in Par(i)} P_{ij}(U_i \leq u_i^{a_{ij}} \mid \mathbf{h}_j)$ is a product of CDFs and, hence, a CDF itself.

MCMC inference can then be carried out over the joint parameter and $\mathbf{H}$ space. Notice that even if all latent variables are marginally independent, conditioning on $\mathbf{U}$ will create dependencies[5], and as such mixing can also be problematic. However, particularly for dense problems where the number of factors is considerably smaller than the number of variables, sampling in the $\mathbf{H}$ space can potentially sound more attractive than sampling in the alternative $\mathbf{Z}$ space.

One important special case are products of Archimedean copulas. An Archimedean copula can be interpreted as the marginal of a latent variable model with a single latent variable, and exchangeable over the observations. A detailed account

[5] As a matter of fact, with one latent variable per factor, the resulting structure is a Markov network where the edge $H_{j_1} - H_{j_2}$ appears only if factors j_1 and j_2 have at least one common argument.

of Archimedean copulas is given by textbooks such as [11, 19], and their relation to exchangeable latent variable models in [7, 15]. Here we provide as an example a latent variable description of the Clayton copula, a popular copula in domains such as finance for allowing stronger dependencies at the lower quantiles of the sample space compared to the overall space.

Example 3 A set of random variables $\{U_1, \ldots, U_p\}$ follows a Clayton distribution with a scalar parameter θ when sampled according to the following generative model [7, 15]:

1. Sample random variable H from a Gamma $(1/\theta, 1)$ distribution
2. Sample p independent and identically distributed (iid) variables $\{X_1, \ldots, X_p\}$ from an uniform $(0, 1)$
3. Set $U_i = (1 - \log(X_i)/H)^{-1/\theta}$ □

This implies that, by using Clayton factors $C_j(\cdot)$, each associated with respective parameter θ_j and (single) gamma-distributed latent variable H_j, we obtain

$$P_{ij}(U_i \leq u_i^{a_{ij}} \mid h_j) = \exp(-h_j(u_i^{-\theta_j a_{ij}} - 1))$$

By multiplying over all parents of U_i and differentiating with respect to u_i, we get:

$$p_i(u_i \mid \mathbf{h}_{Par(i)}) = \left[\prod_{j \in Par(i)} \exp(-h_j(u_i^{-\theta_j a_{ij}} - 1))\right]\left[\sum_{j \in Par(i)} \theta_j a_{ij} h_j u_i^{-\theta_j a_{ij} - 1}\right] \tag{7.5}$$

A MCMC method can then be used to sample jointly $\{\{a_{ij}\}, \{\theta_j\}, \{\mathbf{H}^{(1)}, \ldots, \mathbf{H}^{(d)}\}\}$ given observed data with a sample size of d. We do not consider estimating the shape of the factorization (i.e. the respective graphical model structure learning task) as done in [23].

7.5 Illustration

We discuss two examples to show the possibilities and difficulties of performing MCMC inference in dense and sparse cumulative distribution fields. For simplicity we treat the exponentiation parameters a_{ij} as constants by setting them to be uniform for each variable (i.e. if U_i appears in k factors, $a_{ij} = 1/k$ for all of the corresponding factors). Also, we treat marginal parameters as known in this Bayesian inference exercise by first fitting them separately and using the estimates to generate uniform $(0, 1)$ variables.

The first one is a simple example in financial time series, where we have 5 years of daily data for 46 stocks from the S&P500 index, a total of 1257 data points. We fit a simple first-order linear autoregression model for each log-return Y_{it} of stock i at time t, conditioned on all 46 stocks at time $t - 1$. Using the least-squares estimator, we obtain the residuals and use the marginal empirical CDF to transform the residual data into approximately uniform U_i variables.

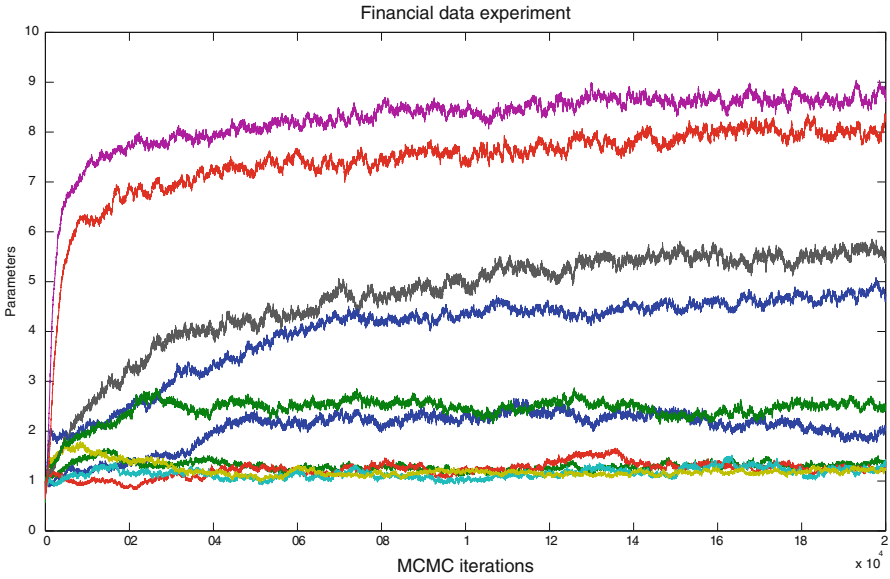

Fig. 7.3 MCMC traces of the 10 parameters for the 46 log-returns data. Convergence is slow, although each step is relatively cheap

The stocks are partitioned into 4 clusters according to the main category of business of the respective companies, with cluster sizes varying from 6 to 15. We define a CDF field using 10 factors: one for each cluster, and one for each pair of clusters using a Clayton copula for each factor. This is not a sparse model[6] in terms of independences among the observed $\{U_1, \ldots, U_{46}\}$. However, in the corresponding latent DAG model there are only 10 latent variables with each observation U_i having only two parents.

We used a Metropolis–Hastings method where each θ_i is sampled in turn conditioning on all other parameters using slice sampling [18]. Latent variables are sampled one by one using a simple random walk proposal. A gamma $(2, 2)$ prior is assigned to each copula parameter independently. Figure 7.3 illustrates the trace obtained by initializing all parameters to 1. Although each iteration is relatively cheap, convergence is substantially slow, suggesting that latent variables and parameters have a strong dependence in the posterior. As is, the approach does not look particularly practical. Better proposals than random walks are necessary, with slice sampling each latent variable being far too expensive and not really addressing the posterior dependence between latent variables and parameters.

[6] Even though it is still very restricted, since Clayton copulas have single parameters. A plot of the residuals strongly suggests that a t-copula would be a more appropriate choice, but our goal here is just to illustrate the algorithm.

Our second experiment is a simple illustration of the proposed methods for a sparse model. Sparse models can be particularly useful to model residual dependence structure, as in the structural equation examples of [23]. Here we use synthetic data on a simple chain $U_1 \leftrightarrow \cdots \leftrightarrow U_5$ using all three approaches: one where we collapse the latent variables and perform MCMC moves using only the observed likelihood calculated by dynamic programming; another where we sample the four continuous latent variables explicitly (the "continuous latent" approach); and the third, where we simply treat our differential indicators as discrete latent variables (the "discrete latent" approach). Clayton copulas with gamma $(2, 2)$ priors were again used, and exponents a_{ij} were once again fixed uniformly. As before, slice sampling was used for the parameters, but not for the continuous latent variables.

Figure 7.4 summarizes the result of a synthetic study with a random choice of parameter values and a chain of five variables (a total of 4 parameters). For the collapsed and discrete latent methods, we ran the chain for 1000 iterations, while we ran the continuous latent method for 10,000 iterations with no sign of convergence. The continuous latent method had a computational cost of about three to four times less than the other two methods. Surprisingly, the collapsed and discrete latent methods terminated in roughly the same amount of wallclock time, but in general we expect the collapsed sampler to be considerably more expensive. The effective sample size for the collapsed method along the four parameters was $(1000, 891, 1000, 903)$ and for the discrete latent case we obtained $(243, 151, 201, 359)$.

7.6 Discussion

Cumulative distribution fields provide another construction for copula functions. They are particularly suitable for sparse models where many marginal independences are expected, or for conditional models (as in [23]) where residual association after accounting for major factors is again sparsely located. We did not, however, consider the problem of identifying which sparse structures should be used, and focused instead on computing the posterior distribution of the parameters for a fixed structure.

The failure of the continuous latent representation as auxiliary variables in a MCMC sampler was unexpected. We conjecture that more sophisticated proposals than our plain random walk proposals should make a substantial difference. However, the main advantage of the continuous latent representation is for problems with large factors and a small number of factors compared to the number of variables. In such a situation perhaps the product of CDFs formulation should not be used anyway, and practitioners should resort to it for sparse problems. In this case, both the collapsed and the discrete latent representations seem to offer a considerable advantage over models with explicit latent variable representations (at least computationally), a result that was already observed for a similar class of independence models in the more specific case of Gaussian distributions [24].

An approach not explored here was the pseudo-marginal method [1], were an in place of the intractable likelihood function we use a positive unbiased estimator. In

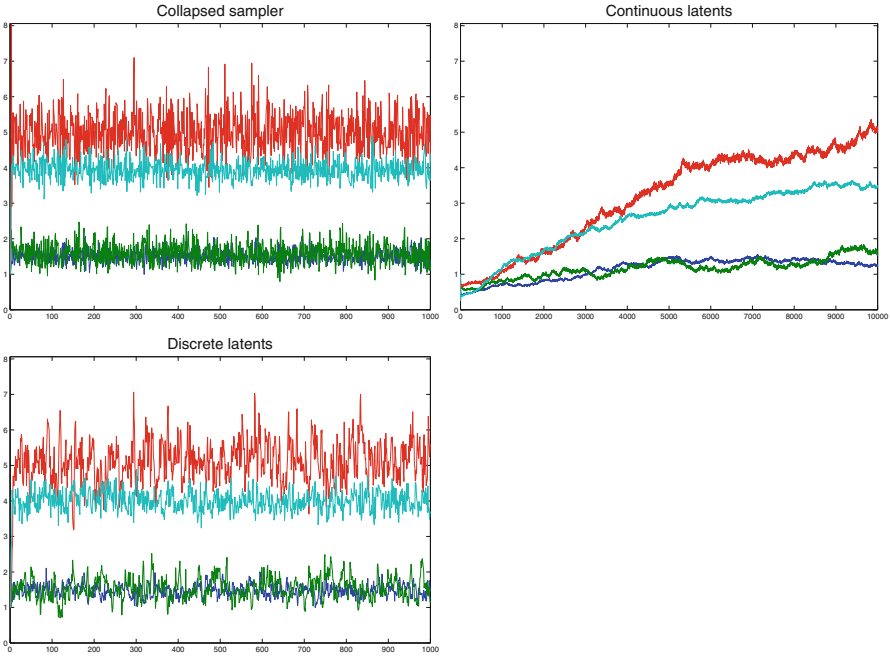

Fig. 7.4 Sampling performance for the synthetic case study using the three different methods

principle, the latent variable formulations allow for that. However, in a preliminary experiment where we used the very naive uniform distribution as an importance distribution for the discrete variables $\mathbf{Z}$, in a 10-dimensional chain problem with 100 data points, the method failed spectacularly. That is, the chain hardly ever moved. Far more sophisticated importance distributions will be necessary here.

Expectation-propagation (EP) [16] approaches can in principle be developed as alternatives. A particular interesting feature of this problem is that marginal CDFs can be read off easily, and as such energy functions for generalized EP can be derived in terms of actual marginals of the model.

For problems with discrete variables, the approach can be used almost as is by introducing another set of latent variables, similarly to what is done in probit models. In the case where dynamic programming by itself is possible, a modification of (7.1) using differences instead of differentiation leads to a similar discrete latent variable formulation (see the Appendix of [22]) without the need of any further set of latent variables. However, the corresponding function is not a joint distribution over $\mathbf{Z} \cup \mathbf{U}$ anymore, since differences can generate negative numbers.

Some characterization of the representational power of products of copulas was provided by [14], but more work can be done and we also conjecture that the point of view provided by the continuous latent variable representation described here can aid in understanding the constraints entailed by the cumulative distribution field construction.

Acknowledgements The author would like to thank Robert B. Gramacy for the financial data. This work was supported by a EPSRC grant EP/J013293/1.

References

1. Andrieu, C., Roberts, G.: The pseudo-marginal approach for efficient Monte Carlo computations. Ann. Stat. **37**, 697–725 (2009)
2. Bedford, T., Cooke, R.: Vines: a new graphical model for dependent random variables. Ann. Stat. **30**, 1031–1068 (2002)
3. Cowell, R., Dawid, A., Lauritzen, S., Spiegelhalter, D.: Probabilistic Networks and Expert Systems. Springer, Heidelberg (1999)
4. Drton, M., Richardson, T.: A new algorithm for maximum likelihood estimation in Gaussian models for marginal independence. Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence (2003)
5. Drton, M., Richardson, T.: Binary models for marginal independence. J. R. Stat. Soc. Ser. B **70**, 287–309 (2008)
6. Elidan, G.: Copulas in machine learning. Lecture notes in statistics (to appear)
7. Hofert, M.: Sampling Archimedean copulas. Comput. Stat. Data Anal. **52**, 5163–5174 (2008)
8. Huang, J., Frey, B.: Cumulative distribution networks and the derivative-sum-product algorithm. Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence (2008)
9. Huang, J., Frey, B.: Cumulative distribution networks and the derivative-sum-product algorithm: models and inference for cumulative distribution functions on graphs. J. Mach. Learn. Res. **12**, 301–348 (2011)
10. Huang, J., Jojic, N., Meek, C.: Exact inference and learning for cumulative distribution functions on loopy graphs. Adv. Neural Inf. Process. Syst. **23**, 874–882 (2010)
11. Joe, H.: Multivariate Models and Dependence Concepts. Chapman-Hall, London (1997)
12. Kirshner, S.: Learning with tree-averaged densities and distributions. In: Platt, J. C., Koller, D., Singer, Y., Roweis, S. T. (eds.) Advances in Neural Information Processing Systems 20, pp. 761–768. Curran Associates, Inc. (2008)
13. Kschischang, F., Frey, B., Brendan, J., Loeliger, H.A.: Factor graphs and the sum-product algorithm. IEEE Trans. Inf. Theory **47**, 498–519 (2001)
14. Liebscher, E.: Construction of asymmetric multivariate copulas. J. Multivar. Anal. **99**, 2234–2250 (2008)
15. Marshall, A., Olkin, I.: Families of multivariate distributions. J. Am. Statist. Assoc. **83**, 834–841 (1988)
16. Minka, T.: Automatic choice of dimensionality for PCA. Adv. Neural Inf. Process. Syst. **13**, 598–604 (2000)
17. Murray, I., Ghahramani, Z., MacKay, D.: MCMC for doubly-intractable distributions. Proceedings of 22nd Conference on Uncertainty in Artificial Intelligence (2006)
18. Neal, R.: Slice sampling. Ann. Stat. **31**, 705–767 (2003)
19. Nelsen, R.: An Introduction to Copulas. Springer-Verlag, New York (2007)
20. Pearl, J.: Probabilistic Reasoning in Expert Systems: Networks of Plausible Inference. Morgan Kaufmann, San Francisco (1988)
21. Richardson, T., Spirtes, P.: Ancestral graph Markov models. Ann. Stat. **30**, 962–1030 (2002)
22. Silva, R.: Latent composite likelihood learning for the structured canonical correlation model. Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence, UAI (2012)
23. Silva, R.: A MCMC approach for learning the structure of Gaussian acyclic directed mixed graphs. In: P. Giudici, S. Ingrassia, M. Vichi (eds.) Statistical Models for Data Analysis, pp. 343–352. Springer, Heidelberg (2013)
24. Silva, R., Ghahramani, Z.: The hidden life of latent variables: Bayesian learning with mixed graph models. J. Machi. Learn. Res. **10**, 1187–1238 (2009)
25. Walker, S.: Posterior sampling when the normalising constant is unknown. Commun. Stat. Simul. Comput. **40**, 784–792 (2011)

Chapter 8
MCMC-Driven Adaptive Multiple Importance Sampling

Luca Martino, Víctor Elvira, David Luengo and Jukka Corander

Abstract Monte Carlo (MC) methods are widely used for statistical inference and stochastic optimization. A well-known class of MC methods is composed of importance sampling (IS) and its adaptive extensions (such as adaptive multiple IS and population MC). In this work, we introduce an iterated batch importance sampler using a population of proposal densities, which are adapted according to a Markov Chain Monte Carlo (MCMC) technique over the population of location parameters. The novel algorithm provides a global estimation of the variables of interest iteratively, using all the generated samples weighted according to the so-called deterministic mixture scheme. Compared with a traditional multiple IS scheme with the same number of samples, the performance is substantially improved at the expense of a slight increase in the computational cost due to the additional MCMC steps. Moreover, the dependence on the choice of the cloud of proposals is sensibly reduced, since the proposal density in the MCMC method can be adapted in order to optimize the performance. Numerical results show the advantages of the proposed sampling scheme in terms of mean absolute error.

8.1 Introduction

Monte Carlo methods are widely used in different fields [4, 15]. Importance sampling (IS) [9, 12] is a well-known Monte Carlo (MC) methodology to compute efficiently integrals involving a complicated multidimensional target probability density function (pdf), $\pi(\mathbf{x})$ with $\mathbf{x} \in \mathbb{R}^n$. Moreover, it is often used in order to calculate the normalizing constant of $\pi(\mathbf{x})$ (also called *partition function*) [12], useful in several

L. Martino (✉) · J. Corander
Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland
e-mail: luca.martino@helsinki.fi

V. Elvira
Department of Signal Theory and Communications, Universidad Carlos III de Madrid, Leganés, Spain

D. Luengo
Department of Signal Theory and Communications, Universidad Politécnica de Madrid, Madrid, Spain

A. Polpo et al. (eds.), *Interdisciplinary Bayesian Statistics,*
Springer Proceedings in Mathematics & Statistics 118, DOI 10.1007/978-3-319-12454-4_8

applications [13], like model selection. The IS technique draws samples from a simple proposal pdf, $q(\mathbf{x})$, assigning weights to them according to the ratio between the target and the proposal, i.e., $w(\mathbf{x}) = \frac{\pi(\mathbf{x})}{q(\mathbf{x})}$. However, although the validity of this approach is guaranteed under mild assumptions, the variance of the estimator depends critically on the discrepancy between the shape of the proposal and the target. For this reason, Markov Chain Monte Carlo (MCMC) methods are usually preferred for large dimensional applications [1, 5, 7, 10]. Nevertheless, MCMC algorithms also present several issues; the diagnostic of the convergence is often difficult, and it is not straightforward to estimate the partition function given the generated samples.

In order to solve these issues, several works are devoted to the design of adaptive IS (AIS) schemes [12], where the proposal density is updated by learning from all the previously generated samples. The population Monte Carlo (PMC) [2] and the adaptive multiple importance sampling (AMIS) [3] methods are two general schemes that combine the proposal adaptation idea with the cooperative use of a population of different proposal pdfs. In PMC, a cloud of particles is propagated and then replicated or "killed" by resampling steps. In AMIS, a single proposal is updated online by taking into account the information provided by the previous weighted samples, as in a standard adaptive IS. Since the proposal is changed, this results in a sequence of proposals. All the samples generated by this sequence of pdfs are adequately weighted and used to build a global estimator. Moreover, the single proposal, adapted online, can be a mixture of proposals itself, but the adaptation procedure becomes much more complicated in this case, involving clustering techniques for instance.

This work is an attempt to mix together the IS and MCMC approaches, while preserving the advantages of both. Thus, we introduce a novel population scheme, *Markov adaptive multiple importance sampling* (MAMIS), which has features and behavior in between AMIS and PMC. MAMIS draws samples from different proposal densities at each iteration, weighting these samples according to the so-called deterministic mixture approach proposed in [11, 14] for a fixed (i.e., nonadaptive) setting. At each iteration, the MAMIS algorithm computes iteratively a *global* IS estimate, taking into account all the generated samples up to that point (similarly to AMIS). The main difference with respect to the existing AMIS and PMC schemes lies in the more streamlined adaptation procedure of MAMIS. MAMIS starts with a cloud of N proposals initialized randomly or according to the prior information available. The algorithm is then divided into groups of T_a iterations (so-called epochs), where the proposals are kept fixed and T_a samples are drawn from each one. At the end of every epoch, $\mathcal{T}$ iterations of an MCMC technique are applied to update the location parameters of the proposal pdfs. In this work, we use the *Sample Metropolis-Hastings algorithm* (SMH) [8, Chap. 6] to improve the positions of the proposals. Moreover, unlike PMC, the novel technique does not require resampling steps to prevent the degeneracy of the mixture, thus avoiding the loss of diversity in the population, which is a common problem for sampling-importance-resampling type algorithms.

The new algorithm increases the robustness with respect to the choice of the proposal parameters, and the adaptation is also simpler than in other adaptive IS methods. One reason for this is that the scale parameters of the proposals are not adapted; as an efficient strategy, we suggest using different (fixed) scale parameters

for the N pdfs in the IS scheme, and adapting the variance of the proposal used in the SMH algorithm. However, even if the proposal of SMH is not properly chosen, after several iterations the cloud of the location parameters will be distributed according to the target distribution, due to the application of a (valid) MCMC technique. In this sense, MAMIS is also easier to analyze from a theoretical point of view than AMIS (see discussion in [3]).

8.2 Problem Statement

In many applications, we are interested in inferring a variable of interest given a set of observations or measurements. Let us consider the variable of interest, $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^n$, and let $\mathbf{y} \in \mathbb{R}^d$ be the observed data. The posterior pdf is then

$$p(\mathbf{x}|\mathbf{y}) = \frac{\ell(\mathbf{y}|\mathbf{x})g(\mathbf{x})}{Z(\mathbf{y})} \propto \ell(\mathbf{y}|\mathbf{x})g(\mathbf{x}), \tag{8.1}$$

where $\ell(\mathbf{y}|\mathbf{x})$ is the likelihood function, $g(\mathbf{x})$ is the prior pdf and $Z(\mathbf{y})$ is the model evidence or partition function (useful in model selection).

In general, $Z(\mathbf{y})$ is unknown, so we consider the corresponding (usually unnormalized) target pdf,

$$\pi(\mathbf{x}) = \ell(\mathbf{y}|\mathbf{x})g(\mathbf{x}). \tag{8.2}$$

Our goal is computing efficiently some moment of $\mathbf{x}$, i.e., an integral measure with respect to the target pdf,

$$I = \frac{1}{Z}\int_{\mathcal{X}} f(\mathbf{x})\pi(\mathbf{x})d\mathbf{x}, \tag{8.3}$$

where

$$Z = \int_{\mathcal{X}} \pi(\mathbf{x})d\mathbf{x}. \tag{8.4}$$

Since both $\pi(\mathbf{x})$ and Z depend on the observations $\mathbf{y}$, the notation $\pi(\mathbf{x}|\mathbf{y})$ and $Z(\mathbf{y})$ would be more precise. However, as the observations are fixed, in the sequel we remove the dependence on $\mathbf{y}$ to simplify the notation.

8.3 Markov Adaptive Multiple Importance Sampling (MAMIS)

The MAMIS algorithm estimates Z and I by drawing samples from a population of proposals whose location parameters are adapted following an MCMC technique. For the sake of simplicity, we only consider a population of Gaussian proposals with fixed covariance matrices and we adapt only the means. However, the underlying idea is more general; many kinds of proposals could be used, including mixtures of different types of proposals.

8.3.1 Overview of the MAMIS Algorithm

1. **Initialization:** Set $t = 1$, $m = 0$, $\hat{I}_0 = 0$ and $L_0 = 0$. Choose N *normalized* Gaussian proposal pdfs,

$$q_i^{(0)}(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_i^{(0)}, \mathbf{C}_i), \qquad i = 1, \dots, N,$$

with mean vectors $\boldsymbol{\mu}_i^{(0)}$ and covariance matrices $\mathbf{C}_i$ ($i = 1, \dots, N$).
Let T be the total number of iterations. Select the number of iterations per epoch, $T_a \geq 1$, and the total number of iterations, $T = MT_a$, with $M \leq T \in \mathbb{Z}^+$ denoting the number of adaptation epochs.
2. **IS steps:**
   a. Draw $\mathbf{z}_i \sim q_i^{(m)}(\mathbf{x})$ for $i = 1, \dots, N$.
   b. Compute the importance weights,

$$w_i = \frac{\pi(\mathbf{z}_i)}{\frac{1}{N}\sum_{j=1}^{N} q_j^{(m)}(\mathbf{z}_i)}, \quad i = 1, \dots, N, \tag{8.5}$$

   and normalize them,

$$\bar{w}_i = \frac{w_i}{S}, \tag{8.6}$$

   where $S = \sum_{j=1}^{N} w_j$.
3. **Iterative IS estimation:** Calculate the "current" estimate of I,

$$\hat{J}_t = \sum_{i=1}^{N} \bar{w}_i f(\mathbf{z}_i), \tag{8.7}$$

and update the *global estimate*, using the recursive formula

$$\hat{I}_t = \frac{1}{L_{t-1} + S}\left(L_{t-1}\hat{I}_{t-1} + S\hat{J}_t\right), \tag{8.8}$$

where $L_t = L_{t-1} + S$.
Note that $\hat{Z}_t = \frac{1}{Nt} L_t$.
4. **MCMC adaptation:** If $t = kT_a$ ($k = 1, 2, \dots, M$), then perform $\mathcal{T}$ iterations of an MCMC technique over the current population of means,

$$\mathcal{P}^{(m)} = \{\boldsymbol{\mu}_1^{(m)}, ..., \boldsymbol{\mu}_N^{(m)}\},$$

to obtain a new population,

$$\mathcal{P}^{(m+1)} = \{\boldsymbol{\mu}_1^{(m+1)}, ..., \boldsymbol{\mu}_N^{(m+1)}\},$$

and set $q_i^{(m+1)} = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_i^{(m+1)}, \mathbf{C}_i)$.

5. **Stopping rule:** If $t < T$, set $t = t + 1$ and repeat from step 2. Otherwise, end. This is the simplest possibility, but more complex rules can be easily designed.
6. **Outputs:** Return the estimate of the desired integral,

$$\hat{I}_T \approx I = \frac{1}{Z} \int_{\mathcal{X}} f(\mathbf{x})\pi(\mathbf{x})d\mathbf{x}, \tag{8.9}$$

as well as the normalizing constant of the target pdf,

$$\hat{Z}_T \approx Z = \int_{\mathcal{X}} \pi(\mathbf{x})d\mathbf{x}. \tag{8.10}$$

The final locations of the Gaussians (i.e., their means, $\boldsymbol{\mu}_i^{(M)}$ for $i = 1, \ldots, N$) could also be used to estimate the locations of the modes of $\pi(\mathbf{x})$ (e.g., to perform maximum a posteriori estimation).

8.3.2 Basic Features and Important Remarks

Note that the MAMIS algorithm is basically an iterated batch importance sampler where, at some transition points (i.e., $t = mT_a$), the cloud of means are moved following an MCMC technique. Let us also remark that the algorithm works on two different time scales, t and m. At the transition iterations between two epochs ($t = mT_a$ with $m = 1, \ldots, M$), the parameters of the proposals, $\boldsymbol{\mu}_i^{(m)}$ for $1 \leq i \leq N$, are updated. Moreover, note that:

1. All the different proposal pdfs should be normalized to provide a correct IS estimation.
2. At each iteration ($t = 1, \ldots, T = MT_a$), MAMIS computes the "current" estimate of the desired integral, $\hat{J}_t$, and updates recursively the global estimates of the desired integral and the normalizing constant, $\hat{I}_t$ and $\hat{Z}_t$ respectively.
3. The "current" estimate $\hat{J}_t$ is obtained using the *deterministic mixture* approach proposed in [11, 14] for a fixed (i.e., nonadaptive) setting. This strategy yields more robust IS estimators.
4. The global estimators, $\hat{I}_T$ and $\hat{Z}_T$, are iteratively obtained by an importance sampling approach using NT total samples drawn (in general) from NT different proposals: N initial proposals chosen by the user, and $N(T-1)$ proposals adapted by the algorithm.
5. Different stopping rules can be applied to ensure that the global estimators produce the desired degree of accuracy, in terms of Monte Carlo variability. For instance, one possibility is taking into account the variation of the estimate over time. In this case, the algorithm could be stopped at any iteration $t^* < T$.

Note that in the previous description the index t could be removed. Indeed, *within an epoch* the proposals do not change, so we could draw T_a i.i.d. samples directly from each proposal and then adapt the proposals using these samples. However, we prefer to maintain the previous description to emphasize the fact that the accuracy of the estimator can be tested at each iteration t, and that the algorithm could be stopped at any time.

8.4 Adaptation via MCMC

In this work, we propose to adapt the location parameters of the proposal pdfs applying a suitable MCMC technique over the cloud of means $\boldsymbol{\mu}_i$. A possible technique is the *SMH* algorithm [8, Chap. 6]; at the transition iterations between two epochs ($t = mT_a$ with $m = 1, \ldots, M$), we apply $\mathcal{T}$ iterations of the SMH algorithm, in order to improve the positions of the means and, as a consequence, to perform the adaptation in the MAMIS algorithm. Consider a generalized target pdf, extended in this way,

$$\pi_g(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N) \propto \prod_{i=1}^{N} \pi(\boldsymbol{\mu}_i),$$

where each marginal $\pi(\boldsymbol{\mu}_i)$, $i = 1, ..., N$, coincides with the true target pdf ($\boldsymbol{\mu}_i \in \mathcal{X} \subseteq \mathbb{R}^n$). At the τ-iteration, we consider *the population* of samples

$$\mathcal{P}_\tau = \{\boldsymbol{\mu}_{1,\tau}, ..., \boldsymbol{\mu}_{N,\tau}\}.$$

At each iteration, the underlying idea of SMH is to replace one "bad" sample in the population with a better one, according to certain suitable probabilities. The algorithm is designed so that, after a burn-in period τ_b, the elements in $\mathcal{P}_{\tau'}$ ($\tau' > \tau_b$) are distributed according to $\pi_g(\boldsymbol{\mu}_{1,\tau'}, \ldots, \boldsymbol{\mu}_{N,\tau'})$, i.e., $\boldsymbol{\mu}_{i,\tau'}$ are i.i.d. samples from $\pi(\mathbf{x})$. For $\tau = 1, ..., \mathcal{T}$, the SMH algorithm consists of the following steps:

1. Draw $\boldsymbol{\mu}_{0,\tau} \sim \varphi(\boldsymbol{\mu})$, where φ is another proposal density, chosen by the user, which could indeed be selected or adapted using the information obtained in the previous steps of MAMIS.
2. Choose a "bad" sample $\boldsymbol{\mu}_{k,\tau}$ from the population (i.e., $k \in \{1, ..., N\}$), according to a probability proportional to $\frac{\varphi(\boldsymbol{\mu}_{k,\tau})}{\pi(\boldsymbol{\mu}_{k,\tau})}$, which corresponds to the inverse of the importance sampling weights.
3. Accept the new population

$$\mathcal{P}_{\tau+1} = \{\boldsymbol{\mu}_{1,\tau+1} = \boldsymbol{\mu}_{1,\tau}, ..., \boldsymbol{\mu}_{k,\tau+1} = \boldsymbol{\mu}_{0,\tau}, ..., \boldsymbol{\mu}_{N,\tau+1} = \boldsymbol{\mu}_{N,\tau}\},$$

with probability

$$\alpha(\boldsymbol{\mu}_{1,\tau}, ..., \boldsymbol{\mu}_{N,\tau}, \boldsymbol{\mu}_{0,\tau}) = \frac{\sum_{i=1}^{N} \frac{\varphi(\boldsymbol{\mu}_{i,\tau})}{\pi(\boldsymbol{\mu}_{i,\tau})}}{\sum_{i=0}^{N} \frac{\varphi(\boldsymbol{\mu}_{i,\tau})}{\pi(\boldsymbol{\mu}_{i,\tau})} - \min\limits_{0 \le i \le N} \frac{\varphi(\boldsymbol{\mu}_{i,\tau})}{\pi(\boldsymbol{\mu}_{i,\tau})}}. \tag{8.11}$$

Otherwise, set $\mathcal{P}_{\tau+1} = \mathcal{P}_\tau$.

4. If $\tau < \mathcal{T}$, set $\tau = \tau + 1$ and repeat from step 1.

Observe that the difference between $\mathcal{P}_\tau$ and $\mathcal{P}_{\tau+1}$ is at most one sample. Note also that $0 \leq \alpha(\boldsymbol{\mu}_{1,\tau}, ..., \boldsymbol{\mu}_{N,\tau}, \boldsymbol{\mu}_{0,\tau}) \leq 1$. Indeed, α depends on the proposed new point $\boldsymbol{\mu}_{0,\tau}$ and the entire population $\boldsymbol{\mu}_{i,\tau}$ $(i = 1, \ldots, N)$, but does not care about which mean $\boldsymbol{\mu}_{k,\tau}$ has been selected for a possible replacement.

An advantage of the SMH technique is that, when the chain has converged, the N samples in the population are independently distributed according to the target. The ergodicity can be proved using the detailed balance condition and considering the extended target pdf [8]. Moreover, for $N = 1$, it is possible to show that SMH becomes the standard MH method with an independent proposal pdf.

It is important to note that the positions of the Gaussians will hardly ever change when the parameters of the SMH proposal are not properly chosen, since the new points will never be accepted. Thus, the performance is never worsened with respect to the standard (fixed) multiple IS framework. Moreover, no diversity in the cloud is lost. In the worst case, we simply waste computational power by performing the MCMC operations. Another interesting point is that only one new importance weight needs to be evaluated at each iteration, since the rest of weights have already been computed in the previous steps (with the exception of the initial iteration, where all the weights need to be computed). Finally, note that the parameters of the proposal $\varphi(\boldsymbol{\mu})$ for the SMH algorithm could be also adapted using one of the many strategies already proposed in the literature [6, 10]. Moreover, in our framework, the adaptation can be improved by using the estimators $\hat{I}_t$ built by MAMIS to update the parameters of the proposal (not just using on-line the samples generated by the MCMC technique).

8.5 Numerical Simulations

8.5.1 Mixture of Gaussians

First of all, we consider a bivariate multimodal target pdf, which is itself a mixture of 5 Gaussians, i.e.,

$$\pi(\mathbf{x}) = \frac{1}{5} \sum_{i=1}^{5} \mathcal{N}(\mathbf{x}; \boldsymbol{\nu}_i, \boldsymbol{\Sigma}_i), \quad \mathbf{x} \in \mathbb{R}^2, \tag{8.12}$$

with means $\boldsymbol{\nu}_1 = [-10, -10]^\top$, $\boldsymbol{\nu}_2 = [0, 16]^\top$, $\boldsymbol{\nu}_3 = [13, 8]^\top$, $\boldsymbol{\nu}_4 = [-9, 7]^\top$, $\boldsymbol{\nu}_5 = [14, -14]^\top$, and covariance matrices $\boldsymbol{\Sigma}_1 = [2,\ 0.6; 0.6,\ 1]$, $\boldsymbol{\Sigma}_2 = [2,\ -0.4; -0.4,\ 2]$, $\boldsymbol{\Sigma}_3 = [2,\ 0.8; 0.8,\ 2]$, $\boldsymbol{\Sigma}_4 = [3,\ 0; 0,\ 0.5]$, and $\boldsymbol{\Sigma}_5 = [2,\ -0.1; -0.1,\ 2]$.

MAMIS Algorithm We apply MAMIS with $N = 100$ Gaussian proposals to estimate the mean (true value $[1.6, 1.4]^\top$) and normalizing constant (true value 1) of the target. We choose deliberately a "bad" initialization of the initial means to test

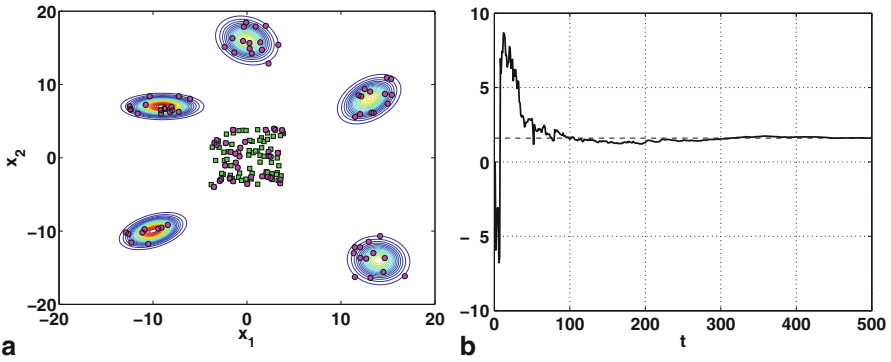

Fig. 8.1 **a** Contour plot of the target $\pi(\mathbf{x})$, the initial $\boldsymbol{\mu}_i^{(0)}$ (green squares) and the final $\boldsymbol{\mu}_i^{(T)}$ (magenta circles) locations of the means of the proposals q_i for a single run of MAMIS with $\lambda = 10$ ($N = 100$, $T = 2000$). **b** Evolution of the estimation of the mean (first comp.) as a function of $t = 0, \ldots, 500$ in one run of MAMIS (with $\lambda = 10$ and $\sigma = 2$). The dashed line depicts the true value (1.6)

the robustness of the algorithm and its ability to improve the corresponding *static* (i.e., nonadaptive) IS approach. Specifically, the initial means are selected uniformly within a square, i.e., $\boldsymbol{\mu}_i^{(0)} \sim \mathcal{U}([-4,4] \times [-4,4])$ for $i = 1, \ldots, N$. A single realization of $\boldsymbol{\mu}_i^{(0)}$ is depicted by the squares in Fig. 8.1a. We recall that we consider proposals

$$q_i^{(m)}(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_i^{(m)}, \mathbf{C}_i), \qquad i = 1, \ldots, N, \quad m = 0, \ldots, M. \tag{8.13}$$

In this example, we use identical isotropic covariance matrices, $\mathbf{C}_i = \sigma^2 \mathbf{I}_2$ with $\mathbf{I}_2$ denoting the 2×2 identity matrix, for every proposal. We test different values of $\sigma \in \{0.5, 1, 2, 10, 70\}$, to gauge the performance of MAMIS. Moreover, we also consider a Gaussian pdf

$$\varphi(\mathbf{x}) = \mathcal{N}(\mathbf{x}; [0,0]^\top, \boldsymbol{\Lambda}), \tag{8.14}$$

with $\boldsymbol{\Lambda} = \lambda^2 \mathbf{I}_2$ and $\lambda \in \{5, 10, 70\}$, as the proposal for the SMH algorithm. We set $T = 2000$ and $T_a \in \{2, 20, 100\}$, i.e., $M = \frac{T}{T_a} \in \{20, 100, 1000\}$.

We also consider $M = 1$ (i.e., $T = T_a$), which corresponds to a standard IS technique with multiple proposals and no adaptation. To maintain a constant computational cost in each simulation, we fix $\mathcal{T} = T_a = \frac{T}{M}$ (the number of iterations of SMH, at the end of each epoch), i.e., the total number of iterations of SMH in the entire MAMIS method is alway $M\mathcal{T} = T$.

All the results are averaged over 3000 independent experiments. Table 8.1 shows the mean absolute error (MAE) in the estimation of the first component of the mean; MAMIS always outperforms the nonadaptive standard IS procedure, with the only exception of $\sigma = 10$, where MAMIS has a negligibly larger error. The MAE in the estimation of the normalizing constant is shown in Table 8.2; in this case, the improvement provided by MAMIS is even more evident. In both cases, the best

Table 8.1 Mean absolute error (MAE) in the estimation of the mean (first component) of the mixture of Gaussians target, using the MAMIS algorithm ($N = 100$)

Std of $\varphi(\boldsymbol{\mu})$	Epochs	Std of $q_i(\mathbf{x})$					
		$\sigma = 0.5$	$\sigma = 1$	$\sigma = 2$	$\sigma = 5$	$\sigma = 10$	$\sigma = 70$
Standard multiple IS	$M = 1$ ($T_a = T$)	5.3566	6.8373	8.3148	0.3926	**0.0886**	0.3376
$\lambda = 5$	$M = 20$ ($T_a = 100$)	4.1945	2.4787	0.9055	0.2356	0.1051	**0.3349**
	$M = 100$ ($T_a = 20$)	4.2037	2.5032	0.8844	0.2372	0.1043	0.3380
	$M = 1000$ ($T_a = 2$)	4.2532	2.4526	0.8429	0.2315	0.1052	0.3376
$\lambda = 10$	$M = 20$ ($T_a = 100$)	0.5775	**0.1767**	0.1666	0.1052	0.0924	0.3502
	$M = 100$ ($T_a = 20$)	0.5839	0.2370	0.1245	0.0751	0.0917	0.3367
	$M = 1000$ ($T_a = 2$)	**0.5755**	0.2224	**0.1062**	**0.0698**	0.0932	0.3371
$\lambda = 70$	$M = 20$ ($T_a = 100$)	3.1817	1.6067	0.6966	0.1441	0.0926	0.3384
	$M = 100$ ($T_a = 20$)	3.0451	1.5679	0.6577	0.1372	0.0917	0.3354
	$M = 1000$ ($T_a = 2$)	3.1425	1.5550	0.6266	0.1303	0.0892	0.3381

Table 8.2 Mean absolute error (MAE) in the estimation of the normalizing constant of the mixture of Gaussians target, using the MAMIS algorithm ($N = 100$)

Std of $\varphi(\boldsymbol{\mu})$	Epochs	Std of $q_i(\mathbf{x})$					
		$\sigma = 0.5$	$\sigma = 1$	$\sigma = 2$	$\sigma = 5$	$\sigma = 10$	$\sigma = 70$
Standard multiple IS	$M = 1$ ($T_a = T$)	266637.6	345.3	3.2584	0.0322	0.0084	0.0322
$\lambda = 5$	$M = 20$ ($T_a = 100$)	0.8426	0.3166	0.0967	0.0193	0.0090	0.0321
	$M = 100$ ($T_a = 20$)	0.8027	0.2827	0.0801	0.0191	0.0091	**0.0313**
	$M = 1000$ ($T_a = 2$)	0.7825	0.2756	0.0740	0.0187	0.0087	0.0324
$\lambda = 10$	$M = 20$ ($T_a = 100$)	0.1252	0.0609	0.0453	0.0089	0.0086	0.0324
	$M = 100$ ($T_a = 20$)	0.1005	0.0360	0.0163	0.0067	0.0083	0.0321
	$M = 1000$ ($T_a = 2$)	**0.0965**	**0.0309**	**0.0114**	**0.0063**	**0.0082**	0.0321
$\lambda = 70$	$M = 20$ ($T_a = 100$)	1.9450	0.4147	0.1116	0.0130	**0.0082**	0.0321
	$M = 100$ ($T_a = 20$)	1.8540	0.3881	0.0958	0.0120	0.0083	0.0320
	$M = 1000$ ($T_a = 2$)	1.7933	0.3802	0.0899	0.0120	0.0083	0.0316

results for each column are shown in bold-face. Note also that the MAE in both cases seems to be almost independent from the value of M (i.e., the number of epochs).

Figure 8.1a also depicts the final locations of the means, $\boldsymbol{\mu}_i^{(T)}$, in one run (with $\lambda = 10$ and $M = 20$ epochs) using circles. Figure 8.1b illustrates the estimation of the mean (first component) as function of the iterations t, in a specific run of MAMIS (with $\lambda = 10$ and $\sigma = 2$). We also compare the performance of MAMIS with a PMC scheme described below.

Population Monte Carlo Scheme In this example, we also apply the mixture PMC scheme proposed in [2], with the same initialization used above for MAMIS. More precisely, we consider a population of samples

$$\{\mathbf{x}_1^{(t)}, \ldots, \mathbf{x}_N^{(t)}\},$$

at the t-th iteration, and propagate them using random walks

$$\mathbf{x}_i^{(t+1)} = \mathbf{x}_i^{(t)} + \epsilon_t, \quad i = 1, \ldots, N,$$

where $\epsilon_t \sim \mathcal{N}(\mathbf{x}; [0,0]^\top, \boldsymbol{\Phi})$, with $\boldsymbol{\Phi} = \phi \mathbf{I}_2$ and $\phi \in \{2, 5, 10, 20, 70\}$.

At each iteration, the resampling step is performed according to the normalized importance weights. The cumulative mean of the cloud $\{x_i^{(t)}\}_{i=1,t=1}^{N,T}$, as well as the cumulative estimate of the normalizing constant, are computed until $T = 2000$. We have not been able to apply the adaptive strategy suggested in [2] in order to select suitable scale parameters, within a population of prechosen values, since it has been difficult to select these values adequately. More specifically, we have not been able to find a set of parameters for this approach that provides reasonable results. It is important to note that the parameter ϕ in PMC plays the role of both λ and σ in MAMIS: it is at the same time a perturbation parameter (like λ in MAMIS) and an IS parameter (like σ in MAMIS).

The corresponding results for the PMC scheme are shown in Table 8.3 for different values of N. We can observe that MAMIS obtains, in general, better results and appears more robust with respect to the variations of the parameters. Indeed, to attain the same performance as MAMIS, PMC needs more computational effort.[1]

[1] A fair comparison of the computational cost of MAMIS and PMC deserves a further discussion. On the one hand, in MAMIS we use NT samples for the estimation and (setting $\mathcal{T} = T_a$, as in this example) we also perform T iterations of the SMH technique, which requires drawing T proposed samples, T samples from multinomial pdfs, and T uniform random variables (RVs; thus, we generate $3T$ additional RVs with respect to a simple iterative IS scheme). Moreover, the target needs to be evaluated at T new points. On the other hand, in PMC we use NT samples in the estimation and we also need to draw NT samples from multinomial densities (resampling steps), thus requiring NT additional RVs with respect to a simple iterative IS scheme.

Table 8.3 Mean absolute error (MAE) in the estimation of the mean of the mixture of Gaussians target (first component) and the normalizing constant, using the PMC algorithm [2] (with $T = 2000$)

Standard PMC-MAE (mean, first comp.)					
N	$\phi = 2$	$\phi = 5$	$\phi = 10$	$\phi = 20$	$\phi = 70$
100	6.2113	**1.3365**	0.1891	0.6359	1.5374
500	5.4231	1.9937	0.0921	0.1437	0.9275
2000	**4.9097**	1.9079	**0.0600**	**0.0433**	**0.3211**
Standard PMC-MAE (normalizing const.)					
N	$\phi = 2$	$\phi = 5$	$\phi = 10$	$\phi = 20$	$\phi = 70$
100	3.7526	0.6034	0.0308	0.0129	0.0320
500	3.5593	0.4624	0.0137	0.0057	0.0145
2000	**3.3276**	**0.3943**	**0.0069**	**0.0028**	**0.0071**

8.5.2 *Banana-Shaped Target Density*

We also test MAMIS with another kind of target pdf, the well-known "banana-shaped" benchmark distribution [6], which can be expressed mathematically as

$$\pi(x_1, x_2) \propto \exp\left(-\frac{1}{2\eta_1^2}(4 - Bx_1 - x_2^2)^2 - \frac{x_1^2}{2\eta_2^2} - \frac{x_2^2}{2\eta_3^2}\right),$$

with $B = 10$, $\eta_1 = 4$, $\eta_2 = 5$, and $\eta_3 = 5$.

We consider again Gaussian proposals, $q_i^{(m)}(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_i^{(m)}, \sigma^2 \mathbf{I}_2)$ with $\mathbf{x} = [x_1, x_2]^\top$, $i = 1, \ldots, N$ and $m = 1, \ldots, M$, for the multiple IS and a Gaussian proposal, $\varphi(\mathbf{x}) = \mathcal{N}(\mathbf{x}; [0, 0]^\top, \lambda^2 \mathbf{I}_2)$, for the SMH algorithm. We choose again deliberately an "inadequate" initialization, $\mu_i^{(0)} \sim \mathcal{U}([-6, -5] \times [-6, -5])$. We use different scale parameters, $\sigma \in \{0.1, 0.5, 1, 5, 10\}$ and $\lambda \in \{5, 70\}$, and also set $N = 100$, $T = 2000$ and $T_a = 10$, i.e., $M = \frac{T}{T_a} = 200$. In order to maintain the computational cost constant in each simulation, we set $\mathcal{T} = T_a = \frac{T}{M}$. We again consider $M = 1$ (i.e., $T = T_a$), which corresponds to a standard IS technique with multiple proposals and no adaptation.

Table 8.4 shows the MAE in the estimation of the first component of the mean of the target (true value ≈ -0.4845). The true value has been computed (in an approximate way) with a standard deterministic numerical method (using a thin grid in the parameter space). The results are averaged over 3000 independent experiments. Table 8.4 also contains the MAE of the PMC scheme, described previously, for different values of the parameters N and ϕ. As in the previous example, it seems that PMC needs more computational effort to obtain the same performance as MAMIS. Figure 8.2 shows the initial and final locations of the means, i.e., $\boldsymbol{\mu}_i^{(0)}$ and $\boldsymbol{\mu}_i^{(T)}$, in a specific run with $\lambda \in \{5, 10\}$.

Table 8.4 Mean absolute error (MAE) in the estimation of the mean of the banana-shaped target (first component), using the MAMIS algorithm ($N = 100$, $T = 2000$) and PMC (with different values of N and $T = 2000$)

Std of $\varphi(\boldsymbol{\mu})$	Epochs	Std of $q_i(\mathbf{x})$				
		$\sigma = 0.1$	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 5$	$\sigma = 10$
Standard multiple IS	$T_a = T$	4.3873	3.3111	1.6394	0.0111	0.0111
$\lambda = 5$	$T_a = 10$	**0.1498**	**0.0207**	**0.0080**	**0.0101**	**0.0101**
$\lambda = 70$	$T_a = 10$	2.0725	1.1833	0.4709	0.0112	0.0112

Standard PMC [2]

N	$\phi = 0.5$	$\phi = 1$	$\phi = 2$	$\phi = 5$	$\phi = 10$	$\phi = 20$	$\phi = 70$
100	0.5924	0.4195	0.1804	0.0416	0.2923	1.6628	14.1891
500	0.5782	0.3862	0.1386	0.0070	0.0346	0.2008	3.9331
2000	**0.5630**	**0.3594**	**0.1061**	**0.0021**	**0.0079**	**0.0343**	**0.9240**

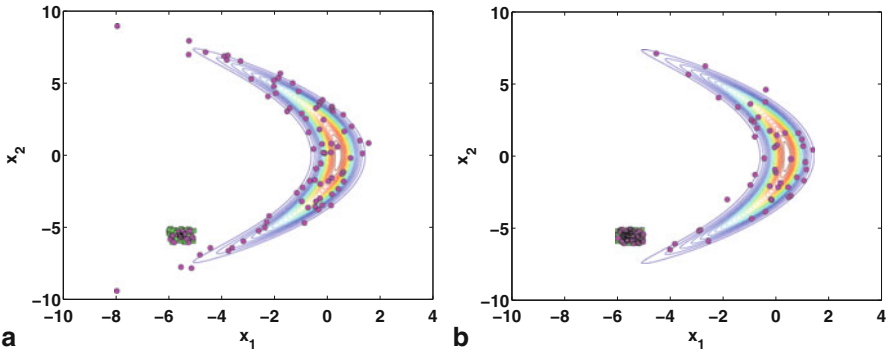

Fig. 8.2 Contour plot of the target $\pi(\mathbf{x})$, the initial $\boldsymbol{\mu}_i^{(0)}$ (green squares) and the final $\boldsymbol{\mu}_i^{(T)}$ (magenta circles) locations of the means of the proposals q_i for a single run of MAMIS ($N = 100$, $T = 2000$); **(a)** with $\lambda = 5$; **(b)** with $\lambda = 10$. With $\lambda = 5$, the number of Gaussians q_i which change their location is larger than with $\lambda = 10$

8.6 Conclusions

We have introduced a novel algorithm, *MAMIS*, which applies the iterative importance sampling (IS) approach using a population of adaptive proposal pdfs. The location parameters of the proposals are adapted according to an MCMC technique.

Although the MAMIS scheme is more general, here we have focused on a specific implementation with a cloud of Gaussian proposal pdfs, adapting their means. Our experiments have shown that MAMIS reduces the dependence on the choice of the different parameters of the proposals.

Indeed, the proposed adaptation procedure improves the results with respect to the corresponding standard nonadaptive IS method, regardless of the variances chosen

initially. We have also compared with respect to a population Monte Carlo (PMC) scheme. MAMIS seems to be more flexible and more robust with respect to the choice of the initial conditions.

In MAMIS, the adaptation does not use resampling procedures, so no diversity in the population is lost (as it happens in PMC). The scale parameters of the proposals, which is crucial for the good performance of an IS algorithm, are not adapted in the multiple IS scheme. In this way, we avoid losing diversity in the population of scale parameters (if different variances are used, as suggested in the black-box implementation), thus maintaining simultaneously explorative and local search behaviors at each iteration (corresponding to large and small variances, respectively).

Acknowledgements This work has been supported by the Spanish government's projects COMONSENS (CSD2008-00010), ALCIT (TEC2012-38800-C03-01), DISSECT (TEC2012-38058-C03-01), OTOSiS (TEC2013-41718-R), and COMPREHENSION (TEC2012-38883-C02-01), as well as by the ERC grant 239784 and AoF grant 251170.

References

1. Andrieu, C., de Freitas, N., Doucet, A., Jordan, M.: An introduction to MCMC for machine learning. Mach. Learn. **50**, 5–43 (2003)
2. Cappé, O., Guillin, A., Marin, J.M., Robert, C.P.: Population Monte Carlo. J. Comput. Graph. Stat. **13**(4), 907–929 (2004)
3. Cornuet, J.M., Marin, J.M., Mira, A., Robert, C.P.: Adaptive multiple importance sampling. Scand. J. Stat. **39**(4), 798–812 (2012)
4. Doucet, A., Wang, X.: Monte Carlo methods for signal processing. IEEE Signal Process. Mag. **22**(6), 152–170 (2005)
5. Fitzgerald, W.J.: Markov chain Monte Carlo methods with applications to signal processing. Signal Process. **81**(1), 3–18 (2001)
6. Haario, H., Saksman, E., Tamminen, J.: An adaptive Metropolis algorithm. Bernoulli. **7**(2), 223–242 (2001)
7. Kotecha, J., Djurić, P.M.: Gibbs sampling approach for generation of truncated multivariate gaussian random variables. Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Phoenix, Arizona, USA (1999)
8. Liang, F., Liu, C., Caroll, R.: Advanced Markov Chain Monte Carlo Methods: Learning from Past Samples. Wiley Series in Computational Statistics, Wiley, USA (2010)
9. Liu, J.S.: Monte Carlo Strategies in Scientific Computing, Springer, Vancouver, Canada (2004)
10. Luengo, D., Martino, L.: Fully adaptive Gaussian mixture Metropolis-Hastings algorithm. Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Vancouver, Canada (2013)
11. Owen, A., Zhou, Y.: Safe and effective importance sampling. J. Am. Stat. Assoc. **95**(449), 135–143 (2000)
12. Robert, C.P., Casella, G.: Monte Carlo Statistical Methods. Springer, USA (2004)
13. Skilling, J.: Nested sampling for general Bayesian computation. Bayesian Anal. **1**(4), 833–860 (2006)
14. Veach, E., Guibas, L.: Optimally combining sampling techniques for Monte Carlo rendering. In SIGGRAPH 1995 Proceedings, Los Angeles, California, USA. pp. 419–428 (1995)
15. Wang, X., Chen, R., Liu, J.S.: Monte Carlo Bayesian signal processing for wireless communications. J. VLSI Signal Process. **30**, 89–105 (2002)

Chapter 9
Bayes Factors for Comparison of Restricted Simple Linear Regression Coefficients

Viviana Giampaoli, Carlos A. B. Pereira, Heleno Bolfarine and Julio M. Singer

Abstract This work compares two simple linear regression slopes that are restricted to an order constraint and to a proper subset of parameter space. Two approaches based on Bayes factors are discussed. The motivation is a practical example designed to evaluate dental plaque reduction. The results indicate that the approach that takes into account the restricted parameter space is more informative than the one with unrestricted parameter space since it allows to obtain more evidence against the null hypothesis.

9.1 Introduction

Classical statistical inference under constrained parametric spaces has been addressed by many authors, among which we mention [25] and [5]. The problem of comparing means of Gaussian distributions with restricted parameter space was considered in [16]. Giampaoli and Singer [17] worked in the same problem under a Bayesian approach, where the restricted parameter space can be handled with less effort. Here, we extend the results to compare slopes that belong to the interval [0,1]. The motivation is a study involving two types of toothbrush with respect to their efficacy in removing dental plaque. The data, listed in Table 9.1, correspond to dental plaque indices measured on 16 preschool children before and after toothbrushing. Eight children used tootbrush A and another eight children used toothbrush B.

V. Giampaoli (✉) · C. A. B. Pereira · H. Bolfarine · J. M. Singer
Instituto de Matemática e Estatística, Universidade de São Paulo, Rua do Matão 1010, Cidade Universitária - São Paulo, SP 05508-090, Brazil
e-mail: vivig@ime.usp.br

C. A. B. Pereira
e-mail: cpereira@ime.usp.br

H. Bolfarine
e-mail: hbolfar@ime.usp.br

J. M. Singer
e-mail: jmsinger@ime.usp.br

A. Polpo et al. (eds.), *Interdisciplinary Bayesian Statistics,*
Springer Proceedings in Mathematics & Statistics 118, DOI 10.1007/978-3-319-12454-4_9

Table 9.1 Dental plaque index

Toothbrush A		Toothbrush B	
Before toothbrushing	After toothbrushing	Before toothbrushing	After toothbrushing
0.62	0.57	0.45	0.40
0.70	0.64	1.15	0.65
0.90	0.80	1.27	1.22
0.95	0.92	0.87	0.80
1.20	1.17	0.82	0.77
0.10	0.10	0.97	0.95
0.57	0.55	0.50	0.47
1.35	1.32	0.73	0.35

Similar pretreatment/posttreatment data was analyzed in [29] with the following model:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon. \tag{9.1}$$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \mathbf{y}_1 = \begin{pmatrix} y_{11} \\ \vdots \\ y_{1n_1} \end{pmatrix}, \mathbf{y}_2 = \begin{pmatrix} y_{21} \\ \vdots \\ y_{2n_2} \end{pmatrix},$$

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{x}_2 \end{pmatrix}, \mathbf{x}_1 = \begin{pmatrix} x_{11} \\ \vdots \\ x_{1n_1} \end{pmatrix}, \mathbf{x}_2 = \begin{pmatrix} x_{21} \\ \vdots \\ x_{2n_2} \end{pmatrix},$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \boldsymbol{\varepsilon} = \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \end{pmatrix}, \boldsymbol{\varepsilon}_1 = \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1n_1} \end{pmatrix} and \boldsymbol{\varepsilon}_2 = \begin{pmatrix} \varepsilon_{21} \\ \vdots \\ \varepsilon_{2n_2} \end{pmatrix},$$

where n_1 (n_2) is the sample size for treatment A (B), y_{1i} (y_{2i}) is the posttreatment measurement of the ith individual submitted to treatment A (B), x_{1i} (x_{2i}) is the pretreatment measurement on the ith individual submitted to treatment A (B), β_1 (β_2) is the slope parameter (a dental plaque reduction index) for treatment A (B) and ε_{1i} (ε_{2i}) is the measurement on the ith individual submitted to treatment A (B). The standard assumption that the random error is normally distributed with null mean vector and covariance matrix $\sigma^2\mathbf{I}_k$, σ^2 being a positive scalar and $\mathbf{I}_k$ denoting an identity matrix of the order $k = (n_1 + n_2) \times (n_1 + n_2)$ is also considered here. Another assumption is that $\beta_i \in [0, 1]$, $i = 1, 2$, since one should expect that the dental plaque index decreases after toothbrusing.

The objective of the study may be assessed via a test of

$$H_1 : \beta_1 = \beta_2 = \beta, \tag{9.2}$$

versus

$$H_2 : \beta_1 > \beta_2, \tag{9.3}$$

under the restriction that $0 \leq \beta_i \leq 1$.

Aoki et al. ([3] and [4]) consider a Bayesian analysis of repeated pretest/posttest data under null intercept errors-in-variables regression models. The proposed Bayesian approach accommodates the correlated measurements and incorporates the restriction that the slopes must lie in the [0,1] interval. Barros et al. [6] used Wald type statistics (see, for example, [28]) for testing (9.2) versus (9.3) under measurement error models, also incorporating the additional assumption that the slopes lie in a bounded interval. The present chapter discusses an alternative approach also using Bayes factors. In this direction, we note that testing the hypotheses (9.2) versus (9.3) is equivalent to comparing the following two models:

$$M = 1 : y_{ik} = \beta x_{ik} + \varepsilon_{ik}, \quad \varepsilon_{ik} \overset{IID}{\sim} N\left(0, \sigma^2\right), \tag{9.4}$$

$$M = 2 : y_{ik} = \beta_i x_{ik} + \varepsilon_{ik}, \quad \varepsilon_{ik} \overset{IID}{\sim} N\left(0, \sigma^2\right), \tag{9.5}$$

$i = 1, 2, k = 1, \ldots n_i$, with $\sigma^2 > 0$, where σ^2 is a nuisance parameter for the testing problem and M is an integer-valued parameter indexing the models, such that $M = 1$ corresponds to H_1 and $M = 2$ corresponds to H_2.

This chapter is organized as follows: in Sect 9.2 we discuss the computation of Bayes factors. In Sect 9.3 we analyze the data in Table 9.1. Finally, in Sect 9.4 we present a brief discussion.

9.2 The Bayes Factor

The Bayes factor is the ratio of predictive densities derived from the two models and may be used for deciding in favor or against each hypothesis. It is defined by

$$B_{21} = \frac{p(\mathbf{y}|M = 2)}{p(\mathbf{y}|M = 1)} = \frac{p(M = 2|\mathbf{y})}{p(M = 1|\mathbf{y})} \frac{\pi_1}{\pi_2}. \tag{9.6}$$

where $p(\mathbf{y}|M = j)$ and $p(M = j|\mathbf{y})$, $j = 1, 2$ denote the predictive (or marginal) density and the posterior probability of model M_j, respectively. $\pi_1 = p(M = 1)$ is the prior model probability for M_1 and $\pi_2 = p(M = 2) = 1 - \pi_1$ is the prior model probability for M_2. Kass and Raftery [20] discuss the use of these factors for hypothesis testing and suggest a rule for deciding for one of the two alternative models.

Although the Bayes factors constitutes an important tool for statistical analysis, its use is not free of controversy. First, it may be severely affected by small variations in the choice of the prior distribution. Also, it lacks interpretation when improper prior distributions are used, as pointed by O'Hagan [24]. Variations commonly employed for model comparison include pseudo-Bayes factors as advocated by Gelfand and Dey [13], posterior Bayes factors, as suggested by Aitkin [1] or criteria that minimize some posterior loss functions as proposed by Gelfand and Ghosh [14]. The intrinsic Bayes factor suggested by Berger and Pericchi [7], with motivations given by Berger and Pericchi [8], Moreno and Liseo [21], Berger and Pericchi [9], and the fractional Bayes factors, discussed by O'Hagan [24] and De Santis [12] are possible alternatives. As suggested by Moreno et al. [22], these techniques, however, use training samples and are unstable when the sample size is small as in the example described previously. The reader is referred to [18] or [23] for a recent review on this topic. Robert and Marin [27] present a detailed analysis of the difficulties associated to some Markov Chain Monte Carlo (MCMC) based techniques. These authors compare such alternatives with the one proposed by Carlin and Chib [11]. We consider two different techniques to obtain the posterior probability associated with each model, namely, $p(M = j|\mathbf{y})$, $j = 1, 2$. The first, proposed by Irony and Pereira [19], provides an analytic expression for the Bayes factor obtained by a direct computation of the posterior distribution; it accommodates any proper prior distribution and does not require Gibbs sampling, since usual methods of numerical integration may be employed. The second, proposed by Carlin and Chib [11], is based on the Gibbs sampler and on MCMC methods for the computations. It requires either the use of conjugate prior distributions or the existence of conditional complete posterior densities. Unfortunately, a direct comparison of the results obtained by the two techniques is quite complicated, since the first technique relies on prior distributions specified for the entire parameter space, while the second requires different prior distributions under the null and the alternative hypotheses.

9.2.1 Computation of Bayes Factor via Predictive Distributions

Note that the null (9.2) and the alternative (9.3) hypotheses may be written as

$$H_j : \boldsymbol{\beta} \in \Omega_j, j = 1, 2, \tag{9.7}$$

with

$$\Omega_1 = \{\boldsymbol{\beta} = (\beta_1, \beta_2) : \beta_1 = \beta_2, 0 \leq \beta_i \leq 1, i = 1, 2\},$$
$$\Omega_2 = \{\boldsymbol{\beta} = (\beta_1, \beta_2) : \beta_1 > \beta_2, 0 \leq \beta_i \leq 1, i = 1, 2\}.$$

The prior density for $\boldsymbol{\beta}$ under $H_j : j = 1, 2$, is defined as

$$g(\boldsymbol{\beta}|M = j) = \frac{g_j(\boldsymbol{\beta})}{\int_{\Omega_j} g_j(\boldsymbol{\beta}) d\boldsymbol{\beta}}, \tag{9.8}$$

where g is a convenient prior density function and g_j denotes the function g restricted to the set Ω_j, i.e., the function with domain Ω_j such that $g_j(\boldsymbol{\beta}) = g(\boldsymbol{\beta})$ if $\boldsymbol{\beta} \in \Omega_j$. Note that the integral $\int_{\Omega_j} g_j d\Omega_j$ represents the volume under the prior density function $g(\boldsymbol{\beta})$ in Ω_j. When the integral is null, as in the case where the dimension of Ω_j is less than the dimension of the parameter space, we consider the corresponding line integral. Naturally, the prior density (9.8) is well defined if $0 < \int_{\Omega_j} g d\Omega_j < \infty$. Under (9.7), the predictive densities are

$$p(\mathbf{y}|M = j) = \frac{\int_{\Omega_j} g_j(\boldsymbol{\beta}) l(\mathbf{y}|\boldsymbol{\beta}) d\boldsymbol{\beta}}{\int_{\Omega_j} g_j(\boldsymbol{\beta}) d\boldsymbol{\beta}}, \tag{9.9}$$

where $l(\mathbf{y}|\boldsymbol{\beta})$ is the likelihood function, so that the posterior probabilities $p(M = j|\mathbf{y})$, $j = 1, 2$, may be obtained from

$$p(M = j|\mathbf{y}) = \frac{\pi_j p(\mathbf{y}|M = j)}{\pi_1 p(\mathbf{y}|M = 1) + \pi_2 p(\mathbf{y}|M = 2)}.$$

9.2.2 Computation of Bayes Factors via MCMC Methods

We now describe an alternative way of computing $p(M = j|\mathbf{y})$, $j = 1, 2$, via the MCMC algorithm proposed by Carlin and Chib [11]. Essentially, they consider M as a component of the random vector $\boldsymbol{\upsilon} = (\beta, (\beta_1, \beta_2), M)$. Hence, it can be sampled via Gibbs methods. After convergence, estimates of $p(M = j|\mathbf{y})$, $j = 1, 2$, may be obtained as the ratios between the number of iterations for which $M = j$ and the total number of iterations. Direct sampling of the marginal distributions themselves or of the joint distribution $p(\boldsymbol{\upsilon}, \mathbf{y})$ is complicated, but sampling the full conditional posterior distributions $p(\beta|(\beta_1, \beta_2), M, \mathbf{y})$, $p((\beta_1, \beta_2)|\beta, M, \mathbf{y})$, and $p(M|\beta, (\beta_1, \beta_2), \mathbf{y})$ is straightforward using the Gibbs sampler. To satisfy the MCMC convergence conditions, we need to specify a full probability model. For such purposes we assume that β and (β_1, β_2) are independent given the model indicator M and that the prior distributions $p(\beta|M = j)$ and $p((\beta_1, \beta_2)|M = j)$, $j = 1, 2$ are proper. Thus,

$$p((\beta, \beta_1, \beta_2)|M = j) = p(\beta|M = j) p(\beta_1, \beta_2|M = j). \tag{9.10}$$

We also assume that $\mathbf{y}$ is independent of (β_1, β_2) given $M = 1$ and of β, given $M = 2$. We complete the Bayesian model specification by choosing proper "pseudoprior" distributions $p(\beta|M = 2)$ and $p(\beta_1, \beta_2|M = 1)$. Because of the conditional independence assumptions, these "pseudoprior" distributions do not interfere with the marginal densities and so their form is irrelevant. The joint distribution of $\mathbf{y}$ and $(\beta, (\beta_1, \beta_2))$ given $M = j$ is

$$p(\mathbf{y}, \beta, (\beta_1, \beta_2)|M = j) = f(\mathbf{y}|\beta, (\beta_1, \beta_2), M = j)\, p(\beta, (\beta_1, \beta_2)|M = j),$$

where $f(\mathbf{y}|\beta, (\beta_1, \beta_2), M = j)$ is a density function corresponding to $f(\mathbf{y}|\beta, M = 1)\pi_1$ for $j = 1$, and $f(\mathbf{y}|(\beta_1, \beta_2), M = 2)\pi_2$ for $j = 2$. To implement the Gibbs

sampler we need the full conditional distributions of β and (β_1, β_2) as well as that of M. Observe that

$$p\left(\beta|(\beta_1,\beta_2),M,\mathbf{y}\right) \propto \begin{cases} f\left(\mathbf{y}|\beta,M=1\right)p\left(\beta|M=1\right), & \text{if } M=1, \\ p\left(\beta|M=2\right), & \text{if } M=2 \end{cases} \tag{9.11}$$

and

$$p\left((\beta_1,\beta_2)|\beta,M,\mathbf{y}\right) \propto \begin{cases} f\left(\mathbf{y}|(\beta_1,\beta_2),M=2\right)p\left((\beta_1,\beta_2)|M=2\right), & \text{if } M=2, \\ p\left((\beta_1,\beta_2)|M=1\right), & \text{if } M=1. \end{cases} \tag{9.12}$$

When $M = 1$, the required conditional distribution for β (9.11) is generated from model 1; otherwise, when $M = 2$, the distribution is generated from the corresponding "pseudoprior" distribution. Similarly, when $M = 1$, the required conditional distribution for (β_1, β_2) (9.12) is generated from the corresponding "pseudoprior" distribution; otherwise, the required distribution is generated from model $M = 2$:

$$p\left(M=1|\beta,(\beta_1,\beta_2),\mathbf{y}\right) = \frac{f\left(\mathbf{y}|\beta,M=1\right))\,p\left(\beta,(\beta_1,\beta_2)|M=1\right)\pi_1}{\sum_{k=1}^{2} p\left(\mathbf{y},\beta,(\beta_1,\beta_2),M=k\right)} \tag{9.13}$$

and

$$p\left(M=2|\beta,(\beta_1,\beta_2),\mathbf{y}\right) = \frac{f\left(\mathbf{y}|(\beta_1,\beta_2),M=2\right)\,p\left(\beta,(\beta_1,\beta_2)|M=2\right)\pi_2}{\sum_{k=1}^{2} p\left(\mathbf{y},\beta,(\beta_1,\beta_2),M=k\right)}. \tag{9.14}$$

Samples of the posterior distribution $p\left(\beta,(\beta_1,\beta_2),M|\mathbf{y}\right)$ may be obtained by sampling the full conditional distributions (9.11)–(9.12) and (9.13)–(9.14), via a Gibbs algorithm. In each iteration, $l = 1, \ldots, N$, a sample of size one, $(\beta^{(l)}, (\beta_1, \beta_2)^{(l)}, M^{(l)})$, is obtained. The ratio

$$\widehat{p}(M=j|\mathbf{y}) = \frac{\text{number of } M^{(l)} = j \text{ in the } N \text{ iterations}}{N}, j = 1, 2,$$

provides a simple estimate of $p(M = j|\mathbf{y})$ that may be used to compute the Bayes factor from (9.6).

Although the form of the "pseudoprior" distributions is theoretically arbitrary, it is convenient to have them close to the conditional densities $p\left(\beta|(\beta_1,\beta_2),M,\mathbf{y}\right)$ and $p\left((\beta_1,\beta_2)|\beta,M,\mathbf{y}\right)$ so that plausible values are generated, even when the assumed model is false. Carlin and Chib [11] recommend separate runs for each model, i.e., considering $\pi_1 = 1$ and $\pi_2 = 0$, and an approximation of the resulting posterior distribution by a "pseudoprior" distribution for the parameter β under model $M = 2$. Subsequently, considering $\pi_2 = 0$ and $\pi_1 = 1$, an approximation of the resulting posterior distribution may also be taken as a "pseudoprior" distribution for the parameters (β_1, β_2) under model $M = 1$.

9.3 Analysis of the Dental Plaque Index Data

9.3.1 Analysis via Predictive Distributions

The likelihood function under the model $M = 2$ is

$$l(\beta_1, \beta_2) = \frac{1}{\sigma^{n_1+n_2}} exp\left[-\frac{1}{2\sigma^2}\left(\sum_{k=1}^{n_1}(y_{1k} - \beta_1 x_{1k})^2 + \sum_{k=1}^{n_2}(y_{2k} - \beta_2 x_{2k})^2\right)\right], \tag{9.15}$$

which can be rewritten as

$$l(\beta_1, \beta_2) = \frac{\tau^{(n_1+n_2)/2}}{2\pi} \exp\left[\frac{-\tau}{2}\left(\sum_{i=1}^{2}\left(s_{i1} - \beta_i s_{i2} + \beta_i^2 s_{i3}\right)\right)\right]$$

$$\tau = \sigma^{-2}, \quad s_{i1} = \sum_{k=1}^{n_i} y_{ik}^2, \quad s_{i2} = \sum_{k=1}^{n_i} 2y_{ik}x_{ik}, \quad s_{i3} = \sum_{k=1}^{n_i} x_{ik}^2 \tag{9.16}$$

for $i = 1, 2$.

Initially, we assumed Beta prior distributions for β_1 and β_2; observe that these are not conjugate distributions for the problem under consideration. The density functions are

$$g(\beta_i) = \frac{1}{B(a_i, b_i)} \beta_i^{a_i - 1} (1 - \beta_i)^{b_i - 1},$$

where $B(a_i, b_i)$ is the Beta function with parameters a_i, b_i, $i = 1, 2$. The joint prior density can be written as

$$g(\beta_1, \beta_2) = C \prod_{i=1}^{2} \beta_i^{a_i - 1} (1 - \beta_i)^{b_i - 1},$$

with $C = [B(a_1, b_1)\, B(a_2, b_2)]^{-1}$.

The denominator of the ratio defining $p(\mathbf{y}|M = 1)$ in (9.9) is

$$\begin{aligned}\int g(\beta, \beta)\, d\overline{\Omega}_1 &= C \int_0^1 \beta^{(a_1+a_2-1)-1} (1 - \beta)^{(b_1+b_2-1)-1}\, d\beta \\ &= C\, B(a_1 + a_2 - 1, b_1 + b_2 - 1),\end{aligned}$$

with $\overline{\Omega}_1 = \{\beta : 0 \leq \beta \leq 1\}$. The corresponding numerator is

$$\int g(\beta, \beta)\, l(\beta, \beta)\, d\overline{\Omega}_1$$

$$= \Gamma\left(\frac{n_1+n_2}{2}+\nu_2\right) C \frac{\nu_2^{\nu_1}}{2\pi\,\Gamma(\nu_1)} \int_0^1 \frac{\beta^{A-1}(1-\beta)^{B-1}}{\left(\frac{1}{2}s(\beta,\beta)+\nu_2\right)^{\frac{n_1+n_2}{2}+\nu_2}} d\beta,$$

with $A = (a_1 + a_2 - 1)$, $B = (b_1 + b_2 - 1)$, $s(\beta_1, \beta_2) = \left(\sum_{i=1}^{2}\left(s_{i1} - \beta_i s_{i2} + \beta_i^2 s_{i3}\right)\right)$, in with s_{i1}, $s_{i2,}$ and s_{i3} defined in (9.16).

The denominator of (9.9) is

$$\int g(\beta_1, \beta_2)\, d\Omega_2$$
$$= C \int_0^1 \beta_1^{a_1-1}(1-\beta_1)^{b_1-1}\left(\int_0^{\beta_1} \beta_2^{a_2-1}(1-\beta_2)^{b_2-1}\, d\beta_2\right) d\beta_1$$

and the corresponding numerator is

$$\int g(\beta_1, \beta_2)\, l(\beta_1, \beta_2)\, d\Omega_2$$
$$= \Gamma\left(\frac{n_1+n_2}{2}+\nu_2\right) C \frac{\nu_2^{\nu_1}}{2\pi\,\Gamma(\nu_1)}$$
$$\int_0^1 \beta_1^{a_1-1}(1-\beta_1)^{b_1-1} \int_0^{\beta_1} \beta_2^{a_2-1}(1-\beta_2)^{b_2-1} \frac{d\beta_2}{\left(\frac{1}{2}s(\beta_1,\beta_2)+\nu_2\right)^{\frac{n_1+n_2}{2}+\nu_2}} d\beta_1.$$

If we consider the uniform prior distribution on [0,1], i.e., taking $a_1 = b_1 = 1$, and $a_2 = b_2 = 1$ and the distribution of τ with hyperparameters equal to their least squares estimators [$\nu_1 = 1$ and $\nu_2 = 0.0169$], we obtain $B_{21} = 0$ suggesting that the null hypotheses should not be rejected.

We assume now that β_1 and β_2 have prior truncated Gaussian distributions with parameters (ξ_1, γ_1) and (ξ_2, γ_2), respectively, i.e, with density functions given by

$$g^*(\beta_i) = \frac{1}{\Phi\left((1-\xi_i)\sqrt{\gamma_i}\right) - \Phi\left(-\xi_i\sqrt{\gamma_i}\right)} \frac{\sqrt{\gamma_i}}{\sqrt{2\pi}} \exp\left(\frac{-\gamma_i}{2}(\beta_i - \xi_i)^2\right),$$

for $i = 1, 2$. Let

$$J(\xi_i, \gamma_i) = \frac{\frac{1}{\sqrt{2\pi}}\left[\exp\left((-\gamma_i/2)\,\xi_i^2\right) - \exp\left((-\gamma_i/2)(1-\xi_i)^2\right)\right]}{\Phi\left((1-\xi_i)\sqrt{\gamma_i}\right) - \Phi\left(-\xi_i\sqrt{\gamma_i}\right)}; \quad \text{then}$$

it follows that the corresponding joint prior density is given by

$$g^*(\beta_1, \beta_2) = C_1 \prod_{i=1}^{2} \frac{\sqrt{\gamma_i}}{\sqrt{2\pi}} \exp\left(\frac{-\gamma_i}{2}(\beta_i - \xi_i)^2\right),$$

where $C_1 = \left[\prod_{i=1}^{2} \Phi\left((1-\xi_i)\sqrt{\gamma_i}\right) - \Phi\left(-\xi_i\sqrt{\gamma_i}\right)\right]^{-1}$. The denominator of the ratio that defines $p(\mathbf{y}|M = 1)$ is

$$\int g^*(\beta, \beta)\, d\overline{\Omega}_1$$

$$= C_1 \frac{\sqrt{\gamma_1 \gamma_2}}{2\pi} \int_0^1 \exp\left(-\left(\frac{\gamma_1}{2}(\beta - \xi_1)^2 + \frac{\gamma_2}{2}(\beta - \xi_2)^2\right)\right) d\beta$$

and the corresponding numerator is

$$\int g(\beta, \beta)\, l(\beta, \beta)\, d\overline{\Omega}_1$$
$$= \Gamma\left(\frac{n_1 + n_2}{2} + \nu_2\right) \frac{\nu_2^{\nu_1}}{2\pi \Gamma(\nu_1)} C_1 \frac{\sqrt{\gamma_1 \gamma_2}}{2\pi}$$
$$\int_0^1 \frac{\exp\left(-\left(\frac{\gamma_1}{2}(\beta - \xi_1)^2 + \frac{\gamma_2}{2}(\beta - \xi_2)^2\right)\right)}{\left(\frac{1}{2}s(\beta, \beta) + \nu_2\right)^{\frac{n_1+n_2}{2}+\nu_2}} d\beta.$$

The denominator of the ratio that defines $p(\mathbf{y}|M = 2)$ is

$$\int g^*(\beta_1, \beta_2)\, d\Omega_2 = C_1 \frac{\sqrt{\gamma_1 \gamma_2}}{2\pi}$$
$$\times \int_0^1 \exp\left(\frac{-\gamma_1}{2}(\beta_1 - \xi_1)^2\right) \left(\int_0^{\beta_1} \exp\left(\frac{-\gamma_2}{2}(\beta_2 - \xi_2)^2\right) d\beta_2\right) d\beta_1,$$

while the corresponding numerator is

$$\int g^*(\beta_1, \beta_2)\, l(\beta_1, \beta_2)\, d\Omega_2$$
$$= \Gamma\left(\frac{n_1 + n_2}{2} + \nu_2\right) \frac{\nu_2^{\nu_1}}{2\pi \Gamma(\nu_1)} C_1 \frac{\sqrt{\gamma_1 \gamma_2}}{2\pi}$$
$$\int_0^1 \exp\left(\frac{-\gamma_1}{2}(\beta_1 - \xi_1)^2\right) \left(\int_0^{\beta_1} \frac{\exp\left(\frac{-\gamma_2}{2}(\beta_2 - \xi_2)^2\right)}{\left(\frac{1}{2}s(\beta_1, \beta_2) + \nu_2\right)^{\frac{n_1+n_2}{2}+\nu_2}} d\beta_2\right) d\beta_1.$$

To analyze the sensitivity of the Bayes factor, we fixed $E(\beta_i) = \widehat{\beta}_i$, and the Bayes factor was obtained for different values for the hyperparameters of the prior distribution of τ, namely of ν_1 and ν_2, so that $E(\tau) = \nu_1/\nu_2 = 59.17$ with precision $(Var(\tau))^{-1} = \left(\nu_1/\nu_2^2\right)^{-1}$. Considering ξ_i, and γ_i, for $i = 1, 2$, equal to their least squares estimators, i.e., $\xi_1 = 0.114$, $\gamma_1 = 337.933$, $\xi_2 = 0.116$, and $\gamma_2 = 349.484$, the results obtained are presented in the Table 9.2.

In all cases, the Bayes factor provides evidence for the rejection of the null hypothesis and it seems to be sensitive to the choice of the hyperparameters of the prior distribution of τ. There is an indication that the Bayes factor decreases as the precision increases. We repeated the computations under unrestricted parameter space with the same choice for the hyperparameter values. The corresponding results are indicated in parentheses in Table 9.2. Note that the restricted model provides stronger evidence against the null hypothesis.

Table 9.2 Bayes factors computed using Irony and Pereira's proposal and in parenthesis the Bayes factor computed on the unrestricted parameter space

v_1	v_2	$\left(v_1/v_2^2\right)^{-1}$	B_{21}
1	0.0169	0.0003	2.26 (0.51)
10	0.1690	0.0030	2.20 (0.62)
100	1.6900	0.0286	1.87 (0.86)
1000	16.900	0.2856	1.47 (0.95)

Table 9.3 Least squares estimates of the parameters in models (9.4) and (9.5)

Model	Parameter	Mean	Standard deviation
1	β	0.8915	0.0387
2	β_1	0.9548	0.0517
	β_2	0.8292	0.0512

9.3.2 Analysis via MCMC Methods

In this section, we compute the Bayes factor for the dental plaque index data using the MCMC approach. The likelihood function under model $M = 1$ is similar to (9.15) with β_1 and β_2 substituted by β. We consider normal prior distributions for the regression parameters with the restriction $0 \leq \beta \leq 1$ for $M = 1$, and $0 \leq \beta_i \leq 1$, $i = 1, 2$, and $\beta_1 > \beta_2$ for $M = 2$. Initially, we use the least squares estimators presented in Table 9.3 as the values for the hyperparameters and we consider the same informative and non-informative prior distributions for τ used in the analysis under the Irony and Pereira approach.

Although both the likelihood and the chosen prior distributions are standard in Bayesian analyses, it is not simple to obtain expressions for the posterior distributions required to compute the Bayes factor. This difficulty is related to the restriction on the parameter space. We used the BUGS (*Bayesian Inference using Gibbs Sampling*) (version 0.6) software developed by Thomas et al. [30] for such purposes.

Running each model separately, we obtained estimates of the posterior means and standard deviations, which we used as values for the hyperparameters of the "pseudoprior" distributions.

We used the methods proposed by Geweke [15] as convergence diagnostics. They are available in CODA *(Convergence Diagnosis and Output Analysis Software for Gibbs sampling)* that serves as an output processor for BUGS. To compute the Bayes factors, we considered a BUGS run of 1000 burn-in iterations and 10,000 updating iterations. These results are presented in Table 9.4.

In all cases, the Bayes factor favors the alternative hypothesis. We recall here that in the Irony and Pereira's method, the prior under the null hypothesis is a direct consequence of the prior for the parameter (restricted) space. Thus, we only need to choose the prior on the complete parameter space. This sensitivity characteristic of the Bayes factor computed under Irony and Pereira's proposal, displayed in Table 9.2, is not observed when computing the Bayes factor with the methodology proposed in [11], (see Table 9.4). We computed the corresponding Bayes factor under the

Table 9.4 Bayes factors computed using Carlin and Chib's proposal

Parameters		Precision	Bayes factor
v_1	v_2	$\left(v_1/v_2^2\right)^{-1}$	B_{21}
1	0.0169	0.0003	2.86
10	0.1690	0.0030	2.97
100	1.6900	0.0286	2.97
1000	16.9000	0.2856	2.92

unrestricted parameter space, obtaining $B_{21} = 3.24$; this value is greater than the one computed under the restricted parameter space ($B_{21} = 2.86$). Thus, the incorporation of restrictions on the parameters increases the evidence against the null hypothesis.

9.4 Discussion

We considered two different approach based on Bayes factors for comparing restricted regression coefficients in normal regression models.

The two approaches differ in the computation of the posterior probabilities $p(M = j|\mathbf{y})$, $j = 1, 2$. Irony and Pereira [19] directly use the predictive density (9.9) while Carlin and Chib [11] use a Gibbs sampling of the full conditional distributions. The prior opinion on the full parameter space is expressed in a certain way via the construction of the "pseudoprior" distributions when working under the approach in [11], and may not be implemented for some distributions. For example, the Beta prior distributions cannot be used because the corresponding full conditional distribution in this case is not acceptable. The methodology proposed by Irony and Pereira [19], on the other hand, could be inappropriate for the comparison of non-nested models, but although it requires a smaller number of hyperparameters.

In view of the differences in formulation, a direct comparison of the methodologies proposed by Irony and Pereira [19] and Carlin and Chib [11] is not simple. However, under the restricted parameter model both approaches highlight evidence against the null hypothesis.

Bayes factors allow objective conclusions, i.e., for the most appropriate prior distribution they are not sensitive to the selection of hyperparameters.

Clearly other alternatives could be considered for this type of problem. In particular we can mention the following methods: (i) the Bayesian reference criterion (BRC) as suggested in [10], (ii) the posterior likelihood ratio presented in [2], and (iii) the deviance information criterion (DIC) proposed in [26]. Finally, we believe that the approaches considered in this chapter can be extended to more general models as, for example, the one considered in [3].

References

1. Aitkin, M.: Posterior Bayes factors. J. Royal. Stat. Soc. B **53**(1), 111–142 (1991)
2. Aitkin, M., Boys, R.J., Chadwick, T.: Bayesian point null hypothesis testing via the posterior likelihood ratio. Stat. Comput. **15**(3), 217–230 (2005)
3. Aoki, R., Bolfarine, H., Singer, J.M.: Null intercept measurement error regression models. TEST **10**(2), 441–457 (2001)
4. Aoki, R., Achcar, J.A., Bolfarine, H., Singer, J.M.: Bayesian analysis of null intercept errors-in-variables regression for pretest/post-test data. J. Appl. Stat. **30**(1), 3–12 (2003)
5. Barlow, R.E., Bartholomew, D.J., Bremmer, J.N., Brunk, H.H.: Statistical Inference Under Order Restrictions. Wiley, New York (1972)
6. Barros, M.K., Giampaoli, V., Lima, C.R.O.: Hypothesis testing in the unrestricted and restricted parametric spaces of structural models. Comput. Stat. Data Anal. **52**, 1196–1207 (2007)
7. Berger, J.O., Pericchi, L.R.: The intrinsic Bayes factor for model selection and prediction. J. Am. Stat. Assoc. **91**(433), 109–122 (1996)
8. Berger, J.O., Pericchi, L.R.: On the justification of default and intrinsic Bayes factor. In: Lee, J.C., et al. (eds.) Modeling and Prediction. Springer, New York (1997)
9. Berger, J.O., Pericchi, L.R.: Training samples in objective Bayesian model selection. Ann. Stat. **32**(3), 841–869 (2004)
10. Bernardo, J.M., Rueda, R.: Bayesian hypothesis testing: a reference approach. Int. Stat. Rev. **70**(3), 351–372 (2002)
11. Carlin, B.P., Chib, S.: Bayesian model choice via Markov chain Monte Carlo methods. J. Royal Stat. Soc. B **57**(3), 473–484 (1995)
12. De Santis, F.: Alternative Bayes factors: sample size determination and discriminatory power assessment. TEST **4**(3), 503–515 (2007)
13. Gelfand, A.E., Dey, D.K.: Bayesian model choice: asymptotics and exact calculations. J. Royal Stat. Soc. B **56**(3), 501–514 (1994)
14. Gelfand, A.E., Ghosh, S.K.: Model choice: a minimum posterior predictive loss approach. Biometrika **85**(1), 1–11 (1998)
15. Geweke, J.: Evaluating the accuracy of sampling-based approaches to calculating posterior moments moments . In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (eds.) Bayesian Statistics, vol. 4. Oxford University Press, Oxford (1992)
16. Giampaoli, V., Singer, J.M.: Comparison of two normal populations with restricted means. Comput. Stat. Data Anal. **4**(3), 511–529 (2004a)
17. Giampaoli, V., Singer, J.M.: Bayes factors for comparing two restricted means: an example involving hypertense individuals. J. Data Sci. **2**(4), 399–418 (2004b)
18. Han, C., Carlin, B.P.: Markov chain Monte Carlo methods methods for computing Bayes factors: a comparative review. J. Am. Stat. Assoc. **96**(96), 1122–1132 (2001)
19. Irony, T., Pereira, C.: Bayesian hypothesis test: using surface integrals to distribute prior information among the hypotheses. Resenhas IME-USP, pp. 27–46 (1998)
20. Kass, R., Raftery, A.: Bayes factors. J. Am. Stat. Assoc. **90**(430), 777–795 (1995)
21. Moreno, E., Liseo, B.: Default priors for testing the number of components of a mixture. J. Stat. Plan. Inference **111**(1), 129–142 (2003)
22. Moreno, E., Torres, F., Casella, G.: Testing equality of regression coefficients in heteroscedastic normal regression models. J. Stat. Plan. Inference **131**, 117–134 (2005)
23. Mukhopadhyay, N., Ghosh, J.K., Berger, O.J.: Some Bayesian predictive approaches to model selection. Stat. Prob. Lett. **73**(4), 369–379 (2005)
24. O' Hagan, A.: Fractional Bayes factors for model comparision (with discussion). J. Royal Stat. Soc. B **57**(1), 99–138 (1995)
25. Perlman, M.D.: One-sided testing problems in multivariate analysis. Ann. Math. Stat. **40**(2), 549–567 (1969)
26. Plummer, M.: Penalized loss functions for Bayesian model comparison. Biostatistics **9**(3), 523–539 (2008)

27. Robert, C.P., Marin, J.: On some difficulties with a posterior probability approximation technique. Bayesian Anal. **3**(2), 427–442 (2008)
28. Sen, P.K., Singer, J.M., Pedroso de Lima, A.C.: From Finite Sample to Asymptotic Methods in Statistics. Cambridge University Press, Cambridge (2009)
29. Singer, J.M., Andrade, D.F.: Regression models for the analysis of pretest-posttest data. Biometrics **53**, 729–735 (1997)
30. Thomas, A., Spielglharter, D., Gilks, W.: Bugs: a program to perform Bayesian inference using Gibbs sampling. Bayesian Stat. **4**(9), 837–842 (1992)

Chapter 10
A Spanning Tree Hierarchical Model for Land Cover Classification

Hunter Glanz and Luis Carvalho

Abstract Image segmentation persists as a major statistical problem, with the volume and complexity of data expanding alongside new technologies. Land cover classification, one of the largest problems in Remote Sensing, provides an important example of image segmentation whose needs transcend the choice of a particular classification method. That is, the challenges associated with land cover classification pervade the analysis process from data pre–processing to estimation of a final land cover map. Multispectral, multitemporal data with inherent spatial relationships have hardly received adequate treatment due to the large size of the data and the presence of missing values. In this chapter we propose a novel, concerted application of methods which provide a unified way to estimate model parameters, impute missing data, reduce dimensionality, and classify land cover. This comprehensive analysis adopts a Bayesian approach which incorporates prior subject matter knowledge to improve the interpretability, efficiency, and versatility of land cover classification. We explore a parsimonious parametric model whose structure allows for a natural application of principal component analysis to the isolate important spectral characteristics while preserving temporal information. Moreover, it allows us to impute missing data and estimate parameters via expectation-maximization. We employ a spanning tree approximation to a lattice Potts model prior to incorporating spatial relationships in a judicious way and more efficiently access the posterior distribution of the pixel labels. We achieve exact inference of the labels via the centroid estimator. We demonstrate this series of analysis on a set of MODIS data centered on Montreal, Canada.

10.1 Introduction

The role of humans as members of many of Earth's ecosystems and contributors to many others has become less difficult to study in recent decades. Multitemporal, remotely sensed data of the entire Earth has become ubiquitous and with certain

H. Glanz (✉)
California Polytechnic State University San Luis Obispo, CA 93407 USA
e-mail: hglanz@calpoly.edu

L. Carvalho
Boston University, Boston, MA 02215, USA
e-mail: lecarval@bu.edu

A. Polpo et al. (eds.), *Interdisciplinary Bayesian Statistics,*
Springer Proceedings in Mathematics & Statistics 118, DOI 10.1007/978-3-319-12454-4_10

technological advances during the 1990s continental and global scale land cover was mapped for the first time using remotely sensed data [5, 11].

To continue informing Earth system models [2, 7, 17] and providing Earth scientists with an accurate picture of global land cover, methods for land cover classification need to adapt to and mirror the sophistication of remote sensing data and technology.

In Statistics, land cover classification attracts a great deal of attention. Despite being a classic problem, classification here involves three significant challenges: large data, spatio-temporal structure, and missing data. The MODIS (Moderate Resolution Imaging Spectroradiometer) instrument images the entire surface of the Earth every 1–2 days [8]. The data of interest involve multispectral observations at each of roughly 1.8 billion one km^2 pixels for around 11 years. Because these data consist of two distinct dimensions, spectral and temporal, models that can exploit this structure will yield increased interpretability and potentially simpler estimation procedures. Additionally, spatial information should capitalize on data in nearby pixels to produce more accurate classifications. Clouds, snow, and other disruptive phenomena prevent clean, high quality images of the Earth's surface. As a result, missing values exist throughout the data. To overcome these challenges and successfully classify land cover, we employ a series of tools which provide physically satisfying and interpretable results with an eye toward computational efficiency.

We begin by proposing a model for the data which takes advantage of spectral and temporal structure present. Because seasonal variation is a first-order property that helps to distinguish many land cover classes, multitemporal information provides critical information that our model isolates. We use an expectation-maximization (EM) procedure [6] to estimate the parameters of this model in the presence of missing data. With temporal information established, we reduce the dimensionality of the data by applying a principal component analysis to the spectral variation. Missing data is then imputed using another EM procedure. Because we pursue a Bayesian approach, we specify a prior on the lattice of pixels to incorporate spatial information. To make posterior inference computationally tractable, we use a third EM procedure to identify an optimal spanning tree approximation to the lattice. With this tree in hand, we proceed to estimate the pixel labels using the *centroid* estimator [3], which better suits this type of high-dimensional discrete inference.

10.2 Likelihood Model and Data Compression

The practical problem of interest here involves labeling an image domain pixelwise with a given set S of discrete labels representing land cover classes. One such example is the International Geosphere–Biosphere Programme which consists of 17 land cover classes [8]. The data consists of multivariate observations at each pixel in the image. In general, let P be a set of pixels in the image L of size $n = p \times q$ and $S = \{1, \ldots, C\}$ a set of C labels.

The classification problem consists of assigning a label from the set S to each node in the set of nodes P.

Our application of interest favors the following matrix normal likelihood, since our main motivation is to isolate the two natural dimensions of the data, *spectral* and *temporal*. The spectral-temporal series X_v at pixel v given its land cover class, θ_v, is:

$$X_v \mid \theta_v = c \overset{\text{ind}}{\sim} \text{Matrix-}N(\mu_c, \Sigma_c, \Sigma_s), \tag{10.1}$$

where Σ_s and Σ_c correspond to separable pieces of variation in the data; in our case spectral and temporal variation, respectively. In (10.1), X_v and μ_c are $B \times T$, Σ_s is $B \times B$, and Σ_c is $T \times T$ for each land cover class c. This matrix normal model isolates the two natural dimensions of the data into an equivalent multivariate normal likelihood of the following form [19]:

$$\text{vec}(X_v) \mid \theta_v = c \overset{\text{ind}}{\sim} N(\text{vec}(\mu_c), \Sigma_s \otimes \Sigma_c). \tag{10.2}$$

To handle identifiability issues in (10.2) we impose $\Sigma_{s_{11}} = 1$. In our application we begin with data from 7 spectral bands collected by MODIS at 46 time points within a single calendar year. Our first step toward compressing the data involves reducing this time series from 46 to 28 time points. This middle 60 % of the year corresponds to spectrally more distinguishable data and physically more vegetated land cover in the region we conduct our simulation study. Hence, for (10.1) and (10.2), $B = 7$ and $T = 28$.

We follow an empirical Bayes approach and use training data, independent of the image of interest, to estimate the parameters in (10.2) prior to the label assignment procedure. To formally estimate the parameters of (10.2) in the presence of missing data we exploit an expectation-maximization procedure derived in [9]. The procedure described here produces accurate estimates while achieving good separability among land cover classes.

10.2.1 Missing Data Imputation and PCA

To alleviate a portion of the computational burden associated with classifying pixels while accounting for missing data, we propose a pre-processing step to impute missing values. We assume that data are missing at random. Using the estimates for the parameters in (10.2), we derive a second expectation-maximization procedure for estimating the missing values in each pixel independent of the neighboring pixels. Treating the pixel label θ_v for pixel v as latent with a prior multinomial distribution, $\theta_v \overset{\text{ind}}{\sim} MN(\exp(\mathbf{h}))$, we compute an update for the missing data Z_v and iterate until convergence:

$$Z_v^{(t+1)} = \left[\sum_l \Pr(\theta_v = l \mid Y_v, Z_v^{(t)}) (\Sigma_l^{-1})_{zz} \right]^{-1} \left[\sum_l \Pr(\theta_v = l \mid Y_v, Z_v^{(t)}) (\Sigma_l^{-1})_{zz} \left(\mu_{l,z_v} - (\Sigma_l^{-1})_{zz}^{-1} (\Sigma_l^{-1})_{yz}^{\top} (Y_v - \mu_{l,y_v}) \right) \right]. \tag{10.3}$$

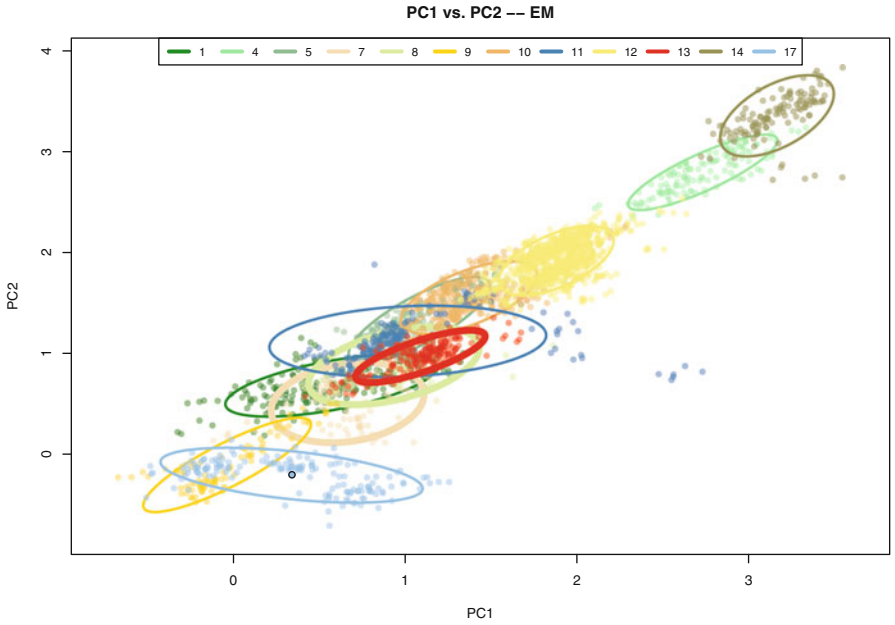

Fig. 10.1 Training data transformed into the space of the first two principal components

The update in (10.3) contains the conditional means given Y_v, but weighted by the concentration submatrices and the posteriors $\Pr(\theta_v = l \mid Y_v, Z_v^{(t)})$. With the observation at each pixel completed with this imputation, we now address the size of the data.

We ultimately seek labels for pixels across the globe, roughly 1.8 billion pixels. While subsets of interest could be analyzed independently, the size of the data remains cumbersome. Experts in the field attest to the correlation between spectral bands and the consequent redundancy of information among them [4]. To isolate the most useful information present in the data in a lower dimensional space we use principal components analysis (PCA) [12]. Because spectral variation transcends the differences between land-cover classes (our C labels), we target only $\hat{\Sigma}_s$ with PCA. Previous empirical work [10] indicates that a vast majority of the spectral variation in the data is captured by the first three principal components (PCs). In fact, not only was approximately 91% of the spectral variation present in the first three PCs, but also the first three components corresponded to the well known Tasseled-Cap Transformation [13]. So, for the remainder of this chapter we denote by X_v, the original data transformed into the space of the first three principal components. In a similar manner, the parameters μ_c and Σ_s will denote transformed parameters (i.e., $B = 3$ instead of $B = 7$).

10.3 Prior Model and Land Cover Classification

We denote a complete set of labels for the pixels in the image by $\boldsymbol{\theta}$. Whereas, in Sect. 10.2.1 our prior on the land cover classes for the entire image amounted to $\Pr(\theta) = \prod_v \Pr(\theta_v) = \exp\{\sum_v \mathbf{h}^\top \mathbf{e}(\theta_v)\}$, we wish to extend this in a way that incorporates spatial information. A Markov random field (MRF) accommodates this interest and so we employ a Potts model [16] prior on $\boldsymbol{\theta}$,

$$\Pr(\boldsymbol{\theta}) \propto \exp\left\{\sum_{v\in P} \mathbf{h}^\top \mathbf{e}(\theta_v) + \eta \sum_{(u,v)\in L} \mathbf{e}(\theta_u)^\top \mathbf{J}\mathbf{e}(\theta_v)\right\}, \tag{10.4}$$

where the image, L, consists of a two-dimensional lattice on the grid of pixels. In (10.4), $\mathbf{h}$ characterizes the global distribution of labels and $\mathbf{J}$ describes the relationship between neighboring pixels. More specifically, $\mathbf{J}_{r,s}$ corresponds to an empirical estimate of the log joint probability of observing labels r and s in adjacent pixels. In an empirical Bayes approach, we obtain the values of $\mathbf{h}$ and $\mathbf{J}$ from training data or previous land cover products similar to, but independent of, the image of interest. Hyperprior η controls the strength of the spatial influence of neighboring pixels. For a thorough investigation of the use of MRFs to classify remote sensing images, we refer the reader to [14] and the references therein.

10.3.1 Posterior Inference

Traditional posterior inference via maximum *a posteriori* (MAP) estimation necessitates the computation of the label configuration which maximizes

$$\Pr(\theta \mid X) = \exp\left\{\sum_v l(X_v \mid \theta_v) + \sum_{v\in P} \mathbf{h}^\top \mathbf{e}(\theta_v) + \eta \sum_{(u,v)\in L} \mathbf{e}(\theta_u)^\top \mathbf{J}\mathbf{e}(\theta_v)\right\} / Z_L(X), \tag{10.5}$$

where l is the log-likelihood specified by (10.2). As we will see in Sect. 10.3.2, we will need the posterior marginal probabilities which means we need to compute

$$Z_L(X) = \sum_{\theta\in S^n} \exp\left\{\sum_v l(X_v \mid \theta_v) + \sum_{v\in P} \mathbf{h}^\top \mathbf{e}(\theta_v) + \eta \sum_{(u,v)\in L} \mathbf{e}(\theta_u)^\top \mathbf{J}\mathbf{e}(\theta_v)\right\}. \tag{10.6}$$

Because of the *high connectivity* of the two-dimensional lattice, the computation of (10.6) is intractable. A popular method in this framework, Iterative Conditional Modes (ICM) [1] maximizes a joint probability and thus does not encounter the difficulty that the lattice presents. However, the value reached by ICM does not necessarily correspond to even the optimal, MAP estimate using (10.5).

Though we conduct our inference using an estimator different from the MAP (the centroid estimator), the ICM solution remains suboptimal. Other variational approaches might attempt to approximate the distribution on the lattice (10.4), with a distribution that eliminates the computational intractability of computing (10.6). In a vein similar to this, we pursue a graph approximation to the lattice, which retains the most important features of the lattice structure while being simpler to compute on.

To compute (10.6) and continue with posterior inference, we approximate the two-dimensional lattice with a *spanning tree*. Minimally connected, this spanning tree approximation allows for more efficient computations. This tree needs to retain the most important spatial information in order to ensure the quality of the approximation as well as preserve the purpose of the hierarchical graphical model.

By approximating the lattice with a spanning tree we introduce an additional layer to the model, yielding

$$\Pr(\boldsymbol{\theta}, X) \propto \sum_{T \in \tau(L)} \Pr(X \mid \boldsymbol{\theta}, T) \Pr(\boldsymbol{\theta} \mid T) \Pr(T), \tag{10.7}$$

where we assume $\Pr(T) \propto 1$ and $\tau(L)$ corresponds to the space of all spanning trees on L. Despite the approximation of L with T in (10.4), the sum over T in (10.7) presents a new computationally intractable piece of the model. To circumvent this, we propose the following approximation:

$$\Pr(\boldsymbol{\theta}, X) \approx \Pr(\boldsymbol{\theta}, X \mid T^*) = \Pr(X \mid \boldsymbol{\theta}, T^*) \Pr(\boldsymbol{\theta} \mid T^*). \tag{10.8}$$

Here, T^* represents a spanning tree optimized to capture the most important spatial relationships in the lattice. To accomplish this, we treat the vertex labels as latent and identify T^* via an EM procedure that iterates

$$\begin{aligned} T^{(t+1)} &= \arg\max_{T \in \tau(L)} \mathbb{E}_{\boldsymbol{\theta} \mid X, T^{(t)}}[\log \Pr(\boldsymbol{\theta}, X, T)] \\ &= \arg\max_{T \in \tau(L)} \mathbb{E}_{\boldsymbol{\theta} \mid X, T^{(t)}} \left[\sum_{(u,v) \in T} \mathbf{e}(\theta_u)^\top \mathbf{J} \mathbf{e}(\theta_v) \right] \\ &= \arg\max_{T \in \tau(L)} \sum_{(u,v) \in T} \mathbb{E}_{\theta_u, \theta_v \mid X, T^{(t)}} [\mathbf{e}(\theta_u)^\top \mathbf{J} \mathbf{e}(\theta_v)]. \end{aligned} \tag{10.9}$$

Thus, to obtain $T^{(t+1)}$, we just need to assign the conditional posterior mean of $\mathbf{e}(\theta_u)^\top \mathbf{J} \mathbf{e}(\theta_v)$ as a weight to each edge $(u, v) \in L$ and then find the corresponding maximum weighted spanning tree [15]. In this way, we maximize the similarity, as measured by the entries of $\mathbf{J}$, of neighboring pixels. Figure 10.2 gives an example result of applying this tree approximation procedure to a toy map.

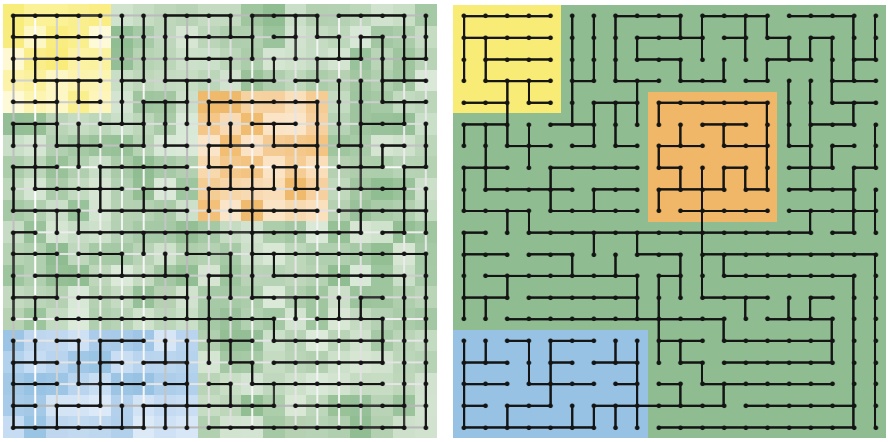

Fig. 10.2 Example result of approximating tree, T^*. Opacity of pixels in left plot indicate strength of the posterior of the true class. Darkness of edges in the left plot indicate the strength of edge weights as shown in 10.9. Final solution is in the right plot

10.3.2 Centroid Estimation

The ubiquitous maximum *a posteriori* estimation assigns to the set of pixels the following:

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\min_{\tilde{\boldsymbol{\theta}} \in \Theta} \mathbb{E}_{\boldsymbol{\theta} \mid X}[I(\tilde{\boldsymbol{\theta}} \neq \boldsymbol{\theta})] = \arg\max_{\tilde{\boldsymbol{\theta}} \in \Theta} \Pr(\tilde{\boldsymbol{\theta}} \mid X). \tag{10.10}$$

In a high-dimensional situation such as this, however, the MAP estimator can myopically find a solution that does not represent the posterior space well. Therefore, we assign a label to each pixel in the image via the *centroid* estimator [3]:

$$\hat{\boldsymbol{\theta}}_C = \arg\min_{\tilde{\boldsymbol{\theta}} \in \Theta} \mathbb{E}_{\boldsymbol{\theta} \mid X}[H(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta})] = \arg\max_{\tilde{\boldsymbol{\theta}} \in \Theta} \sum_{v} \Pr(\tilde{\theta}_v = \theta_v \mid X). \tag{10.11}$$

Here, $H(\cdot, \cdot)$ represents the Hamming loss. Use of this loss function means assigning labels not with the full posterior joint, but with the posterior marginal distribution at each pixel. Not only does this better represent the posterior space, but computation of the centroid is made easy by our use of T^*. Message-passing algorithms allow for quick calculation of the posterior marginal distribution at each pixel. Our centroid estimator assigns the label to pixel v which maximizes the posterior marginal distribution at pixel v.

10.4 Case Study Results

To develop and test our methodology we used Nadir BRDF-adjusted surface reflectance (NBAR) data from MODIS [18]. Specifically, we extracted NBAR data for land cover training sites that are used to produce the MODIS Land Cover Type

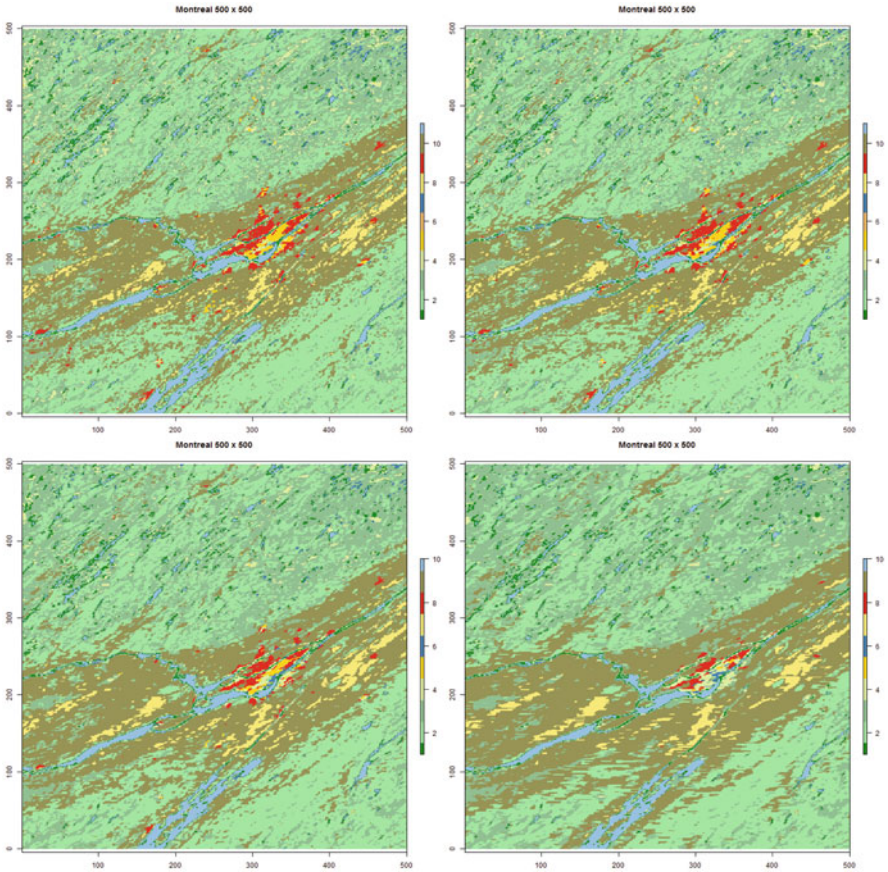

Fig. 10.3 Land cover map of a 500 pixel square region surrounding Montreal, Canada. *Top-left*, $\eta = 1$; *top-right*, $\eta = 2$; *bottom-left*, $\eta = 5$; *bottom-right*, $\eta = 10$

product [8]. These sites are produced by manual interpretation of high resolution imagery and provide exemplars of land cover classes that are used in a supervised classification of global land cover at 500-m spatial resolution. For this analysis we used a subset of the MODIS Land Cover Training site database that includes 204 sites located in the conterminous United States (extending from MODIS tile v04h08 in the northwest to MODIS tile v05h11 in the southeast). These sites encompass 2692 MODIS pixels and all major biomes and land cover types in the lower 48 United States. The data we use here include data from 28 8-day periods in 2005 for MODIS bands 1–7 (i.e., 196 total features). Though the International Geosphere–Biosphere Programme (IGBP) classification scheme contains 17 classes, our training data set consists of only 12. Nevertheless, we proceed with parameter estimation, missing value imputation, and classification.

Figure 10.3 maps the land cover results from our analysis pipeline applied to a 500 pixel $\times$ 500 pixel region surrounding Montreal, Canada. By varying the value

of η in the Potts prior models, we can strengthen or weaken the importance of spatial homogeneity. From the top-left to the bottom-right, the coarseness of the result increases as we increase the value of η from 1 to 10. Note that we lose some potentially important portions of the land cover product by making it more difficult to classify adjacent pixels as two very different classes. For example, some red (urban) patches have disappeared from $\eta = 1$ to $\eta = 10$. The visual quality of the maps in Fig. 10.3 signifies both the success of this method and its versatility. We accurately capture the agricultural swath extending through Montreal as well as the particular types of forest bordering it. Increasing the emphasis on the spatial relationships successfully preserved this overall pattern.

10.5 Conclusion

In this article we present a suite of tools that address the land cover classification problem from start to finish. Modern remote sensing instruments collect multispectral, multitemporal data which necessitate an approach that models this specific structure in a natural way. The choice of a matrix normal likelihood provides increased interpretability while also lending itself to isolated dimension reduction. We estimate the parameters of this likelihood via an expectation-maximization procedure in order to take missing data into account. By only targeting the spectral variation with principal components analysis we retain critical, class-specific temporal information. In order to continue with land cover classification we impute missing data with another EM procedure. Many traditional land cover classification methods assign labels to pixels independent of the information present in surrounding pixels. Because we wish to incorporate information about spatial relationships, we specify a Potts prior on the lattice of pixels. The connectedness of the lattice makes inference via the posterior computationally intractable, and so we approximate the lattice with a representative spanning tree determined with another EM algorithm. The centroid estimator better characterizes the posterior space, which means we assign the labels that maximize the posterior marginal distributions.

We apply the proposed methods to a small area surrounding Montreal, Canada and achieve satisfying results. Future work will include formal calibration of the hyperparameter η. Also, a slight striping pattern can be seen in parts of Fig. 10.3 which may be related to the spanning tree used to derive inference. Consideration needs to be given to how to overcome these potential artifacts of the model.

Acknowledgements Hunter Glanz was supported by funding from NASA under grant number NNX11AG40G. Luis Carvalho was supported by NSF grant DMS-1107067.

References

1. Besag, J.: On the statistical analysis of dirty pictures. J. R. Stat. Soc. **48**(3), 259–302 (1986). (With discussions)
2. Bonan, G.B., Oleson, K.W., Vertenstein, M., Levis, S., Zeng, X., Dai, Y., Dickinson, R.E., Yang, Z.L.: The land surface climatology of the community land model coupled to the NCAR community climate model*. J. Clim. **15**(22), 3123–3149 (2002)
3. Carvalho, L.E., Lawrence, C.E.: Centroid estimation in discrete high-dimensional spaces with applications in biology. Proc. Natl. Acad. Sci. **105**(9), 3209–3214 (2008)
4. Crist, E.P., Cicone, R.C.: A physically-based transformation of Thematic Mapper data—The TM Tasseled Cap. IEEE Trans. Geosci. Remote Sens. **22**(3), 256–263 (1984)
5. DeFries, R., Townshend, J.: NDVI-derived land cover classifications at a global scale. Int. J. Remote Sens. **15**(17), 3567–3586 (1994)
6. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. J. Royal Stat. Soc. Series B (Methodological) **39**(1), 1–38 (1977)
7. Ek, M., Mitchell, K., Lin, Y., Rogers, E., Grunmann, P., Koren, V., Gayno, G., Tarpley, J.: Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model. J. Geophys. Res. **108**(D22), 8851 (2003)
8. Friedl, M.A., Sulla-Menashe, D., Tan, B., Schneider, A., Ramankutty, N., Sibley, A., Huang, X.: MODIS collection 5 global land cover: Algorithm refinements and characterization of new datasets. Remote Sens. Environ. **114**, 168–182 (2010)
9. Glanz, H., Carvalho, L.: An expectation-maximization algorithm for the matrix normal distribution. arXiv preprint arXiv:1309.6609 (2013)
10. Glanz, H., Carvalho, L., Sulla-Menashe, D., Friedl, M.: A parsimonious model for land cover classification and characterization of training data using multitemporal remotely sensed imagery. Submitted (2014)
11. Hansen, M., Defries, R., Townshend, J., Sohlberg, R.: Global land cover classification at 1km spatial resolution using a classification tree approach. Int. J. Remote Sens. **21**(6–7), 1331–1364 (2000)
12. Jolliffe, I.: Principal Component Analysis. Wiley, Hoboken (2005)
13. Lobser, S., Cohen, W.: MODIS tasselled cap: land cover characteristics expressed through transformed MODIS data. Int. J. Remote Sens. **28**(22), 5079–5101 (2007)
14. Moser, G., Serpico, S.B., Benediktsson, J.A.: Land-cover mapping by Markov modeling of spatial–contextual information in very-high-resolution remote sensing images. Proc. IEEE **101**(3), 631–651 (2013)
15. Papadimitriou, C.H., Steiglitz, K.: Combinatorial Optimization: Algorithms and Complexity. Dover, New York (1998)
16. Potts, R.: Some generalized order-disorder transformations. Proc. Camb. Philos. Soc. **48**, 106–109 (1952)
17. Running, S.W., Coughlan, J.C.: A general model of forest ecosystem processes for regional applications I. Hydrologic balance, canopy gas exchange and primary production processes. Ecol. Model. **42**(2), 125–154 (1988)
18. Schaaf, C., Gao, F., Strahler, A., Lucht, W., Li, X., Tsang, T., Strugnell, N., Zhang, X., Jin, Y., Muller, J., Lewis, P., Barnsley, M., Hobson, P., Disney, M., Roberts, G., Dunderdale, M., Doll, C., d'Entremont, R., Hu, B., Liang, S., J.L., P., Roy, D.: First operational BRDF, Albedo Nadir reflectance products from MODIS. Remote Sens. Environ. **83**(1), 135–148 (2002)
19. Srivastava, M., Khatri, C.: An Introduction to Multivariate Statistics. North Holland, New York (1979)

Chapter 11
Nonparametric Bayesian Regression Under Combinations of Local Shape Constraints

Khader Khadraoui

Abstract A nonparametric Bayesian method for regression under combinations of local shape constraints is proposed. The shape constraints considered include monotonicity, concavity (or convexity), unimodality, and in particular, combinations of several types of range-restricted constraints. By using a B-spline basis, the support of the prior distribution is included in the set of piecewise polynomial functions. The first novelty is that, thanks to the local support property of B-splines, many combinations of constraints can easily be considered by identifying B-splines whose support intersects with each constrained region. Shape constraints are included in the coefficients prior using a truncated Gaussian distribution. However, the prior density is only known up to the normalizing constant, which does change with the dimension of coefficients. The second novelty is that we propose to simulate from the posterior distribution by using a reversible jump MCMC slice sampler, for selecting the number and the position of the knots, coupled to a simulated annealing step to project the coefficients on the constrained space. This method is valid for any combination of local shape constraints and particular attention is paid to the construction of a trans-dimensional MCMC scheme.

11.1 Introduction

Estimation of a regression function under combinations of local shape and smoothness constraints is of considerable interest in many applications. Typical examples include, among others, the dose–response curves in medicine, actuarial graduation, construction of the utility function, production functions in industry, etc. The function that provides the best prediction of a dependent variable y conditionally to an independent variable x is the conditional expectation $\mathbb{E}(y|x) = f(x)$. This function is called regression function and the estimation of f from n independent copies of (x, y) is a problem in statistical inference. We consider the case where $(x, y) \in \mathbb{R} \times \mathbb{R}$.

K. Khadraoui (✉)
Department of Mathematics and Statistics, Laval University, Quebec City,
QC G1V 0A6, Canada
e-mail: khader.khadraoui@mat.ulaval.ca

A. Polpo et al. (eds.), *Interdisciplinary Bayesian Statistics,*
Springer Proceedings in Mathematics & Statistics 118, DOI 10.1007/978-3-319-12454-4_11

To study the constrained regression, frequentist and Bayesian approaches are proposed in the literature. Obviously, these approaches have focused on a single shape constraint on an interval determined by the domain of the independent variable. In the frequentist case, there is a rich literature based essentially on kernel methods and regression splines or smoothing splines, which we shall only touch briefly here [16, 17, 18, 21, 25, 27]. For instance, in the context of works based on spline, regression under monotonicity constraint has been studied by Mammen et al. [16] and Ramsey [21]. Mammen et al. [16] used a discretization with smoothing and isotonizing constraints at each stage for estimating a monotone regression function. Mammen and Thomas-Agnan [15] proposed a method based on smoothing splines by first calculating the smoothing spline without constraints and second by projecting the spline in the constrained space using a Sobolev norm. A more general work studying inference under constraints of convexity and monotone in the regression spline was proposed by Meyer [17]. For a description of the most common nonparametric methods under shape constraints, refer [4]. In the Bayesian case, the literature is less explored, relatively recent, and mainly concerns isotonic regression. Both paper of Gelfand and Kuo [7] and Ramgopal et al. [20] proposed a Bayesian method for constrained estimation of dose–response curves. Lavine and Mockus [14] used a Dirichlet prior in a nonparametric Bayesian estimation of isotonic regression function. Their method works only for estimation problems under monotone constraints and cannot be used directly under the presence of flat regions in the dose–response curve. To solve the problem of flat regions, Holmes and Heard [11] proposed a Bayesian approach for isotonic regression using a piecewise constant function with unknown positions and number of knots. Another different approach was proposed by Neelon and Dunson [19] giving an approximation of the regression function by a piecewise linear function along with a correlation between the slopes that have been implemented through the prior distribution. This prior charges the zero slope which allows the method to effectively detect flat regions in the curve. Gunn and Dunson [10] used a Bayesian hierarchical model for modeling unimodal curves. Recently, Shively and Sager [22] proposed two efficient approaches for smooth and monotone regression function: the first is based on the characterization of smooth monotone functions proposed by Ramsay [21] and the second is based on regression splines generated by a truncated polynomial basis. Shively et al. [23] actually generalized the methods from Shively and Sager [22], extending them to situations other than monotone regression that can be summarized by linear constraints (or almost linear constraints) on higher order derivatives (concavity, unimodality, etc).

In this framework, we consider constrained function estimation using free-knot B-spline model. This model is originally proposed by Denison et al. [5] for unconstrained nonparametric Bayesian regression. An interesting feature of using free-knot is that the data are allowed to determine the number and position of knots. The novelty of our work is that constraints include combinations of several types of local shape restrictions, constraints on the value of the regression function and also constraints can be range-restricted. It is worth noting that combinations of constraints has never been considered in the literature, except in [1] by using a polynomial plus an integrated Brownian motion, though it seems of practical interest and realistic.

Generally, shape constraints will be conveniently expressed as (pseudo)-differential inequalities of the regression function f, assuming for the moment that f is sufficiently smooth by controlling the degree of B-spline basis. Important examples are $Df \gtrless 0$ to check monotonicity or unimodality properties as well as $D^2 f \gtrless 0$ for convexity or concavity. We use the reversible jump Markov chain Monte Carlo (MCMC) technique [9] for selecting the number and the position of the knots. Recall that the reversible jump Metropolis–Hastings scheme involves the computations of the ratio of likelihoods and priors for two sets of parameters whose dimension may differ. Contrary to the unconstrained case, our constraints on the coefficients induce numerical difficulties in the computation of the prior ratio. Exactly, in this chapter, constraints are included in the coefficients prior using a truncated Gaussian distribution. Thus, the prior density is only known up to the normalizing constant which does change with the dimension of coefficients. For this reason, we integrated out the coefficients from the reversible jump MCMC chain and used a simulated annealing step that ensures the projection of the coefficients on the constrained space. In this spirit, the reversible jump MCMC slice sampler coupled to the simulated annealing step perform well for any combination of shape constraints.

This chapter is organized as follows. In Sect. 11.2, we introduce the unconstrained Bayesian inference. This can be viewed as an extension of Denison et al. [5] as well as DiMatteo et al. [6] who treated the free-knot case. Section 11.3 outlines the constrained inference for the nonparametric Bayesian regression. Section 11.4 presents a numerical experiment to show the sample properties of the constrained estimator relative to the unconstrained one given in Sect. 11.2.

11.2 The Bayesian Model

In this section, we present the Bayesian model and its specifications. It is assumed that data $(x_i, y_i)_{i=1}^n$ are independent such that $\boldsymbol{y} = (y_1, \ldots, y_n)'$ and $\boldsymbol{x} = (x_1, \ldots, x_n)'$. Clearly, we consider the usual regression model:

$$y_i | x_1, \ldots, x_n \sim p(y_i | f(x_i), \sigma) \qquad i = 1, \ldots, n, \tag{11.1}$$

where $p(y_i|\theta, \sigma)$ is a normal distribution $\mathcal{N}(\theta, \sigma^2)$, f is a real-valued unknown function in $[a, b] \subset \mathbb{R}$ and σ is introduced as a dispersion parameter in the model. To complete the model (11.1), we decompose the function f in a B-spline basis by assuming that f belongs to a class of finite dimension functions. In particular, the function f is modeled by a 4-order B-spline as the class of cubic splines is wide and can be used to approximate any locally smooth function. Thus, for all $x \in [a, b]$, we have the linear combination:

$$f(x) = \sum_{j=1}^{\kappa+4} \beta_j B_{j,\mathbf{t}}(x), \tag{11.2}$$

where $\beta = (\beta_1, \ldots, \beta_{\kappa+4})'$ is the vector of regression coefficients and κ is the dimension of interior knots and $B_{j,\mathbf{t}}$ is a B-spline function. We denote the parameters

space by $\Theta = \cup_{\kappa=1}^{\infty}(\{\kappa\} \times \Theta_\kappa)$ where Θ_κ is a subspace of the Euclidean space $\mathbb{R}^{\kappa+4} \times [a,b]^\kappa \times [0,\infty)$. Also, we denote by $\theta^{(\kappa)} = (\beta_1, \ldots, \beta_{\kappa+4}, t_1, \ldots, t_\kappa, \sigma^2)'$ a generic element of Θ_κ. We shall now construct a probability measure on the parameters space Θ by constructing a prior on the approximating set of regression functions. First, the prior of k is assumed to be a truncated Poisson distribution ($\pi_\kappa(\kappa) \sim \mathcal{P}_{\{E\}}(\lambda)$, $\lambda > 0$ and $\{E\}$ is some discrete set). Next, inside each model κ, the prior for $\beta \in \mathbb{R}^{\kappa+4}$ is defined by a multivariate normal prior. Specifically, the prior of β is specified by the g-prior of Zellner given by

$$\pi_\beta(\beta|\mathbf{t},\kappa,\sigma^2) \sim \mathcal{N}_{\kappa+4}\big(0, \sigma^2 n(B'_{\kappa,\mathbf{t}}B_{\kappa,\mathbf{t}})^{-1}\big), \tag{11.3}$$

where $B_{\kappa,\mathbf{t}}$ denotes the B-spline basis. Furthermore, for knots position, we consider the prior $\pi_{\mathbf{t}}(\mathbf{t}|\kappa) = \kappa!/(b-a)^\kappa$ and for the variance, we adopt an inverse-gamma distribution prior $\pi_{\sigma^2}(\sigma^2) \sim IG(\tau_1,\tau_2)$ where $\tau_1, \tau_2 > 0$. The prior (11.3) was widely discussed in [12]. Concerning the prior (11.3), although arbitrary, such a choice guarantees important properties; in particular, two close functions $B_{j,\mathbf{t}}$ and $B_{j',\mathbf{t}}$ correspond to two coefficients β_j and $\beta_{j'}$ highly correlated. Note that if the functions $B_{j,\mathbf{t}}$ and $B_{j',\mathbf{t}}$ are near, then

$$r_{jj'} = \frac{\int_{[a,b]} B_{j,\mathbf{t}}(x)B_{j',\mathbf{t}}(x)dx}{\big(\int_{[a,b]} B^2_{j,\mathbf{t}}(x)dx \int_{[a,b]} B^2_{j',\mathbf{t}}(x)dx\big)^{1/2}} \approx 1. \tag{11.4}$$

It is easy to remark that the prior correlation between coefficients decreases with $r_{jj'}$. We can interpret the case $r_{jj'} = 0$ as follow: B-spline functions $B_{j,\mathbf{t}}, B_{j',\mathbf{t}}$ are orthogonal and coefficients $\beta_j, \beta_{j'}$ are independent. The priors considered are proper as well. As argued by Denison et al. [5], we can develop Bayesian inference in the same spirit as Green [9]. Concerning the Bayesian unconstrained regression with free knots, we do not claim any originality but essentially follow the methodology that was proposed in [6], in which a careful literature review has been given on the subject. By integrating the variance and the coefficients out, we obtain the likelihood $L(\mathbf{y}|\kappa,\mathbf{t})$. In the sequel, we put $K = \kappa + 4$ and $V = n(B'_{\kappa,\mathbf{t}}B_{\kappa,\mathbf{t}})^{-1}$. Then, for $\beta \in \mathcal{X}_K \subset \mathbb{R}^K$, we can write

$$\begin{aligned} L(\mathbf{y}|\kappa,\mathbf{t}) = \int_{\mathcal{X}_K}\int_0^\infty \big[(2\pi\sigma^2)^{-\frac{n}{2}} \exp\Big\{-\frac{(\mathbf{y}-B_{\kappa,\mathbf{t}}\beta)'(\mathbf{y}-B_{\kappa,\mathbf{t}}\beta)}{2\sigma^2}\Big\}(2\pi\sigma^2)^{-\frac{K}{2}}|V|^{-1/2} \\ \exp\Big\{-\frac{\beta'(V)^{-1}\beta}{(2\sigma^2)}\Big\}\frac{\tau_2^{\tau_1}}{\Gamma(\tau_1)}(\sigma^2)^{-(\tau_1+1)}\exp\Big\{-\frac{\tau_2}{\sigma^2}\Big\}\big]d\beta d\sigma^2 \\ = \frac{\tau_2^{\tau_1}\Gamma(\tau_1^*)(\tau_1^*)^{K/2}\big|\frac{\tau_2^* V^*}{\tau_1^*}\big|^{1/2}}{(2\pi)^{\frac{n}{2}}\Gamma(\tau_1)|V|^{1/2}(\tau_2^*)^{(2\tau_1^*+K)/2}}, \end{aligned} \tag{11.5}$$

where $\Gamma(\cdot)$ denotes the Gamma function and:

$$m^* = (V^{-1} + B'_{\kappa,\mathbf{t}}B_{\kappa,\mathbf{t}})^{-1}(B'_{\kappa,\mathbf{t}}\mathbf{y}) = \big\{(n(B'_{\kappa,\mathbf{t}}B_{\kappa,\mathbf{t}})^{-1})^{-1} + B'_{\kappa,\mathbf{t}}B_{\kappa,\mathbf{t}}\big\}^{-1}(B'_{\kappa,\mathbf{t}}\mathbf{y}) = \frac{n}{1+n}\widehat{\beta};$$

$$V^* = (V^{-1} + B'_{\kappa,\mathbf{t}}B_{\kappa,\mathbf{t}})^{-1} = \left\{(n(B'_{\kappa,\mathbf{t}}B_{\kappa,\mathbf{t}})^{-1})^{-1} + B'_{\kappa,\mathbf{t}}B_{\kappa,\mathbf{t}}\right\}^{-1} = \frac{n}{1+n}(B'_{\kappa,\mathbf{t}}B_{\kappa,\mathbf{t}})^{-1}; \tag{11.6}$$

$$\tau_1^* = \tau_1 + n/2;$$

$$\tau_2^* = \tau_2 + \{\boldsymbol{y}'\boldsymbol{y} - (m^*)'(V^*)^{-1}m^*\}/2.$$

Note that $\widehat{\beta}$ denotes the least squares estimator. We precise that we obtain (11.5) and (11.7) thanks to the standard Bayesian calculation from conjugate priors and in particular from the g-prior. Now, we put $\mathbf{P}_\beta = (\sigma^2, \mathbf{t}, \kappa, \boldsymbol{x}, \boldsymbol{y})$ and $\mathbf{P}_{\sigma^2} = (\beta, \mathbf{t}, \kappa, \boldsymbol{x}, \boldsymbol{y})$. We obtain the full conditional posterior distributions

$$\begin{aligned}\beta|\mathbf{P}_\beta &\sim \mathcal{N}_{\kappa+4}\Big(\frac{n}{1+n}\widehat{\beta}, \frac{n}{1+n}\sigma^2(B'_{\kappa,\mathbf{t}}B_{\kappa,\mathbf{t}})^{-1}\Big),\\ \sigma^2|\mathbf{P}_{\sigma^2} &\sim IG\big(\tau_1^* + K/2, (\beta - m^*)'\{V^*\}^{-1}(\beta - m^*)/2 + \tau_2^*\big),\end{aligned} \tag{11.7}$$

where τ_1^*, τ_2^*, m^*, and V^* are given by (11.7). Consequently, it is possible now to simulate from the posterior distribution, thanks to a reversible jump Metropolis–Hastings within Gibbs sampler algorithm. Precisely, from the likelihood (11.5) and the full conditional distributions (11.2), κ and $\mathbf{t}$ will be computed by the posterior mean from simulations using a reversible jump MCMC move and the coefficients β and σ^2 will be sampled from (11.2) using a Gibbs sampler move. The computation of the reversible jump MCMC scheme requires a likelihood ratio used in the acceptance-rejection probability. Let $(\mathbf{t}, \kappa)$ denote a current state in the reversible jumps move and $(\mathbf{t}^c, \kappa^c)$ denote a candidate state. For example, there is a submove in the reversible jump MCMC move that involves inserting one knot in the vector $\mathbf{t}$. For this type of transition from a current state $(\mathbf{t}, \kappa)$ to a candidate state $(\mathbf{t}^c, \kappa^c = \kappa + 1)$, we obtain the likelihood ratio by

$$\frac{L(\boldsymbol{y}|\mathbf{t}^c, \kappa^c)}{L(\boldsymbol{y}|\mathbf{t}, \kappa)} = (\tau_1^*)^{1/2}\frac{(\tau_2^*)^{(2\tau_1^*+K)/2}}{(\tau_2^{*c})^{(2\tau_1^*+K+1)/2}}\frac{|V^{*c}|^{1/2}}{|V^*|^{1/2}}\frac{|V|^{1/2}}{|V^c|^{1/2}} = (\tau_1^*)^{1/2}\frac{(\tau_2^*)^{(2\tau_1^*+K)/2}}{(\tau_2^{*c})^{(2\tau_1^*+K+1)/2}}, \tag{11.8}$$

where τ_2^{*c} is obtained by replacing $(\mathbf{t}^c, \kappa^c = \kappa + 1)$ in (11.7). We note the likelihood ration simplification $|V^{*c}|^{1/2}|V|^{1/2}/|V^*|^{1/2}|V^c|^{1/2} = 1$, thanks to the use of the g-prior (11.3).

11.3 Constrained Bayesian Inference

The free-knot unconstrained model discussed in Sect. 11.2 can be modified in a straightforward way to allow for constrained regression model. The shape constraints considered in this work will be derived by restricting the first and the second derivatives of (11.2) to be positive respectively to obtain monotonicity and convexity. The unimodality will be derived by restricting the first derivative of (11.2) to have exactly

one root $x^* \in]a, b[$. Suppose that it is intended to control the shape of the unknown function f, defined by (11.2), on an interval $I = [a_0, b_0] \subseteq [a, b]$. For all $x \in [a, b]$, denote by t_{j_x} the smallest knot greater than x. As known implicitly from the first use of the sequence $\mathbf{t}$, it will be convenient to assume that $t_j < t_{j+1}$ for all j, as it enables us to write that $t_{j_x-1} < x \leq t_{j_x}$ for all $x \in [a, b]$. Note that $B_{(j_{a_0}-k),\mathbf{t}}, \ldots, B_{j_{b_0}-1,\mathbf{t}}$ are the B-splines with order k and whose support intersects with $[a_0, b_0]$, that is $[t_j, t_{j+k}] \cap [a_0, b_0] \neq \emptyset$ for $j \in \{(j_{a_0} - k), \ldots, j_{b_0} - 1\}$. Thus, controlling the shape of f on I is reduced to control the shape of the restriction

$$f_{|I}(x) = \sum_{j=(j_{a_0}-k)}^{j_{b_0}-1} \beta_j B_{j,\mathbf{t}}(x), \qquad \text{for all } x \in I. \tag{11.9}$$

By noting $j_0 = (j_{a_0} - k)$ and from [3] (p. 117), it is known that the first derivative of (11.9) is $Df_{|I}(x) = (k-1)\sum_{j=j_0+1}^{j_{b_0}-1} \frac{\beta_j - \beta_{j-1}}{(t_{j+k-1}-t_j)} DB_{j,\mathbf{t}}(x)$ and the second derivative is $D^2 f_{|I}(x) = (k-1)(k-2)\sum_{j=j_0+2}^{j_{b_0}-1} \frac{\beta_j - 2\beta_{j-1}+\beta_{j-2}}{(t_{j+k-2}-t_j)(t_{j+k-1}-t_j)} D^2 B_{j,\mathbf{t}}(x)$. Thus, thanks to the positivity of B-spline functions, it is immediate that

(i) ($f_{|I}$ is monotone) If $\beta_{j_0} \leq \beta_{j_0+1} \leq \cdots \leq \beta_{j_{b_0}-1}$, then, for all $x \in [a_0, b_0]$ and $k \geq 2$ we have $Df_{|I}(x) \geq 0$.

(ii) ($f_{|I}$ is unimodal concave) Let $j_{b_0} - 1 \geq 3$. If $\beta_{j_0+1} - \beta_{j_0} > 0$, $\beta_{j_{b_0}-1} - \beta_{j_{b_0}-2} < 0$, $(\beta_j + \beta_{j-2}) \leq 2\beta_{j-1}$ for $j = j_0+2, \ldots, j_{b_0}-2$, then $Df_{|I}(a_0) > 0$, $Df_{|I}(b_0) < 0$ and $D^2 f_{|I}(x) \leq 0$ for all $x \in [a_0, b_0]$ and $k \geq 3$.

Then, it is clear that a natural way to impose local shape constraints on f is simply by conditioning the prior distribution on some sets. For example, if it is intended to impose a monotone shape constraint on the interval $[a_0, b_0]$, we have the set

$$S = \{f_I \,|\, \beta_j \in \mathbb{R}, \beta_{j_0} \leq \beta_{j_0+1} \leq \cdots \leq \beta_{j_{b_0}-1}\},$$

for unimodality concave restriction, we have the set

$$S = \cap_{j=j_0+2}^{j_{b_0}-1} \{f_I \,|\, \beta_j \in \mathbb{R}, \beta_{j_0} < \beta_{j_0+1}, \beta_{j_{b_0}-2} > \beta_{j_{b_0}-1}, \beta_j - 2\beta_{j-1} + \beta_{j-2} \leq 0\},$$

and for unimodality restriction, we have the set

$$\begin{aligned} S = \cup_{\ell=j_0+2}^{j_{b_0}-2} \{f_I | \beta_j \in \mathbb{R}, \\ \beta_{j_0} = \beta_{j_0+1} < \beta_{j_0+2} \leq \cdots \leq \beta_\ell \geq \beta_{\ell+1} \geq \cdots \geq \beta_{j_{b_0}-2} > \beta_{j_{b_0}-1}\}. \end{aligned}$$

Clearly, it is also easy to impose a succession of increasing, decreasing, concave, or convex parts and to locate the constraints at some parts of the x-axis. For reasons of simplicity, we denote by S the set of vectors $\beta^S = (\beta_j^S)_{j=1}^K$ such that f fulfill the constraint S. The g-prior of the unconstrained coefficients defined in Sect. 11.2 can be generalized in the following manner to handle any constraints: $\pi_{\beta^S}(\beta^S | \mathbf{t}, \kappa, \sigma^2) \propto \mathcal{N}_K\left(0, \sigma^2 n (B'_{\kappa,\mathbf{t}} B_{\kappa,\mathbf{t}})^{-1}\right) \mathbf{1}_{\beta^S \in S}$. Clearly, the prior density of β^S is simply proportional to the density of the unconditioned normal distribution (11.3)

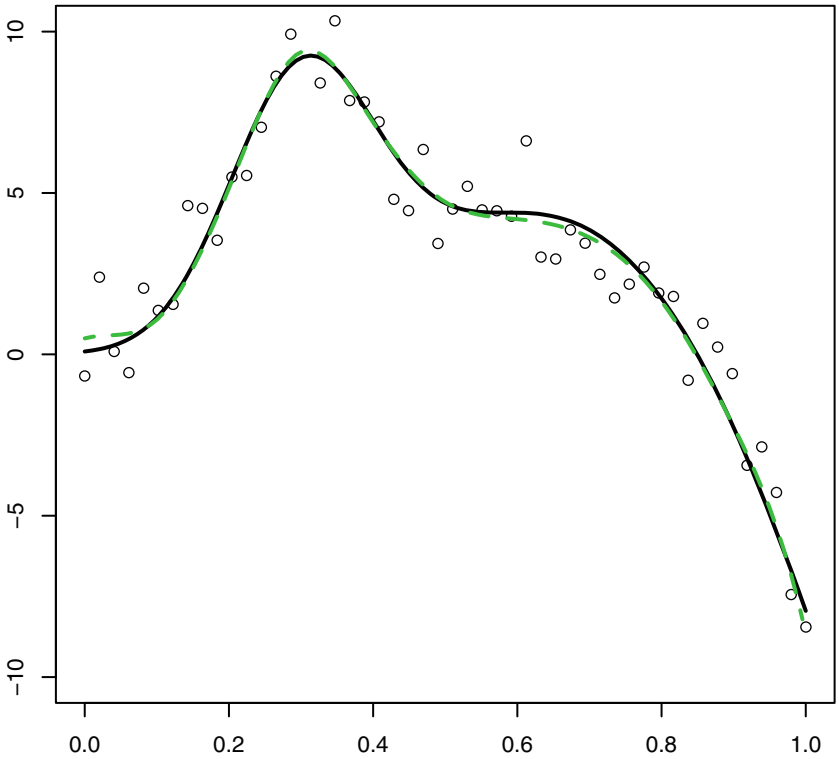

Fig. 11.1 Data (*circle*), true regression function (*solid curve*) and constrained estimate (- - -)

multiplied by the indicator function $\mathbf{1}_{\beta^S \in S}$. It is straightforward to check that the full conditional posterior distribution of β^S is $\mathcal{N}_K\big(\frac{n}{1+n}\widehat{\beta}, \sigma^2 V^*\big)\mathbf{1}_{\beta^S \in S}$, where the matrix V^* is obtained from the unconditioned case. For clarity, let us discuss the following example: f is monotone increasing function on $[0.1, 0.35]$, concave on $[0.3, 0.7]$, equal to 2.25 in 0.8, and no constraint elsewhere. Then, the monotone increasing constraint involves the B-splines $B_{j_{0.1}-k,\mathbf{t}}, \ldots, B_{j_{0.35}-1,\mathbf{t}}$ whose support intersects $[0.1, 0.35]$, while the concavity constraint involves the B-splines $B_{j_{0.3}-k,\mathbf{t}}, \ldots, B_{j_{0.7}-1,\mathbf{t}}$ and the value constraint involves only the B-spline $B_{j_{0.8}-1,\mathbf{t}}$ (it is necessary to consider $t_{j_{0.8}} = \cdots = t_{j_{0.8}+k-1} = 0.8$). Thus, the prior density of coefficients is proportional to $\mathcal{N}_K\big(0, \sigma^2 V\big)\mathbf{1}_{\{\beta^S \in S = S_{[0.1,0.35]} \cap S_{[0.3,0.7]} \cap S_{0.8}\}}$. By way of example, we generate data according to model (11.1) with a true regression function defined by $f(x) = 15x^2 \sin(3.7x) + 2\psi_{0.3,0.1}(x)$, $n = 50$, and $\sigma = 1$, where $\psi_{m,\sigma}$ denotes the normal density of the $\mathcal{N}(m, \sigma)$ distribution (see Fig. 20.1). We assume that it is known that f is unimodal on $[0, 1]$ (first increasing and then decreasing), concave on $[0.55, 1]$, and twice differentiable.

For the constrained Bayesian inference, it is now difficult to construct a trans-dimensional MCMC sampling scheme for the reason that simulations from the

posterior by a reversible jump Metropolis–Hastings algorithm require the exact knowledge of the prior density for β^S. Thus, it is necessary to find the normalizing constant of the truncated g-prior $\pi_{\beta^S}(\beta^S|\mathbf{t},\kappa,\sigma^2)$. As we are unable to compute this normalizing constant, we propose to use a simulated annealing step that ensures the projection of the coefficients on the constrained space. In this spirit, the reversible jump MCMC slice sampler coupled to the simulated annealing step perform well for any combination of shape constraints. To construct the MCMC algorithm for the free-knot model under combinations of shape restrictions, we use the unconstrained coefficients β with prior (11.3) to obtain the constrained coefficients β^S by projecting β in the constrained space S. By denoting $\|\cdot\|_2$ the Euclidean norm, there is a projection operator $\mathbf{P}$ such that $\mathbf{P}\beta = \beta^S$ where

$$\mathbf{P}\beta = \arg\min_{\tilde{\beta}\in S} \|\tilde{\beta}B_{\kappa,\mathbf{t}} - \beta B_{\kappa,\mathbf{t}}\|_2^2 = \arg\min_{\tilde{\beta}\in S} Q(\tilde{\beta}). \tag{11.10}$$

Our main idea from the use of the projection (11.10) is to make inference indirectly on β^S using unconstrained coefficients β where for each vector β we compute $\beta^S = \mathbf{P}\beta$ by solving an optimization problem of the form (11.10) using a simulated annealing step. The constraints can be enforcing, thanks to the proposal step of the simulated annealing move. It is clear that the use of the projection allows to avoid the computation of the normalizing constant of the truncated g-prior. In the sequel, we denote by $\widehat{\beta^S} = \mathbf{P}\widehat{\beta}$ and we aim to compute the likelihoods ratio $L^S(\mathbf{y}|\kappa^c,\mathbf{t}^c)/L^S(\mathbf{y}|\kappa,\mathbf{t})$ in the presence of constraints where

$$\begin{aligned} L^S(\mathbf{y}|\kappa,\mathbf{t}) &= \iint L^S(\mathbf{y}|\kappa,\mathbf{t},\beta^S,\sigma^2)\pi_{\beta^S}(\beta^S|\mathbf{t},\kappa,\sigma^2)\pi_{\sigma^2}(\sigma^2)d\beta^S d\sigma^2 \\ &= \iint L^S(\mathbf{y}|\kappa,\mathbf{t},\mathbf{P}\beta,\sigma^2)\pi_{\beta}(\beta|\mathbf{t},\kappa,\sigma^2)\pi_{\sigma^2}(\sigma^2)d\beta d\sigma^2, \end{aligned} \tag{11.11}$$

and $L^S(\mathbf{y}|\kappa,\mathbf{t},\mathbf{P}\beta,\sigma^2)$ is the constrained likelihood. Then, the constrained version of the likelihood (11.11) will be used to build a reversible jumps simulation chain with stationary distribution $\pi(\mathbf{t},\kappa|\mathbf{y})$. Furthermore, by using Laplace's method, we approximate the integration (11.11) and we can show that the constrained likelihoods ratio can be approximated with an error $O(n^{-1/2})$. For example, let us consider one transition in reversible jump simulations chain from a current state $(\kappa,\boldsymbol{t})$ to a candidate state $(\kappa^c=\kappa+1,\boldsymbol{t}^c)$, the constrained likelihoods ratio is approximated by

$$\frac{L^S(\mathbf{y}|\kappa^c,\boldsymbol{t}^c)}{L^S(\mathbf{y}|\kappa,\boldsymbol{t})} \simeq \frac{1}{\sqrt{n}}\Big(\frac{(\mathbf{y}-\widehat{\beta^S}B_{\kappa,\boldsymbol{t}})'(\mathbf{y}-\widehat{\beta^S}B_{\kappa,\boldsymbol{t}})}{(\mathbf{y}-\widehat{\beta^{Sc}}B^c_{\kappa,\boldsymbol{t}})'(\mathbf{y}-\widehat{\beta^{Sc}}B^c_{\kappa,\boldsymbol{t}})}\Big)^{n/2}, \tag{11.12}$$

where $\widehat{\beta^{Sc}} = \mathbf{P}\widehat{\beta^c}$ and $\widehat{\beta^c} = (B^{c\prime}_{\boldsymbol{t},k}B^c_{\boldsymbol{t},k})^{-1}B^{c\prime}_{\boldsymbol{t},k}\mathbf{y}$. The approximation (11.12) is obtained with an error $O(n^{-1/2})$. The result (11.12) is true for all subsets S of constraints that are considered in this chapter and it can be shown by application of Taylor's theorem and the method of Laplace (for more details we refer the reader to [26]).

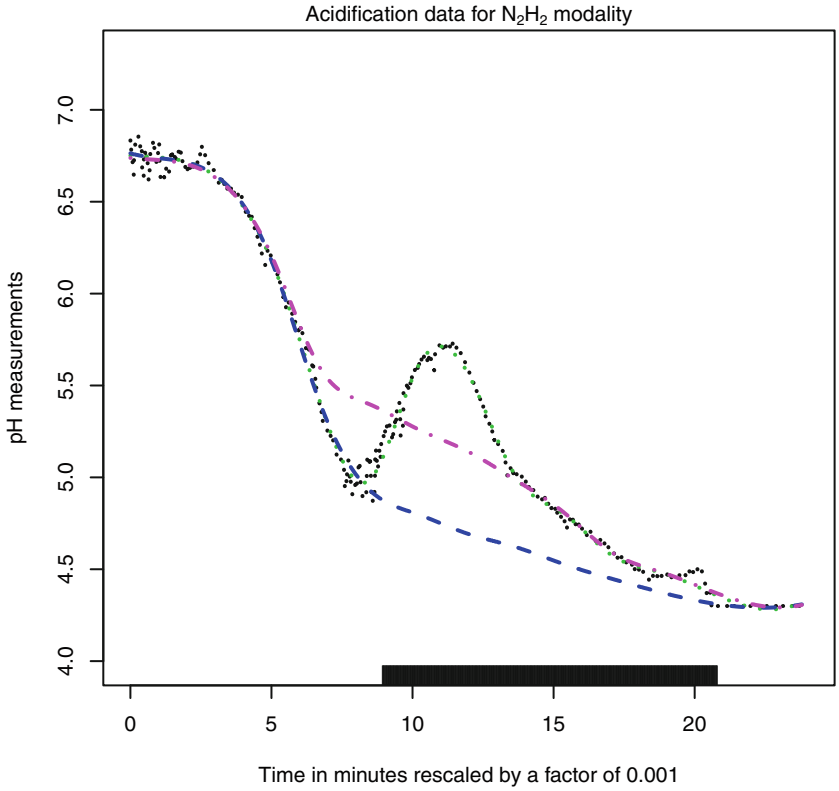

Fig. 11.2 The data: pH vs. time (mn) rescaled by a factor of 10^{-2}. Unconstrained estimate see in Sect. 11.2 ($\cdots$), constrained estimate calculated using all the data (- · -) and the local constrained estimate (under local decreasing constraint on the interval $[8.93, 20.8]$) calculated after removing erroneous measurements (- - -)

11.4 Numerical Experiment

This section outlines a numerical experiment through a real application to acidification kinetics. The control of acidification is an important issue in cheese-making. For instance, the influence of several environmental conditions on acidification kinetics has been studied in [13]. The acidification kinetics under N_2H_2 modality that is presented in Fig. 11.2 consist of $n = 1429$ measures of pH recorded alternatively at 1 and 2 min intervals. It is known, independently of the data, that the pH must be a decreasing function of the time during the experiment. Some errors may occur during the recording process because of the measuring device sensitivity to electrical interference problems. This is the case of data shown in Fig. 11.2: the pH is not a decreasing function on the interval $[8.93, 20.8]$. It is of interest to estimate what would be the true acidification curve if no measurement error had occurred. As shown in Fig. 11.2, we give the unconstrained estimate ($\cdots$), constrained estimate calculated

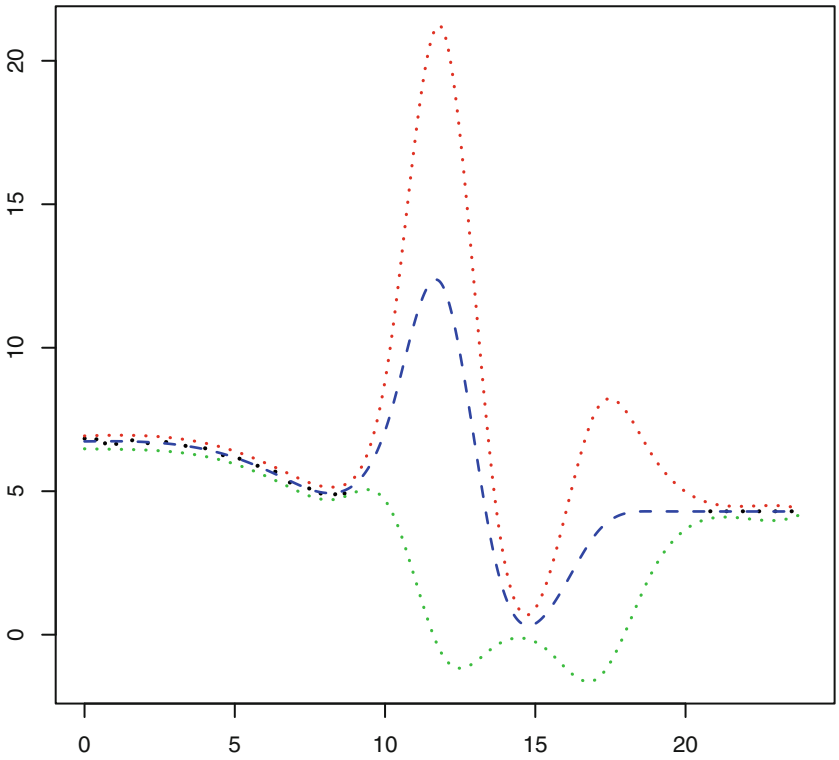

Fig. 11.3 Unconstrained estimate (- - -) seen in Sect. 11.2 calculated after removing erroneous measurements (without data on the interval [8.93, 20.8]) and 95% credibility interval (· · ·)

using all the dataset (- · -) and local constrained estimate (under local decreasing constraint on the interval [8.93, 20.8]) calculated after removing erroneous measurements (- - -). We can note that the constrained and the unconstrained estimates are identical on the interval of the x-axis where the constraints are fulfilled by both the constrained and the unconstrained estimates. We can also note that the constrained estimate calculated using all the dataset (- · -) achieves a compromise between erroneous data and decreasing constraint. In Fig. 11.3, we give the unconstrained estimate (- - -) seen in Sect. 11.2 calculated after removing erroneous measurements (without data on the interval [8.93, 20.8]) and 95 % credibility interval (· · ·). It is clear that unconstrained estimate present an imperfect behavior on the interval [8.93, 20.8]. Finally, we can note that the constrained estimate given in Fig. 11.4 is the most convincing one: it is obtained by removing erroneous measurements and enforcing a local decreasing constraint on [7.98, 20.8]. The only difference between the constrained estimate given in Fig. 11.4 and the constrained estimate (- - -) given in Fig. 11.2 lies in the choice of the position of the local shape constraint. To obtain the estimates, we run the MCMC sampler for 100,000 iterations and retain every 10-th point of the chains. From the 10,000 retained values of each parameter, we compute

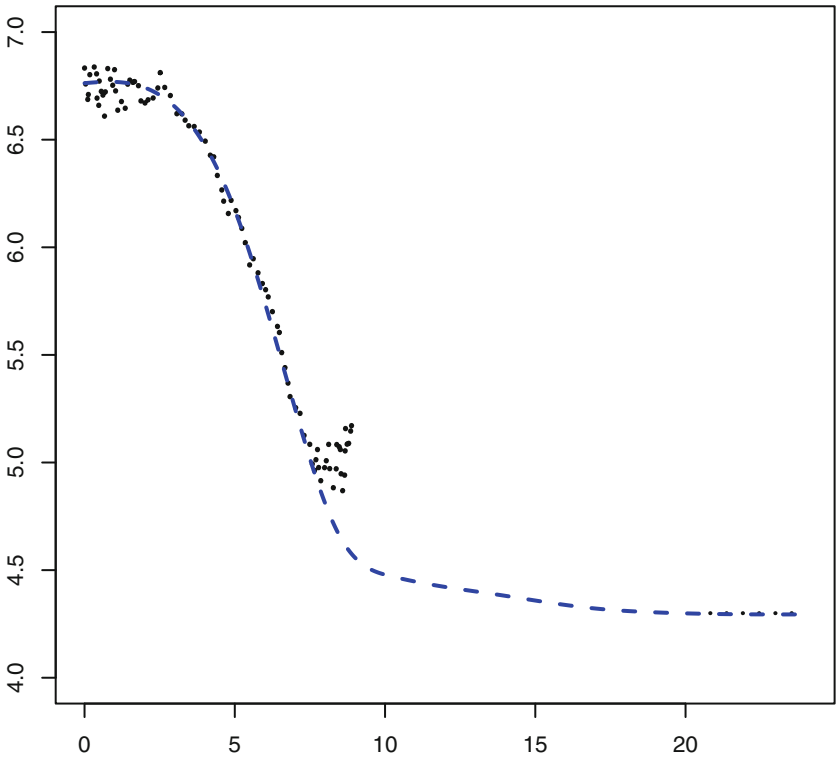

Fig. 11.4 The local constrained estimate (under local decreasing constraint on the interval [7.98, 20.8]) calculated after removing erroneous measurements (- - -)

a 0.05-credible set for β_l for every $l \in \{1, \ldots, K\}$. The computed 0.05-credible set for σ^2 is $[0.0206, 0.0239]$ and the posterior mean is 0.0222 (in the unconstrained case) is $[0.0157, 0.0179]$ and the posterior mean is 0.0168 (where a local decreasing constraint is enforced on $[8.93, 20.8]$) and is $[0.0154, 0.0176]$ and the posterior mean is 0.0165 (where a local decreasing constraint is enforced on $[7.98, 20.8]$).

11.5 Discussion

In this chapter, the Bayesian framework enables to impose local shape constraints on the regression function, thanks to the coefficients prior distribution. The local support of B-splines is an interesting property as a lack of data (or a lack of prior information) in a small region of the x-axis does not affect the inference on the whole definition domain of the independent variable. The performance of the constrained estimate compared to the unconstrained one is shown, thanks to a real numerical application.

Computation of the constrained estimate and simulations from the posterior distribution can be done with no additional difficulties and only requiring a coefficients prior density known up to a normalizing constant. However, the reversible jump MCMC move combined with a simulated annealing move allows only to search for a single optimized solution of the constrained problem and does not provide to obtain a credible set for the estimate. We expect that future work will address the problem of the construction of constrained prior distribution with known normalizing constant. Clearly, such a prior enables us to compute the prior ratio of the acceptance probability in the reversible jump MCMC scheme without resorting to the simulated annealing step. We also project in the future to detect automatically the interval of erroneous measurements seen in the pH data application.

Acknowledgments This article was mainly prepared while the author was working at UMR Mistea of SupAgro Montpellier. I would like to thank Christophe Abraham for his suggestion to study this topic and for his always accurate insights. The author is very grateful to the referees for many useful comments that improved the clarity of the chapter.

References

1. Abraham, C.: Bayesian regression under combinations of constraints. J. Stat. Plan. Inference **142**, 2672–2687 (2012)
2. Bartoli, N., Moral, P.D.: Simulation et algoritmes stochastiques. Cepadues-Editions (2001)
3. de Boor, C.: A Practical Guide to Splines. Springer-Verlag, New-York (2001)
4. Delecroix, M., Thomas-Agnan, C.: Spline and Kernel Regression under Shape Restrictions. Wiley, New-York (2000)
5. Denison, D., Mallick, B., Smith, A.: Automatic Bayesian curve fitting. J. Royal Stat. Soc. (B) **60**, 333–350 (1998)
6. DiMatteo, I., Genovese, C., Kass, R.: Bayesian curve-fitting with free-knot splines. Biometrika **88**, 1055–1071 (2001)
7. Gelfand, A., Kuo, L.: Nonparametric Bayesian bioassay including ordered polytomous response. Biometrika **78**, 355–366 (1991)
8. Gelfand, A., Smith, A., Lee, T.: Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. J. Am. Stat. Assoc. **87**, 523–532 (1992)
9. Green, P.: Reversible jump Markov chain monte carlo computation and Bayesian model determination. Biometrika **82**, 711–732 (1995).
10. Gunn, L., Dunson, D.: A transformation approach for incorporating monotone or unimodal constraints. Biostatistics **6**, 434–449 (2005)
11. Holmes, C., Heard, N.: Generalized monotonic regression using random change points. Stat. Med. **22**, 623–638 (2003)
12. Ibrahim, J., Chen, M.: Power prior distributions for regression models. Stat. Sci. **15**, 46–60 (2000)
13. Jeanson, S., Hilgert, N., Coquillard, M., Seukpanya, C., Faiveley, M., Neuveu, P., Abraham, C., Georgescu, V., Fourcassie, P., Beuvier, E.: Milk acidification by Lactococcus lactis is improved by decreasing the level of dissolved oxygen rather than decreasing redox potential in the potential in the milk prior to inoculation. Int. J. Food Microbiol. **131**, 75–81 (2009)
14. Lavine, M., Mockus, A.: A nonparametric Bayes method for isotonic regression. J. Stat. Plan. Inference **46**, 235–248 (1995)

15. Mammen, E., Thomas-Agnan, C.: Smoothing splines and shape restrictions. Scand. J. Stat. **26**, 239–252 (1999)
16. Mammen, E., Marron, J., Turlach, B., Wand, M.: A general projection framework for constrained smoothing. Stat. Sci. **16**, 232–248 (2001)
17. Meyer, M.: Inference using shape-restricted regression splines. Ann. Appl. Stat. **2**, 1013–1033 (2008)
18. Mukerjee, H.: Monotone nonparametric regression. Ann. Stat. **16**, 741–750 (1988)
19. Neelon, B., Dunson, D.: Bayesian isotonic regression and trend analysis. Biometrics **60**, 398–406 (2004)
20. Ramgopal, P., Laud, P., Smith, A.: Nonparametric Bayesian bioassay with prior constraints on the shape of the potency curve. Biometrika **80**, 489–498 (1993)
21. Ramsay, J.: Estimating smooth monotone functions. J. Royal Stat. Soc. (B) **60**, 365–375 (1998)
22. Shively, T., Sager, T.: A Bayesian approach to non-parametric monotone function estimation. J. Royal Stat. Soc. (B) **71**, 159–175 (2009)
23. Shively, T., Walker, S., Damien, P.: Nonparametric function estimation subject to monotonicity, convexity and other shape constraints. J. Econom. **161**, 166–181 (2011)
24. Tierney, L.: Markov chains for exploring posterior distributions. Ann. Stat. **22**, 1701–1762 (1994)
25. Villalobos, M., Wahba, G.: Inequality constrained multivariate smoothing splines with application to the estimation of posterior probability. J. Am. Stat. Assoc. **82**, 239–248 (1987)
26. Wang, X.: Bayesian free-knot monotone cubic spline regression. J. Comput. Graphical Stat. **17**, 373–387 (2008)
27. Wright, I., Wegman, E.: Nonparametric regression under qualitative smoothness assumptions. Ann. Stat. **8**, 1023–1035 (1980)

Chapter 12
A Bayesian Approach to Predicting Football Match Outcomes Considering Time Effect Weight

Francisco Louzada, Adriano K. Suzuki, Luis E. B. Salasar, Anderson Ara and José G. Leite

Abstract In this chapter we propose a simulation-based method for predicting football match outcomes. We adopt a Bayesian perspective, modeling the number of goals of two opposing teams as a Poisson distribution whose mean is proportional to the relative technical level of opponents. Fédération Internationale de Football Association (FIFA) ratings were taken as the measure of technical level of teams saw well as experts' opinions on the scores of the matches were taken in account to construct the prior distributions of the parameters. Tournament simulations were performed in order to estimate probabilities of winning the tournament assuming different values for the weight attached to the experts' information and different choices for the sequence of weights attached to the previous observed matches. The methodology is illustrated on the 2010 Football Word Cup.

12.1 Introduction

Predicting outcomes of football games has been the focus of research of several researchers, mostly applied to championship leagues. For instance, Keller [8] has fitted the Poisson distribution to the number of goals scored by England, Ireland,

F. Louzada (✉) · A. K. Suzuki
Institute of Mathematics and Computer Science, University of São Paulo—USP, Avenida Trabalhador São-carlense, 400 - Centro, São Carlos 13566-590, SP, Brazil
e-mail: louzada@icmc.usp.br

A. K. Suzuki
e-mail: suzuki@icmc.usp.br

L. E. B. Salasar · A. Ara · J. G. Leite
Departamento de Estatística, Universidade Federal de São Carlos, Rod. Washington Luiz, km 235, São Carlos, SP 13565-905, Brazil
e-mail: luis.salasar@gmail.com

A. Ara
e-mail: alsouzaara@gmail.com

J. G. Leite
e-mail: leite@ufscar.br

A. Polpo et al. (eds.), *Interdisciplinary Bayesian Statistics,*
Springer Proceedings in Mathematics & Statistics 118, DOI 10.1007/978-3-319-12454-4_12

Scotland, and Wales in the British International Championship from 1883 to 1980. Also, Lee [9] considers the Poisson distribution, but allows for the parameter to depend on a general home-ground effect and individual offensive and defensive effects. Moreover, Karlis [7] applied the Skellam's distribution to model the goal difference between home and away teams. The authors argue that this approach does not rely neither on independence nor on the marginal Poisson distribution assumptions for the number of goals scored by the teams. A Bayesian analysis for predicting match outcomes for the English Premiere League (2006–2007 season) is carried out using a log-linear link function and noninformative prior distributions for the model parameters.

Taking another approach, Brillinger [1] proposed to model directly the win, draw, and loss probabilities by applying a trinomial regression model to the Brazilian 2006 Series A championship. By means of simulation, it is estimated for each team the total points, the probability of winning the championship, and the probability of ending the season in the top four places.

In spite of the vast literature directed to League Championship prediction, few articles concern score predictions for the World Cup tournament (WCT) [5, 12, 13]. The WCT is organized by *Fédération Internationale de Football Association* (FIFA, French for International Federation of Football Association), occuring every 4 years. Probably, the shortage of researches on the WCT is due to the limited amount of valuable data related to international matches and also to the fact that few competitions confronts teams from different continents.

A log-linear Poisson regression model which takes the FIFA ratings as covariates is presented by Dyte and Clarke [5]. The authors present some results on the predictive power of the model and also present simulation results to estimate probabilities of winning the championship for the 1998 WCT. Volf [13], using a counting processes approach, modeled the development of a match score as two interacting time-dependent random point processes. The interaction between teams are modeled via a semiparametric multiplicative regression model of intensity. The author has applied his model to the analysis of the performance of the eight teams that reached the quarter-finals of the 2006 WCT. Suzuki et al. [12] proposed a Bayesian methodology for predicting match outcomes using experts' opinions and the FIFA ratings as prior information. The method is applied to calculate the win, draw, and loss probabilities for each match and also to estimate classification probabilities in group stage and winning tournament chances for each team on the 2006 WCT.

In this chapter, we proposed a Bayesian method for predicting match outcomes with use of the experts' opinions and the FIFA ratings as prior information, but differently from [12], for all 48 matches of the first phase (group stage in which each team played three matches) the experts' opinions were provided before the beginning of 2010 WCT. The motivation for such purpose was the difficulty to get the experts' opinions (four sportswriters contributed with their opinions) at the end of each round of group stage. The drawback is that these matches were played within 15 days, mostly in different dates and times. For the second phase, the experts' opinions were provided before each round. Moreover, we incorporate a time-effect weights for the matches, that is, we consider that outcomes of matches which were played

first have less importance than the outcomes of more recent matches. An attractive advantage of our approach is the possibility of calibrating the experts' opinions as well as the importance of previous match outcomes in the modeling, directing for a control on the model prediction capability. Considering a grid of values for the experts' opinions weight a_0 and for the last matches importance, the $p_i's$ values, we can assess the impact of these weights on the model prediction capability.

We used the predictive distributions to perform a simulation based on 10,000 runs of the whole competition, with the purpose of estimating various probability measures of interest, such as the probability that a given team wins the tournament, reaches the final, qualifies to the knockout stage and so on.

The chapter is outlined as follows. In Sect. 12.2, we present the probabilistic model and expressions for priors and posterior distribution of parameters, as well as for predictive distributions. In Sect. 12.3, we present the method used to estimate the probabilities of winning the tournament. In Sect. 12.4, we give our final considerations about the results and further work.

12.2 Probabilistic Model

In the current format, the WCT gathers 32 teams, where the host nation(s) has a guaranteed place and the others are selected from a qualifying phase which occurs in the 3-year period preceding the tournament. The tournament is composed by a *group stage* followed by a *knockout stage*. In the *group stage*, the teams play against each other within their group and the top two teams in each group advance to the next stage. In the *knockout stage*, 16 teams play one-off matches in a single-elimination system, with extra time of 30 min (divided in 2 halves of 15 min each) and penalty shootouts used to decide the winners when necessary.

The probabilistic model is derived as follows. Consider a match between teams A and B with respective FIFA ratings R_A and R_B. In the following we shall assume that, given the parameters λ_A and λ_B, the number of goals X_{AB} and X_{BA} scored by team A and B, respectively, are two independent random variables with

$$X_{AB} \mid \lambda_A \sim \text{Poisson}\left(\lambda_A \frac{R_A}{R_B}\right), \tag{12.1}$$

$$X_{BA} \mid \lambda_B \sim \text{Poisson}\left(\lambda_B \frac{R_B}{R_A}\right). \tag{12.2}$$

In this model, the ratings are used to quantify each team's ability and the mean number of goals A scores against B is directly proportional to team A's rating and inversely proportional to team B's rating. If A and B have the same ratings ($R_A = R_B$), then the mean score for that match is (λ_A, λ_B). So, the parameter λ_A can be interpreted as the mean number of goals team A scores against a team with the same ability and an analogous interpretation applies to λ_B.

We first consider the prior distribution formulation. In order to formulate the prior distribution, a number of experts provide their expected final scores for the incoming

matches which we intend to predict. This kind of elicitation procedure is natural and simple, since the model parameters are directly related to the number of goals and the requested information is readily understandable by the respondents not requiring any extra explanation. We have adopted multiple experts since we believe it aggregates more information than using only one expert.

Assuming independent experts' opinions and following a Poisson distribution, we shall obtain the prior distribution for the parameters using a procedure analogous to the power prior method [2] with the historical data replaced by the experts' expected scores. The proposed elicitation process is based on the assumption that the experts are able to provide plausible outcomes for the incoming matches that could be observed but in fact were not. This elicitation process is in accordance with the Bayesian paradigm for prior elicitation as discussed in [6] and [3] as we will see later in this section. Moreover, although the independence assumption is taken mainly because of mathematical simplicity, we can argue that, at least approximately, the independence assumption holds in our case since the selected experts work at different media and do not maintain any contact.

Suppose we intend to predict a match between teams A and B. Consider that s experts provide their expected scores for m incoming matches of team A and B. Denote by $y_{i,j}$ the jth expert's expected number of goals scored by team A against opponent OA_i and by $z_{i,j}$ the expected number of goals scored by team B against opponent OB_i, $i = 1, \ldots, m$, $j = 1, \ldots, s$.

In the following, we shall assume the probability density functions as the initial information about the parameters given by

$$\pi_0(\lambda_A) \propto \lambda_A^{\delta_0 - 1} \exp\{-\beta_0 \lambda_A\} \text{ and } \pi_0(\lambda_B) \propto \lambda_B^{\delta_0 - 1} \exp\{-\beta_0 \lambda_B\}, \tag{12.3}$$

where $\delta_0 > 0$ and $\beta_0 \geq 0$. Note that if $\delta_0 = 1/2$ and $\beta_0 = 0$ we have the Jeffreys' prior for the Poisson model, if $\delta_0 = 1$ and $\beta_0 = 0$ we have an uniform distribution over the interval $(0, \infty)$ and if $\delta_0 > 0$ and $\beta_0 > 0$ we have a proper gamma distribution. The first two cases are usual choices to represent noninformative distribution for the parameter.

Updating this initial prior distribution with the experts' expected scores, we obtain the power prior of λ_A

$$\pi\left(\lambda_A | \mathcal{D}_0\right) \propto \lambda_A^{a_0 \sum_{i=1}^{m} \sum_{j=1}^{s} y_{i,j} + \delta_0 - 1} \exp\left\{ - \left(a_0 s \sum_{i=1}^{m} \frac{R_A}{R_{OA_i}} + \beta_0 \right) \lambda_A \right\}, \tag{12.4}$$

where $0 \leq a_0 \leq 1$ represents a "weight" given to experts' information and $\mathcal{D}_0$ denotes all the experts' expected scores. Thus, if $0 < a_0 \leq 1$, the prior distribution of λ_A is

$$\lambda_A | \mathcal{D}_0 \sim \text{Gamma}\left(a_0 \sum_{i=1}^{m} \sum_{j=1}^{s} y_{i,j} + \delta_0, a_0 s \sum_{i=1}^{m} \frac{R_A}{R_{OA_i}} + \beta_0 \right),$$

and if $a_0 = 0$ the prior for λ_A is the initial Jeffreys' prior (12.3) and corresponds to disconsider all the experts' information. In particular, if $a_0 = 1$ the prior for λ_A

equals to the posterior which would be obtained if all the expected scores were in fact real data. Thus, the a_0 parameter can be interpreted as a degree of confidence in the experts' information.

The elicitation of the prior distribution (12.4) can also be viewed in the light of the Bayesian paradigm of elicitation [6], when we consider the likelihood for the experts' information

$$L'(\lambda_A|y_{1,1},\ldots,y_{m,s}) \propto \prod_{i=1}^{m}\prod_{j=1}^{s}\left[\lambda_A^{y_{i,j}}\exp\left\{-\lambda_A\frac{R_A}{R_{OA_i}}\right\}\right]^{a_0} \tag{12.5}$$

and combine it with the initial noninformative prior distribution (12.3) by applying the Bayes theorem. The likelihood (12.5) provide information for the parameter through the experts' information, which are treated like data, i.e, we assume them to follow a Poisson distribution.

Analogously, the prior distribution of λ_B is

$$\pi\left(\lambda_B|\mathcal{D}_0\right) \propto \lambda_B^{a_0\sum_{i=1}^{m}\sum_{j=1}^{s}z_{i,j}+\delta_0-1}\exp\left\{-\left(a_0 s\sum_{i=1}^{m}\frac{R_B}{R_{OB_i}}+\beta_0\right)\lambda_B\right\}. \tag{12.6}$$

It is important to note that by the way the experts present their guesses there is possibility of contradictory information. There is some literature on prior elicitation of group opinions directed towards to remove such inconsistencies. According to O'Hagan et al. [10], they range from informal methods, such as Delphi method [11], which encourage the experts to discuss the issue in the hope of reaching consensus, to formal ones, such as weighted averages, opinion polling, or logarithmic opinion pools. For a review of methods of pooling expert opinions see [6].

Now the posterior and predictive distributions are presented. Our interest is to predict the number of goals that team A scored against team B, using all the available information (hereafter denoted by $\mathcal{D}$). This information is originated from two sources: the experts' expected score and the actual scores of matches already played. So, we may be in two distinct situations: (i) we do have the experts' information but no matches have been played, and (ii) we have both the experts' opinions and the scores of played matches.

In situation (i), we only have the experts' information. So, from the model (12.1) and the prior distribution (12.4), it follows that the prior predictive distribution of X_{AB} is

$$X_{AB} \sim \text{NB}\left(a_0\sum_{i=1}^{m}\sum_{j=1}^{s}y_{i,j}+\delta_0,\ \frac{a_0 s\sum_{i=1}^{m}\frac{R_A}{R_{OA_i}}+\beta_0}{a_0 s\sum_{i=1}^{m}\frac{R_A}{R_{OA_i}}+\frac{R_A}{R_B}+\beta_0}\right), \qquad k=0,1,\ldots, \tag{12.7}$$

where $NB(r,\gamma)$ denotes the negative binomial distribution with probability function given by

$$f(k;r,\gamma)=\frac{\Gamma(r+k)}{k!\Gamma(r)}(1-\gamma)^k\gamma^r,\qquad k=0,1,\ldots,$$

with parameters $r>0$ and $0<\gamma<1$.

Analogously, from model (12.2) and the prior distribution (12.6), it follows that the prior predictive distribution of X_{BA} is given by

$$X_{BA}\sim \text{NB}\left(a_0\sum_{i=1}^{m}\sum_{j=1}^{s}z_{i,j}+\delta_0,\frac{a_0 s\sum\limits_{i=1}^{m}\frac{R_B}{R_{OB_i}}+\beta_0}{a_0 s\sum\limits_{i=1}^{m}\frac{R_B}{R_{OB_i}}+\frac{R_B}{R_A}+\beta_0}\right),\qquad k=0,1,\ldots. \tag{12.8}$$

In situation (ii), assume that team A has played k matches, the first against team C_1, the second against team C_2, and so on until the kth match against team C_k. Suppose also that, given λ_A, $X_{A,C_1},\ldots,X_{A,C_k}$ are independent Poisson random variables with parameters $\lambda_A\frac{R_A}{R_{C_1}},\ldots,\lambda_A\frac{R_A}{R_{C_k}}$. Hence, from model (12.1) it follows that the weighted likelihood is given by

$$L^{\mathbf{p}}(\lambda_A|\mathcal{D})=\prod_{i=1}^{k}P\left[X_{A,C_i}=x_A^i\right]^{p_i}\propto\exp\left\{-\lambda_A\sum_{i=1}^{k}p_i\frac{R_A}{R_{C_i}}\right\}\lambda_A^{\sum\limits_{i=1}^{k}x_A^i p_i}, \tag{12.9}$$

where x_A^i are the number of goals scored by A against the ith opponent, $i=1,\ldots,k$, and $\mathbf{p}=(p_1,\ldots,p_k)$, $0<p_i<1$, is the vector of fixed weights assigned to each match in order to decrease the influence of past matches.

From the likelihood (12.9) and the prior distribution (12.4), it follows that the posterior distribution of λ_A is

$$\lambda_A|\mathcal{D}\sim\text{Gamma}\left(a_0\sum_{i=1}^{m}\sum_{j=1}^{s}y_{i,j}+\sum_{l=1}^{k}p_l x_A^l+\delta_0,\sum_{l=1}^{k}p_l\frac{R_A}{R_{C_l}}+a_0 s\sum_{i=1}^{m}\frac{R_A}{R_{OA_i}}+\beta_0\right), \tag{12.10}$$

which implies by the model (12.1) that the posterior predictive distribution of X_{AB} is

$$X_{AB}|\mathcal{D}\sim\text{NB}\left(a_0\sum_{i=1}^{m}\sum_{j=1}^{s}y_{i,j}+\sum_{l=1}^{k}p_l x_A^l+\delta_0,\frac{\sum\limits_{l=1}^{k}p_l\frac{R_A}{R_{C_l}}+a_0 s\sum\limits_{i=1}^{m}\frac{R_A}{R_{OA_i}}+\beta_0}{\sum\limits_{l=1}^{k}p_l\frac{R_A}{R_{C_l}}+a_0 s\sum\limits_{i=1}^{m}\frac{R_A}{R_{OA_i}}+\frac{R_A}{R_B}+\beta_0}\right). \tag{12.11}$$

Analogously, the posterior distribution of λ_B is given by

$$\lambda_B|\mathcal{D}\sim\text{Gamma}\left(a_0\sum_{i=1}^{m}\sum_{j=1}^{s}z_{i,j}+\sum_{l=1}^{k}p_l x_B^l+\delta_0,\sum_{l=1}^{k}p_l\frac{R_B}{R_{D_l}}+a_0 s\sum_{i=1}^{m}\frac{R_B}{R_{OB_i}}+\beta_0\right), \tag{12.12}$$

where x_B^l is the number of goals team B scores against the lth opponent, $D_l, l = 1, \ldots, k$. Hence, from the model (12.2) and the posterior (12.12), it follows that the posterior predictive distribution of X_{BA} is

$$X_{BA}|\mathcal{D} \sim \text{NB}\left(a_0 \sum_{i=1}^{m}\sum_{j=1}^{s} z_{i,j} + \sum_{l=1}^{k} p_l x_B^l + \delta_0, \frac{\sum_{l=1}^{k} p_l \frac{R_B}{R_{D_l}} + a_0 s \sum_{i=1}^{m} \frac{R_B}{R_{OB_i}} + \beta_0}{\sum_{l=1}^{k} p_l \frac{R_B}{R_{D_l}} + a_0 s \sum_{i=1}^{m} \frac{R_B}{R_{OB_i}} + \frac{R_B}{R_A} + \beta_0} \right). \tag{12.13}$$

At this point, it is important to note that matches taken to construct the prior distribution are distinct from those considered to the likelihood function, that is, the matches already played have their contribution (though their final scores) included in the likelihood function but not in the prior.

12.3 Methods

In this section, we shall consider the competition divided into seven rounds, where the first three rounds are in the group stage (first phase) and the last four in the knockout stage (second phase). The four experts' were asked for their expected final scores for the matches in five distinct times: just before the beginning of tournament and just before each of the four rounds in the knockout stage. At the beginning of competition, experts provided their expected final scores for all matches in the group stage at once, while in the knockout stage they provide their expected final scores only for matches in the incoming round. To account for the mean experts' opinion, we have chosen $a_0 = 1/(3 * 4) = 1/12$ (for the group stage) and $a_0 = 1/4$ (for the knockout stage), in the sense that the posterior distribution of the parameter is the same as that, which would be obtained if we took one observation equal to the mean expected score from the sampling distribution. It is important to note that the experts were selected from different sports media, in order to make their guesses as independent as possible.

For the knockout stage, teams can play for additional 30 min if they remain level after the 90 min regulation time and if the result persists the teams proceed to a penalty shootout decision. For the extra time, we considered a Poisson distribution with parameter multiplied by one third to account for the shrinkage of time, which is equivalent to 30 min of extra time. That is, one third of the overall match time (90 min). For penalty shootout, we simulated a Bernoulli random variable proportional to the ratio of parameter estimates (posterior means).

The exact calculation of probabilities is possible just for the case of a single match prediction. We calculate the probabilities exactly from the predictive distributions. The probabilities regarding qualifying chances, winning tournament chances among others must be performed by simulation, since they may involves many combinations of match results.

Table 12.1 Relative differences (in %) of the De Finetti measure of our pool of experts fixing a_0 at 0 and 0.25, relatively to a fictitious pool of perfect experts

Round	1st	2nd	3th	4th	5th	6th	7th
Pool of perfect experts	0.59	0.59	0.47	0.43	0.55	0.55	0.47
Pool of experts with $a_0 = 0.25$ (in %)	10.67	7,69	34.29	9.22	16,43	8.49	3.41
Pool of experts with $a_0 = 0$ (in %)	143.74	36.68	36.73	18.83	17.81	14.20	7,94

A method used to measure the goodness of a prediction is to calculate the De Finetti distance [4] which is the square of the Euclidean distance between the point corresponding to the outcome and the one corresponding to the prediction. It is useful to consider the set of all possible forecasts given by the simplex set

$$S = \left\{(P_W, P_D, P_L) \in [0,1]^3 : P_W + P_D + P_L = 1\right\}.$$

Observe that the vertices $(1,0,0)$, $(0,1,0)$, and $(0,0,1)$ of S represent the outcomes win, draw, and loss, respectively. Thus, if a prediction is $(0.2, 0.65, 0.15)$ and the outcome is a draw $(0,1,0)$, then the De Finetti distance is $(0.2-0)^2 + (0.65-1)^2 + (0.15-0)^2 = 0.185$. Also, we can associate to a set of predictions the average of its De Finetti distances, known as the De Finetti measure. So, we shall consider the best among some prediction methods the one with the least De Finetti measure.

To assess the impact of the experts' information on the quality of the predictions, Table 12.1 displays the relative differences (in %) of the De Finetti measure of our pool of experts fixing a_0 at 0 and 0.25, denoting respectively the total absence of experts' opinion and the amount of experts' information as considered in our method, relatively to a fictitious pool of "perfect" experts, who always forecast the exact score for each one of the matches, with a_0 fixed 0.25.

At the initial rounds, the use of experts' information greatly improves prediction, with the De Finetti measure with $a_0 = 0.25$ always closer to the results of the "perfect" experts' opinion. However, as observed data enters the model, with the progress of the competition, the gain of using expert information is decreasing. This feature is in fully agreement with our initial motivation to consider experts' information in our modeling: filling the lack of information when there is shortage of objective information (data) available. Note that De Finetti measure without considering the experts' opinion ($a_0 = 0$) in the first and second round is much larger than such measure for an equiprobable predictor, which assigns equal probability to all outcomes. This is another evidence in favor of the usefulness of our modeling. Our model also joined a prediction model competition for the matches of the group stage of the 2010 WCT organized by the Brazilian Society of Operational Research, the World Cup 2010 Football Forecast Competition, reaching the first place. The inclusion of subjective information into our modeling through expert's opinions was crucial for such achievement. For the knockout stage matches, experts' information did not improve prediction, which can be explained in part by the small difference of skill level between teams and by the lack of confidence on Spain and Netherlands teams who defeated the traditional teams of Germany and Brazil, respectively.

12.3.1 Predictions for the Whole Tournament

We use the predictive distributions to perform a simulation of 10,000 replications of the whole competition. The purpose of this simulation is to estimate probabilities that a given team wins the tournament. The probabilities are estimated by the percentage of times the event

Considering only the four finalists (Spain, Netherlands, Germany, and Uruguay) we obtained, just before each round of the knockout stage, the winning tournament probabilities assuming different values for the a_0 weight attached to the experts' information and different choices for the sequence of p_i's weights attached to the previous observed matches. Table 12.2 presents the obtained results for this simulation study. Various different time weighting values were considered within the range (0, 1), including someone that, from the practical point of view, may not make sense, but allows us to realize the impact of the p_i's in the winning tournament probabilities. From these results, we can observe that, for a fixed value of a_0, different values for the p_i's does not alter significantly the predictions, particularly in the advanced stages of the championship. On the other hand, for fixed values of the p_i's, we see a noticeable influence on predictions according to changes in the a_0 value. For instance, observe the probabilities for Netherlands and Germany in the semifinals, and Netherlands and Spain in the quarterfinals.

12.4 Final Remarks

In this chapter, we propose a Bayesian simulation methodology for predicting match outcomes of the 2010 Football World Cup, which makes use of the FIFA ratings and experts' opinions. FIFA ratings system are based on previous 4-year performance of teams. The drawbacks of this ratings system is the great changes in the formation of teams in such a large period of time and the small number of games played between teams of different continent in comparison with those played by teams of the same continent. Other measures of strength of teams should be considered further and compared with the FIFA ratings. Moreover, the development of two ratings for teams, possibly via experts' information or even FIFA documentation, one for attack and another for defense, could improve prediction since there are teams which have strong defense but weak attack and vice versa. A simple possibility to incorporate those abilities would be to add one parameter for each team directly on the Poisson model, as it was made for instance in [13]. This major embedding can be seen as a direct generalization of our modeling and should be considered in future research in the field.

The prior distributions are updated every round, providing flexibility to the modeling, once the experts' opinion are influenced by all previous events to the match. The use of expert's opinions may compensate, at least in part, for the lack of information of the factors which can influence a football match during a competition, such as tactic disciplines, team psychological conditions, referee, player injured or suspended, amongst others.

Table 12.2 Percentage of tournament wins

	Team	p_i's	a_0						
			0	0.10	0.25	0.50	0.75	0.90	1.00
Before round of sixteen	Spain	(0.10, 0.25, 1.00)	17.44	13.98	10.71	7.92	5.96	5.39	5.40
		(0.25, 0.50, 1.00)	15.04	12.77	10.01	7.54	5.90	5.68	5.41
		(0.50, 0.75, 1.00)	12.77	11.09	9.22	6.68	5.96	5.23	5.33
		(0.80, 0.90, 1.00)	10.86	9.58	8.56	6.94	6.09	5.69	5.03
		(1.00, 1.00, 1.00)	9.98	9.43	7.56	6.31	5.39	5.23	4.94
	Netherlands	(0.10, 0.25, 1.00)	9.99	12.63	15.41	18.62	19.99	19.46	20.43
		(0.25, 0.50, 1.00)	9.64	11.54	14.24	16.53	18.46	19.17	19.16
		(0.50, 0.75, 1.00)	9.47	10.90	13.71	15.79	17.42	18.62	17.91
		(0.80, 0.90, 1.00)	9.84	11.72	12.95	15.86	17.60	17.73	18.43
		(1.00, 1.00, 1.00)	10.02	11.66	13.51	16.07	17.12	18.18	18.46
	Germany	(0.10, 0.25, 1.00)	2.93	3.15	3.93	4.01	4.58	4.47	4.74
		(0.25, 0.50, 1.00)	3.94	4.41	4.46	4.81	4.81	4.74	4.74
		(0.50, 0.75, 1.00)	5.67	5.90	5.75	5.88	5.55	5.79	6.09
		(0.80, 0.90, 1.00)	8.38	8.46	7.38	7.44	6.68	6.59	7.20
		(1.00, 1.00, 1.00)	9.95	9.50	9.49	7.94	7.72	7.56	7.14
	Uruguay	(0.10, 0.25, 1.00)	3.03	5.23	6.94	8.62	9.85	10.96	10.54
		(0.25, 0.50, 1.00)	4.33	5.28	6.48	8.85	10.17	9.73	10.70
		(0.50, 0.75, 1.00)	4.00	5.19	6.40	8.33	8.78	8.93	9.35
		(0.80, 0.90, 1.00)	3.54	4.75	5.54	7.17	8.33	8.85	8.47
		(1.00, 1.00, 1.00)	3.61	4.00	5.04	6.43	7.81	7.89	8.02

Table 12.2 (continued)

	Team	p_i's	a_0						
Before quarter-finals	Spain	(0.10, 0.25, 0.50, 1.00)	16.84	18.70	20.31	20.36	20.64	20.57	21.44
		(0.25, 0.50, 0.75, 1.00)	18.12	18.95	20.06	21.29	20.73	20.70	21.47
		(0.70, 0.80, 0.90, 1.00)	15.69	16.33	17.71	18.86	19.80	20.09	19.99
		(1.00, 1.00, 1.00, 1.00)	15.22	15.75	16.41	18.00	18.36	18.42	19.25
	Netherlands	(0.10, 0.25, 0.50, 1.00)	9.98	8.18	7.18	5.28	4.49	4.18	3.83
		(0.25, 0.50, 0.75, 1.00)	10.20	8.76	7.35	6.27	5.15	4.68	4.36
		(0.70, 0.80, 0.90, 1.00)	11.38	10.19	8.89	7.23	5.99	5.31	5.64
		(1.00, 1.00, 1.00, 1.00)	11.25	10.57	9.28	7.60	6.28	6.44	6.01
	Germany	(0.10, 0.25, 0.50, 1.00)	18.29	17.92	17.20	16.73	16.40	16.39	16.32
		(0.25, 0.50, 0.75, 1.00)	15.17	16.28	15.40	15.23	15.71	15.56	16.16
		(0.70, 0.80, 0.90, 1.00)	16.64	17.40	15.98	16.65	16.37	16.77	16.37
		(1.00, 1.00, 1.00, 1.00)	17.26	17.69	17.35	16.70	17.46	17.13	16.53
	Uruguay	(0.10, 0.25, 0.50, 1.00)	6.48	6.73	6.39	6.42	6.60	6.70	6.97
		(0.25, 0.50, 0.75, 1.00)	7.39	7.38	6.91	6.95	6.66	7.02	6.86
		(0.70, 0.80, 0.90, 1.00)	7.08	6.95	6.87	6.95	6.51	6.56	6.49
		(1.00, 1.00, 1.00, 1.00)	6.60	6.81	6.41	6.70	6.98	6.28	6.36
Before semi-finals	Spain	(0.10, 0.25, 0.50, 0.75, 1.00)	14.57	15.05	15.09	15.59	15.56	15.57	15.41
		(0.60, 0.70, 0.80, 0.90, 1.00	16.32	16.58	16.61	17.26	16.38	16.34	16.53
		(1.00, 1.00, 1.00, 1.00, 1.00)	16.76	16.33	17.14	16.98	16.16	16.38	16.65
	Netherlands	(0.10, 0.25, 0.50, 0.75, 1.00)	31.07	28.31	25.26	21.88	20.61	19.86	19.60
		(0.60, 0.70, 0.80, 0.90, 1.00)	30.75	28.64	27.01	24.20	22.17	20.76	20.54
		(1.00, 1.00, 1.00, 1.00, 1.00)	30.77	29.42	27.45	25.06	23.85	22.92	22.26

Table 12.2 (continued)

	Team	p_i's	a_0						
	Germany	(0.10, 0.25, 0.50, 0.75, 1.00)	47.44	49.10	52.37	54.42	55.38	55.98	55.81
		(0.60, 0.70, 0.80, 0.90, 1.00)	44.33	45.83	47.21	49.49	51.35	52.66	52.97
		(1.00, 1.00, 1.00, 1.00, 1.00)	43.76	44.77	46.28	48.29	50.24	51.30	51.40
	Uruguay	(0.10, 0.25, 0.50, 0.75, 1.00)	6.92	7.54	7.28	8.11	8.45	8.59	9.18
		(0.60, 0.70, 0.80, 0.90, 1.00)	8.60	8.95	9.17	9.05	10.10	10.24	9.96
		(1.00, 1.00, 1.00, 1.00, 1.00)	8.71	9.48	9.13	9.67	9.75	9.40	9.69
Before final	Spain	(0.10, 0.25, 0.50, 0.75, 0.90, 1.00)	37.50	38.9	40.38	43.16	45.20	46.45	46.57
		(0.50, 0.60, 0.70, 0.80, 0.90, 1.00)	39.66	40.64	41.32	44.04	46.48	46.71	47.41
		(1.00, 1.00, 1.00, 1.00, 1.00, 1.00)	40.50	41.71	43.21	44.66	45.58	46.28	46.86
	Netherlands	(0.10, 0.25, 0.50, 0.75, 0.90, 1.00)	62.50	61.10	59.62	56.84	54.80	53.55	53.43
		(0.50, 0.60, 0.70, 0.80, 0.90, 1.00)	60.34	59.36	58.68	55.96	53.52	53.29	52.59
		(1.00, 1.00, 1.00, 1.00, 1.00, 1.00)	59.50	58.29	56.79	55.34	54.42	53.72	53.14

The method may be used to calculate the win, draw, and loss probabilities at each single match, as well as to simulate the whole competition in order to estimate, for instance, probabilities of classification at group stage, of reaching the knockout stage or the final match, and of winning the tournament.

Moreover, the method presents a high performance within a simulation structure since known predictive distributions are obtained. This enables a rapid generation of predictive distribution values and consequently the probabilities of interest are obtained quickly.

Overall, the Bayesian simulation methodology with different weight values for the played matches and different weight values for the expert's opinions provides a better idea on the impact of the latest matches and the different weights assigned to the experts' opinion on the estimated probabilities of interest, evidencing the advantage of incorporating time-effect weights for the match results. In our analysis the weights of the experts' opinion are fixed and known. As further work it may be considered one distinct value of a_0 for each expert and round allowing changes of the values over the rounds.

One interpretation that can be made is that, for a particular expert, if a_0 increases over the rounds, the confidence of the information given by this expert increases as well.

Alternatively, if a_0 decreases, that means the information provided by this expert in previous rounds were not reliable. Furthermore, we can assume that a_0 is a random variable and use a full hierarchical structure specifying a parametric distribution for the parameter a_0, like a beta distribution, as suggested in [2].

Acknowledgments The research is partially supported by the Brazilian Government Agencies: CNPq, CAPES, and FAPESP.

References

1. Brillinger, D.R.: Modelling game outcomes of the Brazilian 2006 series A championship as ordinal-valued. Braz. J. Probab. Stat. **22**, 89–104 (2008)
2. Chen, M.H., Ibrahim, J.G.: Power prior distributions for regression models. Stat. Sci. **15**, 46–60 (2000)
3. Clemen, R.T., Winkler, R.L.: Combining probability distributions from experts in risk analysis. Risk Anal. **19**, 187–203 (1999)
4. De Finetti, B.: Probability, Induction and Statistics. Wiley, London (1972)
5. Dyte, D., Clarke, S.R.: A ratings based Poisson model for World Cup Soccer simulation. J. Opl. Res. Soc. **51**, 993–998 (2000)
6. Genest, C., Zidek, J.V.: Combining probability distributions: a critique and an annotated bibliography. Stat. Sci. **1**, 114–148 (1986)
7. Karlis, D., Ntzoufras, I.: Bayesian modelling of football outcomes: using the Skellam's distribution for the goal difference. IMA J. Manag. Math. **20**, 133–145 (2009)
8. Keller, J.B.: A characterization of the Poisson distribution and the probability of winning a game. Am. Stat. **48**, 294–298 (1994)
9. Lee, A.: Modeling scores in the premier league: is Manchester United really the best?. Chance **10**, 15–19 (1997)

10. O'Hagan, A., Buck, C.E., Daneshkhah, A., Eiser, J.R., Garthwaite, P.H., Jenkinson, D.J., Oakley, J.E., Rakow, T.: Uncertain Judgements: Eliciting Experts' Probabilities. Wiley, London (2006)
11. Pill, J.: The Delphi method: substance, context, a critique and an annotated bibliography. Socio-Econ. Plan. Sci. **5**, 57–71 (1971)
12. Suzuki, A.K., Salasar, L.E.B., Louzada-Neto, F., Leite, J.G.: A Bayesian approach for predicting match outcomes: the 2006 (Association) Football World Cup. J. Oper. Res. Soc. **61**, 1530–1539 (2010)
13. Volf, P.: A random point process model for the score in sport matches. IMA J. Manag. Math. **20**, 121–131 (2009)

Chapter 13
Homogeneity Tests for 2 × 2 Contingency Tables

Natalia Oliveira, Marcio Diniz and Adriano Polpo

Abstract Using the likelihood ratio statistic, we develop a significance index, called *P*-value, to test the hypothesis of homogeneity in 2 × 2 contingency tables. The *P*-value does not depend on asymptotic distributions, and is based on the elimination of the nuisance parameter. Therefore, we obtain the exact distribution of the likelihood ratio statistic in a way that is, moreover, compatible with the likelihood principle. For a better understanding of significance indices to test homogeneity, we perform a study comparing the *P*-value with some indices (likelihood ratio test (LRT), chi-square test) and with the full Bayesian significance test (FBST). This comparative study shows an interesting relation between all the analyzed indices, Bayesian and frequentist.

13.1 Introduction

Hypothesis testing is widely conducted in several fields of applied sciences. Among the known procedures to test hypotheses, the *p*-value is one of the most used by all kind of researchers. However, one can question if it is truly a good and reasonable index. The literature is rich [4, 5] in providing examples where the use of *p*-values leads to harming consequences or where it is based on delicate assumptions. More specifically, considering the homogeneity hypothesis in contingency tables, the evaluation of *p*-values is based on asymptotic considerations. Is it a reliable approximation? Is it possible to construct an exact *p*-value to test such hypothesis? Trying to answer these questions, we develop an exact index, called *P*-value, for the likelihood ratio test (LRT) and also evaluate the quality of indices based on asymptotic approximations to test the homogeneity hypothesis in 2 × 2 contingency tables.

N. Oliveira (✉) · M. Diniz · A. Polpo
Federal University of Sao Carlos, Rod. Washington Luiz, km 235, Sao Carlos 13565-905, SP, Brazil
e-mail: nat.nlo@gmail.com

M. Diniz
e-mail: marcio.alves.diniz@gmail.com

A. Polpo
e-mail: polpo@ufscar.br

A. Polpo et al. (eds.), *Interdisciplinary Bayesian Statistics,*
Springer Proceedings in Mathematics & Statistics 118, DOI 10.1007/978-3-319-12454-4_13

Table 13.1 Contingency table 2×2

	C_1	C_2	Total
X	x	$n_x - x$	n_x
Y	y	$n_y - y$	n_y

After establishing the definition of P-value, our first concern was to assess if it is a good alternative as an exact p-value for the LRT. This is done analyzing the relationship between the P-value and the asymptotic p-value as the sample size increases. Another way to evaluate that is to compare the P-value with the e-value, the significance index that results from the the full Bayesian significance test (FBST) suggested by Pereira and Stern [6] and revised by Pereira et al. [8].

Since Diniz et al. [3] showed an asymptotic relationship between the p-value (for the LRT) and the e-value, it is expected that, as the sample size increases, the relationship between P-value and e-value will become evident.

After evaluating, according to some criterion, the quality of P-value, we also analyze asymptotic approximations, that is, we try to answer the question: Which are the smaller sample sizes that make the exact indices closely related to the asymptotic indices? The same analysis done to study the P-value is then conducted to evaluate asymptotic approximations.

This chapter is organized as follows. The next section describes the homogeneity test for 2×2 contingency tables. Then, we define the P-value and comment on its computation for this specific test. Closing the section we review other procedures to test hypotheses used by the frequentist (chi-square and asymptotic LRT) and Bayesian (FBST) schools. Section 12.3 presents some comparisons among the indices derived from these procedures considering some tables of fixed (sample) size. We close with some comments and perspectives on future research.

13.2 Homogeneity Test for 2×2 Contingency Tables and Significance Indices

In this section, we describe the homogeneity test for 2×2 contingency tables. Consider an experiment whose outcome is classified according to two dimensions: in one dimension it can assume categories X or Y and in the other it can assume categories C_1 or C_2. In this case, the experiment can result in any one of the four pairs: (X, C_1), (X, C_2), (Y, C_1), or (Y, C_2). After performing the experiment a given number of times, one can display Table 13.1, where $n_x(n_y)$ is the number of times outcome $X(Y)$ was observed and, among those $x(y)$ were of category C_1.

There is more than one way to statistically model this table. In this work, we assume that n_x experiments of category X and n_y of category Y are performed, that is, the margins of the table are known or given in advance. In this context, the simplest statistical model assumes that X follows a Binomial(n_x, θ_x) distribution and Y a Binomial(n_y, θ_y) distribution.

The homogeneity hypothesis for 2×2 tables considers that, in one dimension, both categories are equal or *homogeneous*, that is,

$$\mathrm{H} : \theta_x = \theta_y = \theta. \tag{13.1}$$

Following the notation used in Table 13.1 and under hypothesis (13.1), the likelihood function is specified by

$$L(\theta \mid x, y, n_x, n_y) = \frac{n_x! n_y!}{x! y! (n_x - x)! (n_y - y)!} \theta^{x+y} (1-\theta)^{n_x+n_y-x-y}, \tag{13.2}$$

in which $\theta \in [0, 1]$.

Given that the LRT statistic will be used in the sequel, we derive it now. Being $\mathcal{X} \times \mathcal{Y}$ the sample space, Θ the unrestricted parameter space and Θ_H the parameter space under (13.1), the LRT statistic for a given sample point $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is

$$\begin{aligned} \lambda(x, y) &= \frac{\sup_{\theta \in \Theta_H} L(\theta \mid x, y)}{\sup_{\theta \in \Theta} L(\theta \mid x, y)} = \frac{\sup\limits_{\theta \in \Theta_H} L(\theta_x, \theta_y \mid x, y, n_x, n_y)}{\sup\limits_{\theta \in \Theta} L(\theta_x, \theta_y \mid x, y, n_x, n_y)} \\ &= \frac{\left(\frac{x+y}{n_x+n_y}\right)^{x+y} \left(\frac{n_x+n_y-x-y}{n_x+n_y}\right)^{n_x+n_y-x-y}}{\left(\frac{x}{n_x}\right)^{x} \left(\frac{n_x-x}{n_x}\right)^{n_x-x} \left(\frac{y}{n_y}\right)^{y} \left(\frac{n_y-y}{n_y}\right)^{n_y-y}}. \end{aligned} \tag{13.3}$$

Now we proceed to the description of the indices used in this chapter, namely: the P-value, the p-values derived from chi-square test and the asymptotic LRT and the e-value derived from the FBST.

13.2.1 P-value

The ideas behind the P-value, as presented here, were first discussed by Pereria and Wechsler [7]. We implement their ideas and compute the P-value based on the predictive distribution of (X, Y)— the marginal distribution of (X, Y) before the observations were collected. Following this approach, it is possible to obtain the exact distribution of the LRT statistic and to avoid the violation of the likelihood principle.

To compute this index, it is necessary to derive the sampling distribution of the test statistic under $\boldsymbol{H}$. Therefore, we want to find the distribution—that does not depend on θ—of the contingency table under $\boldsymbol{H}$. For this, we consider that θ is a nuisance parameter and then integrate (13.2), the likelihood function under $\boldsymbol{H}$, over θ, to eliminate it, that is,

$$h(x, y) = \int_0^1 L(\theta \mid x, y, n_x, n_y) \mathrm{d}\theta = \frac{\binom{n_x}{x}\binom{n_y}{y}}{\binom{n_x+n_y}{x+y}} \frac{1}{(n_x + n_y + 1)}. \tag{13.4}$$

To have a probability distribution, we must find the normalization constant. The distribution is thus defined by

$$\Pr(X = x, Y = y \mid \boldsymbol{H}) = \frac{h(x, y)}{\sum_{i=0}^{n_x} \sum_{j=0}^{n_y} h(i, j)}. \tag{13.5}$$

To calculate $\sum_{i=1}^{n_x} \sum_{j=1}^{n_y} h(i, j)$, an algorithm is used to obtain $h(i, j)$ for all possible tables with margins n_x, n_y and then we sum over all these values.

Note that $\Pr(X = x, Y = y \mid \boldsymbol{H})$ does not depend on θ. Therefore, the P-value is obtained directly from this distribution, by

$$P\text{-value} = \Pr(\lambda(X, Y) \leq \lambda(x, y) \mid \boldsymbol{H}) = \sum_{(i,j):\ \lambda(X,Y) \leq \lambda(x,y)} \sum \Pr(X = i, Y = j \mid \boldsymbol{H}),$$

in which $\lambda(x, y)$ is the observed test statistic, given by (13.3).

The next sections briefly describe some other tests and their respective significance indices.

13.2.2 Chi-Square Test

This is the most used test for the homogeneity hypothesis in 2×2 contingency tables.

Let $n_{i\cdot}$ be total of the ith line, $n_{\cdot j}$ the total for the jth column, n the total of the table, ℓ the number of lines, and c the number of columns. The test statistic is given by

$$\chi^2_{obs} = \sum_{i=1}^{\ell} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

where O_{ij} is the observed count of the i, jth cell and $E_{ij} = \frac{n_{i\cdot} \times n_{\cdot j}}{n}$ is the expected value of the same cell, under the null hypothesis. Assuming some regularity conditions [9, Chap. 17], the statistic χ^2_{obs} has asymptotic distribution given by $\chi^2_{(c-1)\times(\ell-1)}$. Since we are considering the case $c = \ell = 2$, the asymptotic p-value for this test is given by

$$p\text{-value} = \Pr(\chi^2_1 \leq \chi^2_{obs}).$$

13.2.3 Asymptotic LRT

Another important index to take into account is the LRT asymptotic p-value. This index for an observed point $\boldsymbol{x} \in \mathcal{X}$ is the probability, under $\boldsymbol{H}$, of the event

$$S_x = \{z \in \mathcal{X} : \lambda(z) \leq \lambda(\boldsymbol{x})\}. \tag{13.6}$$

Considering the usual regularity conditions [2, 9, 10], $-2\ln[\lambda(\boldsymbol{x})]$ follows, asymptotically, a chi-square distribution with $m-h$ degrees of freedom, where m is the dimension of the parameter space and h is the dimension of the null hypothesis [1].

13.2.4 FBST

Let $\pi(\theta \mid \boldsymbol{x})$ be the posterior density or distribution function of θ given the observed sample, $\boldsymbol{x}$, and $T(\boldsymbol{x}) = \{\theta \in \Theta : \pi(\theta \mid \boldsymbol{x}) \geq \sup_{\theta\in\Theta_H} \pi(\theta \mid \boldsymbol{x})\}$ called the tangent set to the hypothesis. The evidence measure supporting the hypothesis $\theta \in \Theta_H$ is defined as $Ev(\Theta_H, \boldsymbol{x}) = 1 - \Pr(\theta \in T(\boldsymbol{x}) \mid \boldsymbol{x})$, and called e-value.

Considering the notation established in Sect. 13.2, we first derive the posterior distribution of (θ_x, θ_y). As prior for such vector, we consider a product of the uniforms, the joint posterior being given by

$$\theta_x, \theta_y \mid x, y, n_x, n_y \sim \text{Beta}(x+1, n_x - x + 1)\text{Beta}(y+1, n_y - y + 1). \quad (13.7)$$

To compute $\sup_{\theta_x,\theta_y\in H} \pi(\theta_x, \theta_y \mid x, y, n_x, n_y)$, we should maximize

$$\pi(\theta \mid x, y, n_x, n_y) = \binom{n_x}{x}\binom{n_y}{y}(n_x+1)(n_y+1)\theta^{x+y}(1-\theta)^{n_x+n_y-x-y},$$

with respect to θ. Therefore,

$$\sup_{\theta\in(0,1)} \pi(\theta \mid x, y, n_x, n_y) = \sup_{\theta_x,\theta_y\in H} \pi(\theta_x, \theta_y \mid x, y, n_x, n_y)$$
$$= \frac{(n_x+1)!(n_y+1)!}{x!y!(n_x-x)!(n_y-y)!}\left(\frac{x+y}{n_x+n_y}\right)^{x+y}\left(\frac{n_x+n_y-x-y}{n_x+n_y}\right)^{n_x+n_y-x-y},$$

and T, the tangent set to the hypothesis is

$$T(x, y, n_x, n_y) = \{\theta \in (0,1) : \pi(\theta \mid x, y, n_x, n_y) \geq \sup_{\theta\in(0,1)} \pi(\theta \mid x, y, n_x, n_y)\},$$

finally leading to e-value,

$$e\text{-value} = 1 - \Pr[\theta \in T(x, y, n_x, n_y)].$$

The computation of the asymptotic e-value is rather simple and almost identical to the asymptotic LRT p-value: it just requires the evaluation of $\Pr[-2\ln(\lambda(X,Y)) \leq -2\ln(\lambda(x,y))]$. But unlike the asymptotic LRT p-value, under regularity assumptions, the asymptotic distribution of $-2\ln(\lambda(X,Y))$ is shown to be χ^2 with m degrees of freedom, where m is the dimension of the parameter space. See [3] and [8] for more on the asymptotic e-value.

13.3 Comparing the Indices

In this section, we study the behavior of the proposed P-value when compared with other indices. We remark the fact that this is not a simulation study since for each sample size we evaluate all indices for all possible configurations of 2×2 contingency tables.

First, considering that all indices presented in the previous section are based on the same statistic, $\lambda(x, y)$, we plot the indices against the value of the respective statistic. The studied indices are the P-value, the LRT asymptotic p-value, the chi-square p-value, and e-values (exact and asymptotic). The sample sizes considered are $(n_x, n_y) = \{(5, 5), (10, 10), (17, 29), (30, 30), (100, 100), (300, 300), (1000, 1000)\}$.

Figure 13.1 shows the comparisons. It is clear that the exact and the asymptotic e-values are quite similar, even for small sample sizes, being close to the 45º line (grey circles and triangles). The P-value, and the LRT and chi-square asymptotic p-values (diamonds, squares, and dark circles, respectively) are also very similar to each other.

The difference found between e-values—exact and asymptotic—and all the three p-values is a consequence of the construction of the indices. While e-values consider m degrees of freedom, p-values consider $m - h$, where h is the dimension of Θ_H. This situation is more clear while noticing the asymptotic relationship between e-value and the LRT p-value highlighted by Diniz et al. [3].

Bearing this relationship in mind, it is natural to question if there is some relation between the P-value and the e-value. To verify that, Fig. 13.2, displays graphs with P-values in the horizontal axis and exact e-values in the vertical axis. The grey line represents the asymptotic relationship between e-value and the LRT p-value. From the figures, it is clear that, as the sample size increases, both indices are attracted towards the asymptotic line. In this case, we think that the asymptotic relation between the e-value and the LRT p-value given in [3] is also being verified for the P-value and the e-value.

13.4 Discussion and Conclusion

In this chapter, we implement the ideas of Pereira and Wechsler [7] to compute the P-value, a significance index to test the homogeneity hypothesis in 2×2 contingency tables, based on the LRT statistic. Our index does not violate the likelihood principle since the distribution of the test statistic does not depend on the parameter θ. Moreover, the comparison study provide us the behavior of the proposed index, showing that it is very similar to the asymptotic p-values, and to the e-value.

The main disadvantage of the P-value is that we need to evaluate the probability of every contingency table in the sample space, given the margins. For 2×2 tables this is not time consuming, at least for the sample sizes we considered for in the comparative study. For larger tables (3×3, 4×3, ...) it may be even impossible to compute it, since the amount of tables to be evaluated increases exponentially.

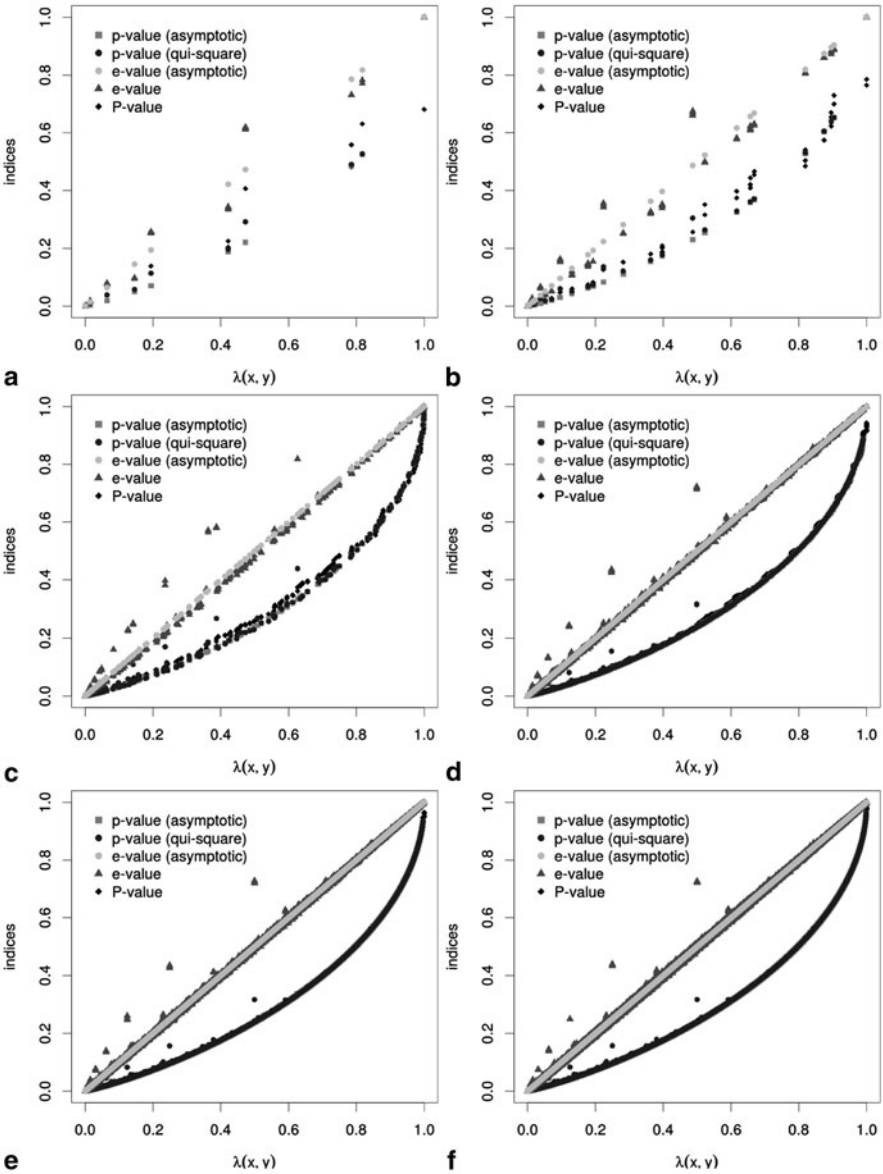

Fig. 13.1 Comparison between significance indices for contingency tables (*horizontal axis*: LRT statistic; *vertical axis*: significance indices; *red*: *P*-value; *green*: *p*-value asymptotic; *magenta*: χ^2 *p*-value; *blue*: *e*-value; *cyan*: asymptotic *e*-value)

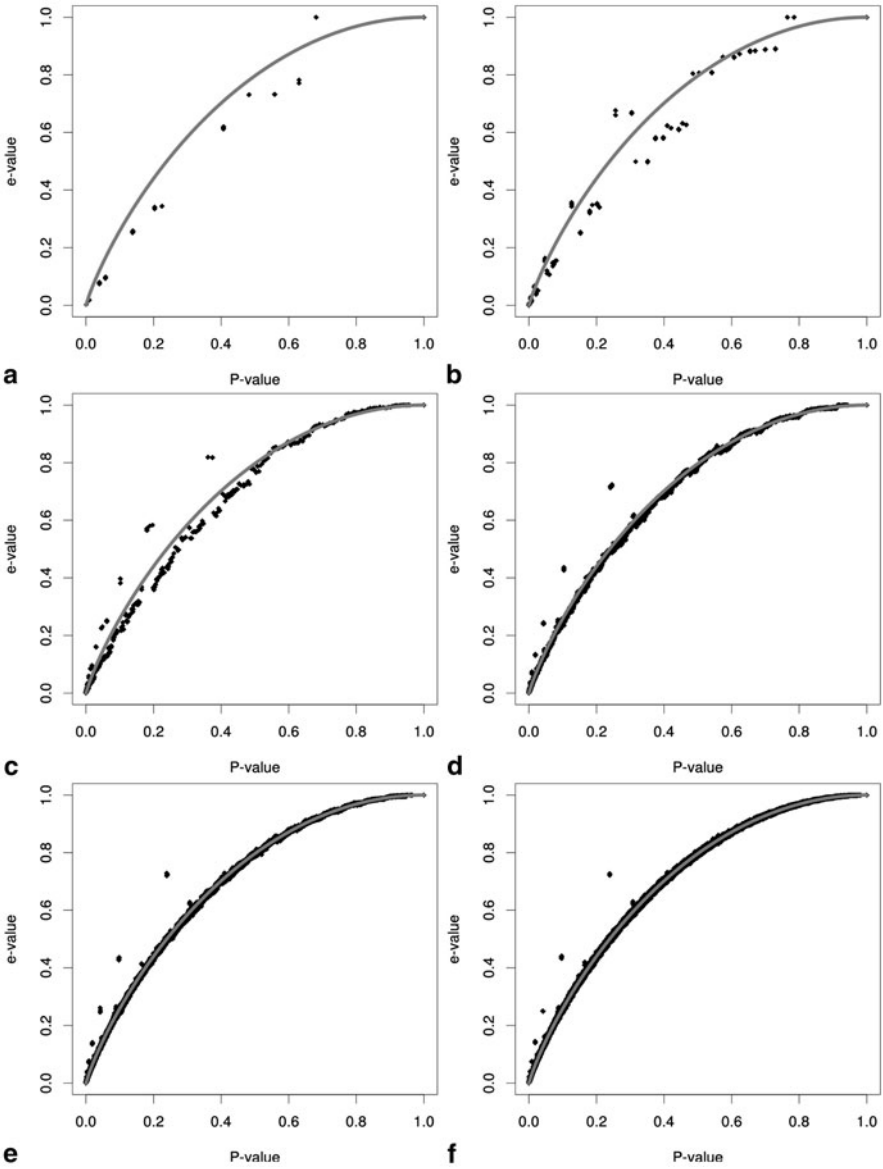

Fig. 13.2 Comparison between *P*-value and *e*-value for 2 × 2 contingency tables (the *red line* is the asymptotic relationship between LRT's *p*-value and *e*-value, *dots* on *top* of this *line* represent equality between indices)

However, the study shows that the P-value is very similar to the asymptotic p-values, even for small sample sizes. Then, since the asymptotic approximation turns out to be reasonable in unexpected conditions, we can consider that asymptotic indices (p-values and e-values) produces reliable results even with small sample sizes.

As a perspective for future research, we may address the homogeneity test for contingency tables. We can also compare the results with other options to evaluate the exact p-value for contingency tables.

References

1. Casella, G., Berger, R.: Statistical Inference, 2nd edn. Duxbury Press, Pacific Grove (2001)
2. Chernoff, H.: On the distribution of the likelihood ratio. Ann. Math. Stat. **25**(3), 573–578 (1954)
3. Diniz, M.A., Pereira, C.A.B., Polpo, A., Stern, J., Wechesler, S.: Relationship between Bayesian and frequentist significance indices. Int. J. Uncertain. Quantif. **2**(2), 161–172 (2012). doi:10.1615/Int.J.UncertaintyQuantification.2012003647
4. Lin, M., Lucas, H.C., Shmueli, G.: Research commentary-too big to fail: large samples and the p-value problem. Inf. Syst. Res. **24**(4), 906–917 (2013). doi:10.1287/isre.2013.0480. http://pubsonline.informs.org/doi/abs/10.1287/isre.2013.0480
5. Moran, J., Solomon, P.: A farewell to p-values? Crit. Care Resusc. **6**, 130–137 (2004)
6. Pereira, C., Stern, J.: Evidence and credibility: a full Bayesian test of precise hypothesis. Entropy **1**, 104–115 (1999)
7. Pereira, C.A.B., Wechsler, S.: On the concept of p-value. Braz. J. Probab. Stat. **7**, 159–177 (1993)
8. Pereira, C.A., Stern, J., Wechsler, S.: Can a significance test be genuinely Bayesian? Bayesian Anal. **3**(1), 19–100 (2008)
9. van der Vaart, A.: Asymptotic Statistics. Cambridge University Press, Cambridge (1998). http://books.google.com.br/books?id=UEuQEM5RjWgC
10. Wilks, S.S.: The large-sample distribution of the likelihood ratio for testing composite hypotheses. Ann. Math. Stat. **9**, 60–62 (1938)

Chapter 14
Combining Optimization and Randomization Approaches for the Design of Clinical Trials

Victor Fossaluza, Marcelo de Souza Lauretto, Carlos Alberto de Bragança Pereira and Julio Michael Stern

Abstract Intentional sampling methods are non-randomized procedures that select a group of individuals for a sample with the purpose of meeting specific prescribed criteria. In this paper, we extend previous works related to intentional sampling, and address the problem of sequential allocation for clinical trials with few patients. Roughly speaking, patients are enrolled sequentially, according to the order in which they start the treatment at the clinic or hospital. The allocation problem consists in assigning each new patient to one, and only one, of the alternative treatment arms. The main requisite is that the profiles in the alternative arms remain similar with respect to some relevant patients' attributes (age, gender, disease, symptom severity and others). We perform numerical experiments based on a real case study and discuss how to conveniently set up perturbation parameters, in order to yield a suitable balance between *optimality*—the similarity among the relative frequencies of patients in the several categories for both arms, and *decoupling*—the absence of a tendency to allocate each pair of patients consistently to the same arm.

> We describe a possible allocation that the experimenter judges to be free of covariate interference as *haphazard*. Randomization may be a convenient way of producing a haphazard design. We argue that it is the haphazard nature, and not the randomization, that is important. It seems therefore that a reasonable approximation to an optimal design would be to select a haphazard design. ... a detailed Bayesian consideration of possible covariates would almost certainly not be robust in that the analysis might be sensitive to small changes in judgments about covariates.
>
> Lindley [16, pp. 438–439] - The Role of Randomization in Inference.

V. Fossaluza (✉) · C. A. de Bragança Pereira · J. M. Stern
Institute of Mathematics and Statistics,
University of São Paulo, São Paulo, Brazil
e-mail: victorf@ime.usp.br

M. de Souza Lauretto
EACH—USP, São Paulo, Brazil
e-mail: marcelolauretto@usp.br

C. A. de Bragança Pereira
e-mail: cpereira@ime.usp.br

J. M. Stern
e-mail: jstern@ims.usp.br

A. Polpo et al. (eds.), *Interdisciplinary Bayesian Statistics,*
Springer Proceedings in Mathematics & Statistics 118, DOI 10.1007/978-3-319-12454-4_14

14.1 Introduction

Lindley [17, pp. 47–48] illustrates the celebrated Simpson's paradox with a medical trial example. In such example, the association between two variables, Treatment and Recovery from a given illness, is reversed if the data is aggregated or disaggregated over a confounding variable, sex, see also [33]. Randomized and double-blind (masked) clinical trials are designed to shield the experiment from undesired bias effects caused by undue interference or deliberate manipulation of confounding variables. The introduction of randomized tests, first proposed in Peirce and Jastrow [23] and popularized by Fisher [11], has established a new paradigm for statistically valid empirical research. However, the standard sampling methods by randomized design are not always appropriate; for example, they have limited application when cost, ethical or inherent rarity constraints only allow the use of very small samples.

Intentional sampling methods are non-randomized procedures that select or allocate groups of individuals with the purpose of meeting specific prescribed criteria. Such methods can overcome some of the aforementioned limitations of standard randomized designs for statistical experiments. However, intentional or purposive sampling methods pose several interesting questions concerning statistical inference, as extensively discussed in Basu and Ghosh [4], see also Schreuder et al. [27, Sect. 6.2], Brewer [7] and Särndal [8] and following discussions in Madow et al. [18].

This paper focus on sequential allocation methods, following previous research in the field of intentional sampling presented in [12, 15]. Particularly, we discuss an allocation scheme that combines aspects of intentional and randomized sampling methods.

The paper is organized as follows. Section 14.2 provides a brief discussion of sampling design under the perspective of linear regression superpopulation model. Section 14.3 elucidates how to use linear regression models to handle compositional data, which is of direct interest for our case study. Section 14.4 illustrates our approach of combining purposive sampling with random perturbation techniques for *providing samples which are approximately balanced*, as stated by Royall and Pfeffermann [26, p. 20], in an application case concerning a clinical trial allocation. Section 14.5 presents and discusses our numerical experiments and results.

14.2 Linear Regression Superpopulation Model

We will introduce the basic ideas for our approach in the context of the linear regression superpopulation model, as presented by Royall and Pfeffermann [26], Pereira and Rodrigues [24] and Tam [30].

$$\mathrm{E}\left(\begin{bmatrix} y_s \\ y_r \end{bmatrix}\right) = \begin{bmatrix} X_s \\ X_r \end{bmatrix} \beta\,, \quad \mathrm{Cov}\left(\begin{bmatrix} y_s \\ y_r \end{bmatrix}\right) = \begin{bmatrix} V_{s,s} & V_{s,r} \\ V'_{s,r} & V_{r,r} \end{bmatrix}.$$

Row i of the $n \times m$ matrix X contains the explanatory variables for individual i, and is known for the entire population. The response variable, y_i, is observed for the individuals in a given sample, S, indexed by i in $s = [1, 2, \ldots, m]$, and unobserved for the remaining individuals, indexed by i in $r = [m+1, m+2, \ldots, n]$. (Whatever the sample, $s = [i_1, i_2, \ldots, i_m]$, we can always reorder the indices so to place them first.) We partition all vectors and matrices of the model accordingly and assume that $V_{s,s} > 0$.

We seek an optimal linear predictor, $p'y_s$, for a quantity of interest, $\kappa = q'y$, and define the auxiliary matrices: $t' = [t'_s, t'_r] = [p' - q'_s, -q'_r]$ and $M = \left(X'_s V_{s,s}^{-1} X_s\right)^{-1}$.

Since probability expectation is a linear operator, that is, for any random (vector) variable, z, $\mathrm{E}(Az + b) = A\mathrm{E}(z) + b$, one can compute the expected value of the prediction error,

$$\mathrm{E}\left(t'y\right) = \left(t'X\right)\beta = \left(t'_s X_s - q'_r X_r\right)\beta.$$

Hence, for a general parameter β, the predictor p is unbiased if and only if it obeys the *balance* constraint,

$$t'X = t'_s X_s - q'_r X_r = p'X_s - q'X = 0.$$

Solving the normal linear system for the minimum variance unbiased estimator problem at hand yields the solution

$$t^*_s = (p^* - q_s) = V_{s,s}^{-1}\left(V_{s,r} + X_s U\left(X'_r - X'_s V_{s,s}^{-1} V_{s,r}\right)\right) q_r.$$

Finally, we can write the optimal (minimum-variance unbiased) prediction for the quantity of interest, $\kappa = q'y$, as

$$\hat{\kappa} = q'_s y_s + q'_r\left(X_r\hat{\beta} + V_{r,s}V_{s,s}^{-1}\left(y_s - X_s\hat{\beta}\right)\right), \text{ where } \hat{\beta} = MX_s V_{s,s}^{-1} y_s, \text{ and}$$

$$\mathrm{Var}\left(\hat{\kappa}\right) = q'_r\left(V_{r,r} - V_{r,s}V_{s,s}^{-1}V'_{r,s}\right)q_r + q'_r\left(X_r - V_{r,s}V_{s,s}^{-1}X_s\right)M\left(X_r - V_{r,s}V_{s,s}^{-1}X_s\right)' q_r.$$

The balance and optimality conditions obtained above can be used for choosing a good predictor p, but they can also be used to select a "good" sample $s = [i_1, i_2, \ldots, i_m]$. Many survey sampling studies are interested in population totals, where $q = \mathbf{1}$, that is, $\kappa = \mathbf{1}'y$ where $\mathbf{1}$ is the column vector of ones of appropriate dimension in the context. In this case, the balance equation takes the form $p'X_s = \mathbf{1}'X$.

Robustness is also a desirable characteristic of a survey design. Imagine our model is misspecified, say by omission of important covariates, Z, in the expanded covariate matrix $\widetilde{X} = [X, Z]$. Without loss of generality, assume that we use "orthogonal" covariates, for witch $X'Z = 0$ and $Z'Z = I$. We would like to still be able to make a useful prediction. In general, this desire is just wishful thinking if we know nothing about the ignored covariates in Z. Now, assume that we fix p as the *expansion* predictor, $p = (N/n)\mathbf{1}$, and choose a *representative* sample, X_s, for which the sample totals are (approximately) balanced, that is, $p'X_s \approx \mathbf{1}'X$. If we are lucky enough,

the balance equation will also (approximately) hold for the omitted covariates, that is, $(N/n)\mathbf{1}'Z_s \approx \mathbf{1}'Z$. For further developments of this idea, see [24, 26, 30].

But how do we manage to get lucky? According to Lindley [16, pp. 438–439], that seems to be the role of randomization in experimental design. For complementary and mutually supportive views of the role of randomization in statistical design of experiments, see [6, 28, 29]. There is a vast literature in design-based random sampling that aims to achieve this goal. In this paper, we explore the use of purposive sampling. For a more extensive discussion of this approach and a complete application case, see [15].

14.3 Compositional Models and Simplex Geometry

We begin this section reviewing some basic notions of compositional models and Simplex geometry, as presented in [1, 2]. The open (m-1)-Simplex is the set $S^{m-1} = \left\{x \in R^m \mid x > 0 \wedge \mathbf{1}'x = 1\right\}$, where $\mathbf{1}$ in the vector of ones of appropriate dimension. The *closure-to-unity* transformation, from R^m_+ to S^{m-1}, the *additive logratio transformation*, from S^{m-1} to the unrestricted R^{m-1} space, and the *centered logratio transformation*, from S^{m-1} to a hyperplane through the origin of R^m, are defined as

$$\text{clu}(x) = (1/\mathbf{1}'x)x, \quad \text{alr}(x) = \log((1/x_m)[x_1, \ldots x_{m-1}]),$$
$$\text{and clr}(x) = \log((1/g(x))[x_1 \ldots x_m]), g(x) = (x_1 x_2 \ldots x_m)^{1/m}.$$

It is easy to check the normalization conditions stating that $\mathbf{1}'\text{clu}(x) = 1$ and $\mathbf{1}'\text{clr}(x) = 0$, as well the following expressions of the inverse transformations

$$\text{alr}^{-1}(z) = \text{clu}(\exp([z_1, \ldots z_{m-1}, 0])), \text{clr}^{-1}(z) = \text{clu}(\exp([z_1, \ldots z_m])).$$

We can introduce the *power* (scalar multiplication) operator, $*$, and the *perturbation* (vector summation) operation, $\oplus$, providing a vector space structure for the Simplex, namely, $\alpha * x = \text{clu}([x_1^\alpha, \ldots x_m^\alpha])$ and $x \oplus y = \text{clu}([x_1 y_1, \ldots x_m y_m])$. The perturbation operation can be interpreted as the effect of proportional decay rates in y over the fractional composition in x. The power operation can be interpreted as the α-times repeated effect of proportional decay rates. The perturbation operation defines an Abelian (commutative) group, where the identity element is $e = (1/m)\mathbf{1}$, and the inverse of a perturbation is given by $x^{-1} = \text{clu}([1/x_1, \ldots 1/x_m])$. Hence, we define the difference $x \ominus y = \text{clu}([x_1/y_1, \ldots x_m/y_m])$.

Next, we equip the Simplex with a metric structure. More specifically, we search for a distance function, $D_S(x, y)$, that exhibits the invariance properties that are most adequate for the purpose of compositional analysis. The most important of these invariance properties are:

- Perturbation invariance: For any perturbation, z, $D_S(x \oplus z, y \oplus z) = D_S(x, y)$.
- Permutation invariance: For any permutation matrix, P, $D_S(Px, Py) = D_S(x, y)$.
- Power scaling: For any $\alpha > 0$, $(1/\alpha)D_S(\alpha * x, \alpha * y) = D_S(x, y)$.

The following distance function exhibits all of these desirable invariance properties, as well as the standard properties required from a distance function, like positivity, symmetry and the triangular inequality.

$$D^2_{S(x,y)} = [\mathrm{clr}(x) - \mathrm{clr}(y)]' I [\mathrm{clr}(x) - \mathrm{clr}(y)] =$$
$$= [\mathrm{alr}(x) - \mathrm{alr}(y)]' H^{-1} [\mathrm{alr}(x) - \mathrm{alr}(y)], H_{i,j} = 2\delta_{i,j} + 1(1 - \delta_{i,j}).$$

We can further extend the mathematical structure over the Simplex to a vector (finite Hilbert) space, defining the inner product

$$\langle x, y \rangle_S = \mathrm{clr}(x)' I \mathrm{clr}(y) = \mathrm{alr}(x)' H^{-1} \mathrm{alr}(y).$$

Defining the norm, $\|x\|_S^2 = \langle x, x \rangle_S$, it is easy to check that $D_S(x, y) = \|x \ominus y\|_S$ and that $\|x\|_S = D_S(x, e)$. Finally, we can compose the additive logratio transformation with an orthogonalization operation that translates Aitchison's inner product for the Simplex to the standard inner product for the unrestricted Euclidean space. For example, we can use the Cholesky factorization $L'L = H$ to define the *isometric logratio transformation* $\mathrm{ilr}(x) = L^{-t}\mathrm{alr}(x)$, where L^{-t} denotes the transpose of L^{-1}. In this case, since $H^{-1} = L^{-1}L^{-t}$, we have

$$\langle x, y \rangle_S = \mathrm{ilr}(x)' L (L^{-1}L^{-t}) L' \mathrm{ilr}(y) = \mathrm{ilr}(x)' I \mathrm{ilr}(y) = \langle \mathrm{ilr}(x), \mathrm{ilr}(y) \rangle_E .$$

All these compatible vector space structures make it easy to develop linear regression models for compositional data, see [10, 21]. In the most simple terms, in the Euclidean space we can use the standard linear properties of the expectation operator, and consequent transformation properties for the covariance operator, as reviewed in the previous section. These properties suffice to prove Gauss–Markov theorem, allowing the computation of optimal unbiased estimators via least-squares linear algebra, see [32, Sect. 14.4]. Hence, we can map compositional data, naturally presented in S^{m-1}, into R^m or R^{m-1}, analyse the data with linear regression models and, if so desired, map the models back to the Simplex.

The centered logratio transformation from S^{m-1} to R^m requires regression models under linear equality constraints. Meanwhile, the additive logratio transformation allows the use of standard (unconstrained) regression models in R^{m-1}. Moreover, in many practical applications, the last coordinate in the Simplex, x_m, is a significant proportion that may represent the preponderant component, an aggregate of many residual or indiscriminated components, a dispersion medium, etc. In this case, the coordinates generated by the additive logratio transformation have an intuitive interpretation. Furthermore, under appropriate conditions, the random variates corresponding to these coordinates in the statistical model have probability distributions with interesting statistical properties, as analysed in [3], [25, Sect. 7]. Finally, the isometric logratio transformation provides orthogonal coordinates in the unrestricted Euclidean space. However, in many practical applications, these orthogonal coordinates have a less intuitive interpretation than the oblique coordinates given by the additive logratio transformation.

Nevertheless, all these approaches will define compatible statistical models, render coherent statistical inferences, and mutually support each other for rich interpretations. In particular, it is easy to translate to the context of compositional data the results obtained in the last section concerning best unbiased predictors and well balanced samples. This will be the starting point of the next section.

14.4 Haphazard Intentional Allocation for Clinical Trials

In the setting discussed in Sect. 14.2, we choose purposively a sample S that represents well the covariates X, that is, so that $(N/n)\mathbf{1}'X_s \approx \mathbf{1}'X$. (We could even seek a sample that, simultaneously, also approximately minimizes the variance of our prediction $\hat{\kappa}$ for $\kappa = \mathbf{1}'y$.) However, at the same time, we would like to use some sort of randomization technique in order to obtain a sample S that is haphazard with respect to the omitted covariates, so that it is (probably) balanced, that is, $(N/n)\mathbf{1}'Z \approx \mathbf{1}'Z$. In this section, we introduce a technique for conciliating these goals, in a clinical trial case study.

The case study discussed in this work is the allocation of patients with obsessive-compulsive disorder (OCD) between two treatment arms, see [12]. Patients are enrolled sequentially, according to the order in which they start the treatment at the clinic or hospital. The allocation problem consists in assigning each new patient to one, and only one, of two alternative treatments (arms). A requisite stated by the trial coordinators is that profiles in the alternative arms remained similar with respect to some relevant patients' factors. In other words, it was expected that the compositional vectors (i.e. relative frequencies of patients in each variable category) remained similar each other as new patients were allocated. The available clinical trial dataset consists of $T = 277$ patients.

Roughly speaking, the factors and respective number of classes considered are: (1) Current patient's *age* (a): three classes; (2) Treatment *history* (h): three classes; (3) OCD symptom *severity* (v): nine classes; and (4) *Gender* (g): two classes. A more detailed description on these factors and respective categories may be found in Fossaluza et al. [12].

After some patients are already in treatment, we denote by n_i^a, n_i^h, n_i^v and n_i^g the quantities of patients already allocated to arm i belonging to each category of factors *age*, *history*, *severity* and *gender*. For example, $n_1^a = [n_{1,1}^a, n_{1,2}^a, n_{1,3}^a]$ denotes the quantity vector of patients in arm 1 belonging to the three age classes.

In order to yield allocations with approximately the same number of patients in each arm, we also consider, besides the previous factors, the sample *size* (z) in each arm. With that purpose we define q_i as the total number of patients allocated to arm i, and the vector of total allocation to arm i and its complement, $n_i^z = [q_i, (q_1+q_2-q_i)]$.

The complete profile of arm i, $i = 1, 2$, is stored in the concatenated vector $n_i = [n_i^a, n_i^h, n_i^v, n_i^g, n_i^z]$. In order to avoid empty categories in the allocation process, we may add to vector n a *ground-state* or *weak-prior*, see [25], in the form of vector $w = [w^a, w^h, w^v, w^g, w^z]$. For any character ξ in the set $\{a, h, v, g, z\}$, where factor w^ξ has $\kappa(\xi)$ categories, we take $w^\xi = [1/\kappa(\xi), \ldots, 1/\kappa(\xi)]$. From vectors

n and w, we obtain the *regularized proportions* vector: $p_i = [p_i^a, p_i^h, p_i^v, p_i^g, p_i^z]$, where $p_i^\xi = clu(n_i^\xi + w_i^\xi), \xi \in \{a, h, v, g, z\}$.

We define the heterogeneity measure between arms 1 and 2 by the function:

$$\Delta(p_1, p_2) = \frac{1}{5}\left(D_s(p_1^a, p_2^a) + D_s(p_1^h, p_2^h) + D_s(p_1^v, p_2^v) + D_s(p_1^g, p_2^g) + D_s(p_1^z, p_2^z)\right). \tag{14.1}$$

Let us consider a new patient that enrolls the study and must be allocated to one of arms 1 or 2. We denote by x^a, x^h, x^v, x^g and x^z the binary vectors indicating to which categories the new patient belongs. For example, in vector $x^a = [x_1^a, x_2^a, x_3^a]$, $x_k^a = 1$ if and only if the patient belongs to age category $k, k \in [1, \ldots, \kappa(a)]$. Vector x^z is set as $x^z = [1, 0]$. So, the relevant information about the new patient is carried by the vector $x = [x^a, x^h, x^v, x^g, x^z]$.

The arm allocation decision for the new patient is taken as follows.

1. For $j = 1, 2$, consider the allocation of the new patient, x, in arm j, that is, for $i = 1, 2$, make $m_i = n_i + \delta(i, j)x$ and perform the following steps:
   a) For $i = 1, 2$ and $\xi \in \{a, h, v, g, z\}$, compute the regularized proportions
   $$p_i^\xi = clu(m_i^\xi + w_i^\xi);$$
   b) For $i = 1, 2$, set $p_i = [p_i^a, p_i^h, p_i^v, p_i^g, p_i^z]$;
   c) For $i = 1, 2$, set $b_i = [u_i, 1 - u_i]$, where u_i are independently generated from $Uniform(0, 1)$ distribution;
   d) For $\epsilon \in [0, 1]$, compute the ϵ-perturbed distance
   $$d_\epsilon(j) = (1 - \epsilon)\Delta(p_1, p_2) + \epsilon D_s(b_1, b_2).$$
2. Choose the allocation j that minimizes $d_\epsilon(j)$, assign the new patient to the corresponding arm, and update vector n accordingly.

The perturbation parameter ϵ introduces a random component in the allocation method. The higher the value of ϵ, the higher the proportion of randomness in the allocation. For $\epsilon = 0$, we have a deterministic intentional allocation scheme, as described in [12], and for $\epsilon = 1$, we have the pure random allocation method, which consists in assign each patient randomly (with probability 0.5) to one of the two arms.

14.5 Numerical Experiments

In order to evaluate the performance of our allocation procedure, we conducted some numerical experiments described below.

We generated $P = 300$ random permutations of the original OCD data, each one representing a possible sequence of patients arriving to the hospital or clinic. For each random permutation and for each value of $\epsilon \in \{0.005, 0.01, 0.05, 0.25, 1\}$, we

ran the haphazard intentional allocation method $H = 300$ times, each one expected to yield a different allocation configuration. At each patients' permutation we also ran once the deterministic intentional sampling procedure, by setting $\epsilon = 0$.

Two criteria were used to analyse the performance of the haphazard intentional allocation method: *Optimality* and *Decoupling*. The first criterion, *Optimality*, is based on the distance Δ defined in Eq. 14.1 and concerns the difference among the relative frequencies of patients in the several categories for both arms.

The second criterion, *Decoupling*, concerns the absence of a tendency to allocate each pair patients to the same arm. In this work, we use the Yule's coefficient of association (Q), see [34], in the following way. After each batch of H runs of haphazard allocations, for each pair of patients, A and B, we build a 2×2 contingency table z where z_{ij} denotes the number of runs patient A was assigned to arm i and patient B was assigned to arm j ($i, j = 1, 2$).

The Yule' coefficient, given by $Q = (z_{11}z_{22} - z_{21}z_{12})/(z_{11}z_{22} + z_{21}z_{12})$, measures the balance among the number of pairs in agreement and disagreement. It ranges in the interval $[-1, 1]$; equals zero when the numbers of agreement and disagreement pairs is equal; and is maximum (-1 or $+1$) in the presence of total negative (complete disagreement) or positive (complete agreement) association.

Figures 14.1 and 14.2 present the 5, 25, 50, 75, 95% empirical percentiles of Δ and Q, respectively. In Fig. 14.1, the quantiles for Δ span the H haphazard allocations. In Fig. 14.2, the quantiles for Q span the $T(T-1)/2$ pairs of patients, where the value of Q for each pair is computed over the H haphazard allocations. Each bar height corresponds to the median over the P random permutations, and the vertical line in each bar represent the corresponding (5%, 95%) percentiles. Continuous and dashed horizontal lines in Fig. 14.1 represent, respectively, the median of distance Δ for the deterministic intentional allocation method, $\epsilon = 0$, and the (5%, 95%) percentiles over P random permutations. Figure 14.2 omits the percentile 50%, since the Yule's coefficient medians were close to zero for all allocation methods.

Figure 14.1 shows a clear difference between the optimality (Δ) achieved by the haphazard intentional allocation method and the pure random method. Notice that, for $\epsilon \leq 0.01$, even the 95% percentiles of Δ for the haphazard intentional method are lower than the 5% percentile for the pure random method.

We also notice that, for the same range of ϵ, the distribution of distances Δ achieved by the haphazard intentional method comes close to the distribution provided by the deterministic method, only showing moderate degradation in the 95% percentile.

Figure 14.2 shows that, for the lower range of ϵ, the absolute values of Yule's association coefficient (Q) tend to be high, indicating that the haphazard intentional allocation method, with too small an ϵ, tends to allocate the same pairs of patients in the same arms, that is, it fails to achieve the desired decoupling property. On the other hand, moderate values of ϵ attenuate these dependencies, making the haphazard intentional allocation method perform in the decoupling criterion almost as well as the the pure random method. Indeed, for $\epsilon \geq 0.05$, the distribution of Yule's Q is very close to the pure random method, which provides our benchmark for decoupling performance.

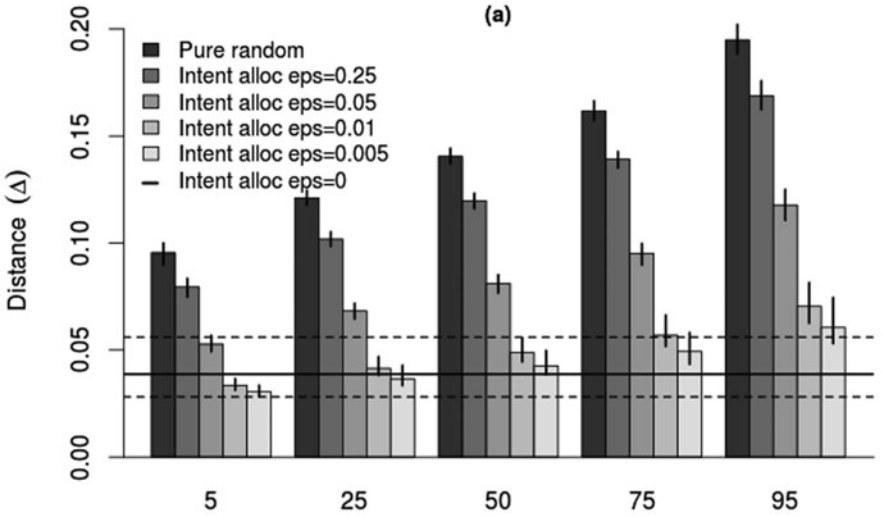

Fig. 14.1 The 5, 25, 50, 75 and 95% percentiles for Δ optimality, with $\epsilon \in \{0, 0.005, 0.01, 0.05, 0.25, 1\}$

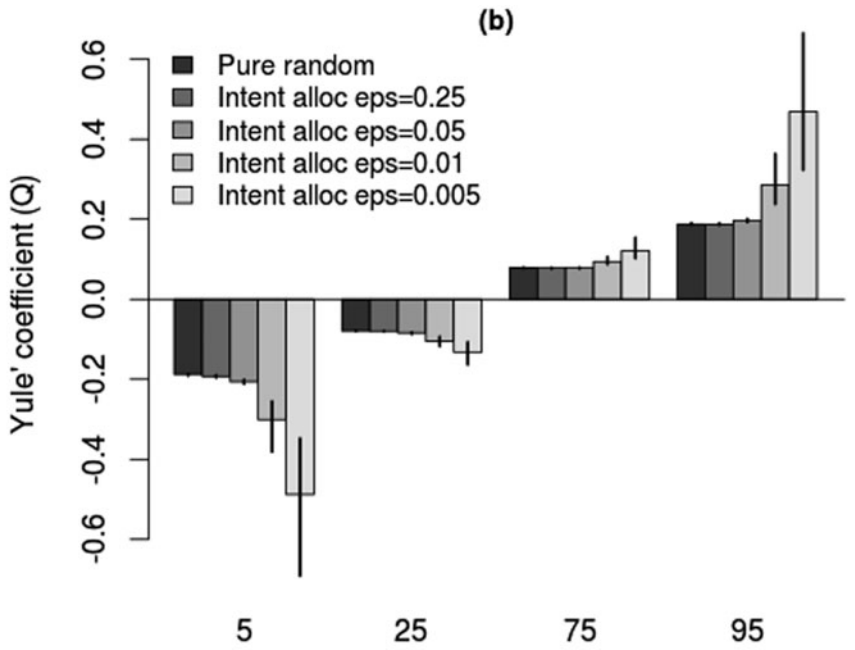

Fig. 14.2 The 5, 25,50,75 and 95% percentiles for Yule's Q decoupling, with $\epsilon \in \{0, 0.005, 0.01, 0.05, 0.25, 1\}$

It is worth mentioning that, for the haphazard intentional allocation method, the percentiles intervals (represented by vertical lines in each bar) are, in general, similar to the pure random method (except for the 5 and 95% percentiles of Q for $\epsilon \leq 0.01$). This result suggests that the proposed method is highly adaptive, in the sense that its performance indicators (optimality and decoupling) are little sensitive to patients permutations.

The results of these numerical experiments indicate that, under an appropriate calibration of the perturbation parameter ϵ, the haphazard intentional allocation method proposed in this paper has the remarkable property of being able to conciliate the performance on optimality achieved by the deterministic intentional allocation with the performance on decoupling achieved by the pure random allocation method.

14.6 Acknowledgements and Final Remarks

The present article does not include Bayesian models incorporating prior information although, formally, such models only generalize the implicit uninformative priors. Despite our focus in this paper have been on the conceptual and practical discussions concerning randomization and intentional sampling, it is perfectly possible to extend our analysis to more general Bayesian models, a work we intend to do following [5]. The article *Six Approaches to Enumerative Survey Sampling*, by Brewer and Särndal [8], see also [27, Sect. 6.2], has been used for the last 30 years as a classification scheme concerning, among other things, the role of randomization in survey sampling. However, it is not straightforward to fit the allocation method we have just presented in that classification scheme. We hope to explore this theme in following articles.

As duly noted by an anonymous referee, an important topic for further research concerns a comparative analysis of the logical status of all the aforementioned randomization, intentional and mixed-randomization sampling methods according to several possible theoretical and epistemological frameworks. In the standard Bayesian decision theoretical framework, randomization methods may not be incompatible with optimal decisions, and may even be able to address some extra-theoretical demands. However, they can never find a direct intra-theoretical justification; see [9, Sect. 8.5, pp. 128–130]. Nevertheless, the standard decision theoretical framework can be expanded by taking into account games with multiple adversarial opponents, see [19, 20], [29, Sect. 6.8]. Bonassi et al. [6] explore this expanded decision-theoretical framework, survey the pertinent literature, and suggest interesting approaches for further development. Finally, the function and logical status of randomization methods can be analyzed in the framework of systems theory, see for example [22, pp. 16–20, 340–348] and [28].

All the aforementioned theoretical frameworks offer alternative ways to deal with concepts related to decoupling, separation, shielding from undue interference, or defensive strategies against players with hostile objectives. Hence, logical analyses conducted in these frameworks should be able to provide guidelines for the coherent

development and application of intentional but haphazard sampling methods in the scope of Bayesian statistics.

The anonymous referee also stresses the importance of carefully analyzing the ethical consequences of using alternative sampling methods. In the context of clinical trials, the experimenter must always consider at least two competing (if not conflicting) objectives: On one hand, the primary objective of any clinical trial is the acquisition of valid or objective knowledge, see; Chap. 5. On the other hand, providing appropriate health care for all patients participating in the trial is a second goal that should never be neglected, see [13, 14]. The use of intentional sampling methods is a technical solution that has the potential of facilitating the reconciliation of such multiple objectives. Moreover, these considerations can be extended to multi-phase and adaptive trials. All these are topics that certainly deserve further research.

The authors are grateful for support received from EACH-USP, the School of Humanities Arts and Sciences; IME-USP, the Institute of Mathematics and Statistics of the University of São Paulo; FAPESP, the São Paulo Research Foundation (grants Reg-2012/04788-9 and CEPID-2013/07375-0); and CNPq, the Brazilian National Council for Scientific and Technological Development (grants PQ-306318-2008-3 and PQ-302046-2009-7).

References

1. Aitchison, J.: The Statistical Analysis of Compositional Data. Chapman & Hall, London (1986)
2. Aitchison, J.: The single principle of compositional data analysis, continuing fallacies, confusions and misunderstandings and some suggested remedies CODAWORK08 –3rd Compositional Data Analysis Workshop, Girona (2008)
3. Aitchison, J., Shen, S.M.: Logistic-normal distributions: some properties and uses. Biometrika **67**, 261–272 (1980)
4. Basu, D., Ghosh, J.K. (eds.): Statistical Information and Likelihood, A Collection of Essays by Dr. Debabrata Basu (Lecture Notes in Statistics 45). Springer, New York (1988)
5. Bolfarine, H., Pereira, C.A.B., Rodrigues, J.: Robust linear prediction in finite populations: a Bayesian perspective. Sankhya, B **49**(1), 23–35 (1987)
6. Bonassi, F.V., Nishimura, R., Stern, R.B.: In defense of randomization: a subjectivist Bayesian approach. AIP Conf. Proc. **1193**, 32–39 (2009)
7. Brewer, K.R.W.: Combined Survey Sampling Inference: Weighing of Basu's Elephants. Hodder Arnold Publication, London (2002)
8. Brewer, K. R. W., Särndal, C. E.: Six approaches to enumerative survey sampling. Incomplete Data Sample Surv. **3**, 363–368 (1983)
9. DeGroot, M.H.: Optimal Statistical Decisions. McGraw-Hill, New York (1970).
10. Egozcue, J.J., Daunis-i-Estadella, J., Pawlowsky-Glahn, V., Hron, K., Filzmoser, P.: Simplicial regression: the normal model. J. Appl. Probab. Stat. **6**, 87–108 (2001)
11. Fisher, R.A.: The Design of Experiments, 8th edn (1966). Oliver and Boyd, London (1935)
12. Fossaluza, V., Diniz, J.B., Pereira, B.B., Miguel, E.C., Pereira, C.A.B.: Sequential allocation to balance prognostic factors in a psychiatric clinical trial. Clinics **64**, 511–518 (2009)
13. Kadane, J.B.: Bayesian Methods and Ethics in a Clinical Trial Design. Wiley, New York (1996)
14. Kadane, J.B., Sedransk, N.: Toward a more ethical clinical trial. Trabajos de Estadistica Y de Investigacion Operativa **31**(1), 329–346 (1980).
15. Lauretto, M.S., Nakano, F., Pereira, C.A.B., Stern, J.M.: Intentional Sampling by goal optimization with decoupling by stochastic perturbation. AIP Conf. Proc. **1490**, 189–201 (2012)

16. Lindley, D.V.: The role of randomization in inference. In: Asquith, P., Nickles, T. (eds.) Proceedings of the Biennial Meeting of the Philosophy of Science Association, Vol. 2, 431–446. University of Chicago Press (1982)
17. Lindley, D.V.: Making Decisions. Wiley, New York (1991)
18. Madow, W.G., Olkin, E., Rubin, D.B.: Incomplete Data in Sample Surveys (Vol. 3), Academic Press, New York (1983)
19. Morgenstern, O.: Game Theory. Dictionary of the History of Ideas Vol.2 p. 264–275 (2008)
20. Morgenstern, O., Neumann, J.: The Theory of Games and Economic Behavior. Princeton University Press, Princeton (1947)
21. Pawlowsky-Glahn, V., Egozcue, J.J.: Geometric approach to statistical analysis on the simplex. Stoch. Environ. Res. Risk Assess. **15**(5), 384–398 (2001)
22. Pearl, J.: Causality: Models, Reasoning, and Inference. Cambridge University Press, Cambridge (2000)
23. Peirce, C.S., Jastrow, J.: On small differences of sensation. den Mem. Natl. Acad. Sci. **3**, 75–83 (1884)
24. Pereira, C.A.B., Rodrigues, J.: Robust linear prediction in finite populations. Int. Stat. Rev. **3**, 293–300 (1983)
25. Pereira, C.A.B., Stern, J.M.: Special characterizations of standard discrete models. RevStat Stat. J. **6**, 199–230 (2008)
26. Royall, R.M., Pfeffermann, D.: Balanced samples and robust Bayesian inference in finite population sampling. Biometrika **69**(2), 401–409 (1982)
27. Schreuder, H.T., Gregoire, T.G., Wood, G.B.: Sampling Methods for Multiresource Forest Inventory. Wiley, New York (1993)
28. Stern, J.M.: Decoupling, sparsity, randomization, and objective Bayesian inference. Cybern. Hum. Knowing **15**, 49–68 (2008)
29. Stern, J.M.: Cognitive Constructivism and the Epistemic Significance of Sharp Statistical Hypotheses in Natural Sciences. arXiv:1006.5471 (2011).
30. Tam, S.M.: Characterization of best model-based predictors in survey sampling. Biometrika **73**(1), 232–235 (1986)
31. Valliant, R., Dorfman, A.H., Royall, R.M.: Finite Population Sampling and Inference: A Prediction Approach. Wiley, New York (2000)
32. Whittle, P.: Probability via Expectation. Springer, New York (2000)
33. Yule, G.U.: Notes on the theory of association of attributes in statistics. Biometrika **2**(2), 121–134 (1903)
34. Yule, G.U.: On the methods of measuring association between two attributes. J. R. Stat. Soc. **75**(6), 579–652 (1912)

Chapter 15
Factor Analysis with Mixture Modeling to Evaluate Coherent Patterns in Microarray Data

Joao Daniel Nunes Duarte and Vinicius Diniz Mayrink

Abstract The computational advances over the last decades have allowed the use of complex models to analyze large data sets. The development of simulation-based methods, such as the Markov chain Monte Carlo (MCMC) method, has contributed to an increased interest in the Bayesian framework as an alternative to work with factor models. Many studies have applied the factor analysis to explore gene expression data with results often outperforming traditional methods for estimating and identifying patterns and metagene groups related to the underlying biology. In this chapter, we present a sparse latent factor model (SLFM) using a mixture prior (sparsity prior) to evaluate the significance of each factor loading; when the loading is significant, the effect of the corresponding factor is detected through patterns displayed along the samples. The SLFM is applied to investigate simulated and real microarray data. The real data sets represent the gene expression for different types of cancer; these include breast, brain, ovarian, and lung tumors. The proposed model can indicate how strong is the observed expression pattern allowing the measurement of the evidence of presence/absence of the gene activity. Finally, we compare the SLFM with two simpler gene detection methods available in the literature. The results suggest that the SLFM outperforms the traditional methods.

15.1 Introduction

In recent years, Bayesian framework has become an important alternative to explore factor analysis, being the main reason, the development of iterative Markov chain Monte Carlo (MCMC) simulation methods, which have allowed the use of complex models and large data sets. In particular, Bayesian factor modeling has brought interesting improvements to the analysis of gene expression data. In the literature, we can easily find studies applying this model on microarray data and, in most cases, their results show better identification of patterns and hidden genomic structures.

J. D. Nunes Duarte (✉)· V. D. Mayrink
Departamento de Estatistica, ICEx, UFMG, Av. Antonio Carlos 6627, Belo Horizonte, MG, Brazil
e-mail: joaodaniel@ufmg.br

V. D. Mayrink
e-mail: vdm@est.ufmg.br

A. Polpo et al. (eds.), *Interdisciplinary Bayesian Statistics,*
Springer Proceedings in Mathematics & Statistics 118, DOI 10.1007/978-3-319-12454-4_15

For example, West [16] introduced sparse latent factor models (SLFM) as a natural extension of the sparse regression model. The study applies prior distributions for variable selection and demonstrates the ability of latent factor models to describe the pattern/profile of signatures on genomic expressions. Lucas et al. [6] also applied sparse hierarchical prior distributions and obtained substantial improvements in the identification of complex patterns of covariance between genes. The paper explores the Bayesian methodology for large-scale regression, analysis of variance, and latent factor models. Carvalho et al. [2] used sparsity priors for addressing the dimension reduction in latent factor models applied to gene expression data. Stochastic simulation and evolutionary stochastic search methods are used to deal with the uncertainty about the number of latent factors. This same problem is studied in [5] using reversible jumps MCMC methods.

In this chapter, we improved the SLFM presented in [9] by using a univariate distribution for each element of the α vector, thus giving the model more flexibility on pattern detection. We also implemented this model in the statistical software R [12].

The focus of this study is to perform inferences with SLFM. We will apply this model to capture the underlying structure of gene expression data and use results of variable selection to classify arrays of data regarding the presence or absence of a particular gene. This model will be assessed in terms of: (i) quality of parameter estimation and (ii) quality of classification of gene presence.

We use real gene expression data from breast, lung, ovarian, and brain tumors. We also explore simulated data to evaluate model accuracy; in this case, real classification of gene activity is known and chosen by the researcher. Section 15.2 describes the SLFM and indicates the MCMC setup. Section 15.3 shows the analysis using the sparse latent factor model-univariated (SLFM-U) which assumes the sparsity prior for each loading separately. Finally, Section 15.4 presents the conclusions.

15.2 Sparse latent factor model

Factor analysis has been used for dimension reduction and to study patterns and structures in the data set. The latent factor models, presented in [16], split the variation of predictor variables into components that represent recurring patterns, and separate them from the inherent variation in each variable, which we call noise. Large data sets, e.g., microarray data, require strategies to facilitate the analysis, given the large number of parameters proposed in a usual statistical model for this case. Sparse models are widely used in this context because they are designed to identify zeros or nonsignificant components. An SLFM is discussed below.

15.2.1 The SLFM

Let X be the matrix containing the data set. Assume that X has dimensions $m \times n$, where each row i may represent an individual, a variable , a feature, etc. Each column

j represents a sample (vector of observations). Consider $X = \alpha\lambda + \epsilon$ where α is the loadings matrix, λ is the factor score matrix, and ϵ represents the noise matrix ($m \times n$). The α matrix is $m \times L$ and λ is $L \times n$, where L is the number of factors in the model. We assume that $\epsilon_{ij} \sim N(0, \sigma_i^2)$ and $\sigma^2 = (\sigma_1^2, \sigma_2^2, \ldots, \sigma_m^2)$. Note that the variance of ϵ is indexed by i suggesting that it is constant along the samples but differs from row to row.

Let us assume $L = 1$ to illustrate a simpler case with only one factor. Accordingly, α will be a column vector ($m \times 1$) and λ will be a row vector ($1 \times n$). The following prior distributions are specified: $\lambda_j \sim N(0, 1)$ and $\sigma_i^2 \sim IG(a, b)$. The $N(0, 1)$ is chosen to fix the magnitude of λ and then avoid identification problems related to the values of α and λ. We use a sparse prior for the loading α_i: $\alpha_i \sim (1 - Z_i)\delta_0(\alpha_i) + Z_i N(0, \omega)$, where $Z_i \sim Bernoulli(q_i)$ and $q_i \sim Beta(\lambda_1, \lambda_2)$. As each loading in α is sampled from an univariate distribution, we call this model SLFM-U.

The likelihood can be expressed in two different ways. Calculations of posterior distributions are simplified depending on the chosen version of the likelihood function.

- Likelihood 1: Denote $X_{\cdot j}$ as the jth column of X. Then $(X_{\cdot j}|\alpha, \lambda_j, \sigma^2) \sim N_m[\alpha\lambda_j, D]$, where $D = diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_m^2)$. Assuming conditional independence between the columns of X, we have:

$$p(X|\alpha, \lambda, \sigma^2) = \prod_{j=1}^{n} p(X_{\cdot j}|\alpha, \lambda_j, \sigma^2).$$

- Likelihood 2: Let $X_{i\cdot}$ be the ith row of X. Then $(X'_{i\cdot}|\alpha_i, \lambda, \sigma_i^2) \sim N_n[\lambda'\alpha_i, \sigma_i^2 I_n]$. Supposing conditional independence between rows of X, we have:

$$p(X|\alpha, \lambda, \sigma^2) = \prod_{i=1}^{m} p(X_{i\cdot}|\alpha_i, \lambda, \sigma_i^2).$$

Assuming independence between rows and columns of X, the full conditional distributions used in the Gibbs sampling are as follows: if $Z_i = 0$, then $(\alpha_i|\alpha_{-i}, \lambda, \sigma^2, Z, X) \sim \delta_0(\alpha_i)$; if $Z_i = 1$, then $(\alpha_i|\alpha_{-i}, \lambda, \sigma^2, Z, X) \sim N(M_\alpha, V_\alpha)$, with $V_\alpha = [\frac{1}{\omega} + \frac{1}{\sigma_i^2}\sum_{j=1}^{n} \lambda_j^2]^{-1}$ and $M_\alpha = V_\alpha[\frac{1}{\sigma_i^2}\sum_{j=1}^{n} \lambda_j X_{ij}]$. The parameter q_i represents the prior probability that a given variable has null effect; its full conditional distribution is: $(q_i|Z) \sim Beta(\gamma_1^*, \gamma_2^*)$, with $\gamma_1^* = \gamma_1 + \sum_{i=1}^{m} Z_i$ and $\gamma_2^* = \gamma_2 + m - \sum_{i=1}^{m} Z_i$. The full conditional distribution of $Z_i = 1$ is given by:

$$q_i^* = p(Z_i = 1|\alpha, \lambda, \sigma^2, q_i, X) = \frac{q_i}{q_i + \frac{N[0|M_\alpha, V_\alpha]}{N[0|0, \omega]}(1 - q_i)}. \tag{15.1}$$

In order to calculate the full conditional distribution of λ, we use Likelihood 1 to reach the following formulation: $(\lambda_j|\alpha, \lambda_{-1}, \sigma^2, X) \sim N(M_\lambda, V_\lambda)$, where $V_\lambda = (\alpha' D^{-1}\alpha + 1)^{-1}$ and $M_\lambda = V_\lambda(\alpha' D^{-1} X_{\cdot j})$. The full conditional distribution of σ^2, $i = 1, \ldots, m$ is: $(\sigma_i^2|\alpha, \lambda, \sigma_{-i}^2, X) \sim IG(A, B)$, with $A = a + (n/2)$ and $B = b + (1/2)(X_{i\cdot}X'_{i\cdot} - 2\alpha_i\lambda X'_{i\cdot} + \alpha_i\lambda\lambda'\alpha'_i)$.

15.2.2 MCMC and Analysis Criterion

This section discusses the MCMC algorithm used to generate posterior samples of the parameters. The Gibbs sampling takes into account the full conditional distributions, discussed in the previous section, to obtain the samples. The algorithm is as follows: First, iterate over i: (i) sample σ_i^2 from $IG(A, B)$, (ii) sample Z_i, (iii) sample q_i from $Beta(\gamma_1^*, \gamma_2^*)$, (iv) sample α_i from $N(M_\alpha, V_\alpha)$ if $Z_i = 1$, and (v) make $\alpha_i = 0$ if $Z_i = 0$. Finally, iterate over j to sample λ_j from $N(M_\lambda, V_\lambda)$.

In this study, we consider the posterior mean as the point estimate summarizing the information of the chains. However, since we are using a sparse prior for α, it is necessary to estimate q_i in first place and decide whether α_i will be set to zero or estimated from the sample related to the nonzero component of the mixture. The identification of rows not affected by the factor considers the following criterion: The factor effect over the ith row of X is null, if q_i^* in (15.1) is small (less than 0.5) for the ith factor loading. The main aim is to classify data matrices in terms of presence or absence of patterns. This classification will also be based on the posterior probabilities q_i^* in (15.1). The matrix classification follows the criterion: First, calculate the high probability density (HPD) intervals for q_i^*; α_i is said "absent" if the superior limit of the HPD interval is less than 0.5, it is "marginal" if the HPD interval includes to 0.5, and it is "present" if the inferior limit of the HPD interval is greater than 0.5. Finally, calculate the number of loadings in each category; the whole matrix X is said to display a coherent pattern if the category "present" is the most frequent.

In terms of matrix classification for most rows, "present" indicates the occurrence of a consistent pattern, suggesting gene activity and "absent" indicates that no pattern is observed (no activity). The matrix classification "marginal" suggests an intermediate observed pattern which is not strong or weak enough to guide our decision (inconclusive gene activity).

15.3 Gene Expression Analysis

15.3.1 Affymetrix Technology

The data used in the analyses are related to GeneChip Affymetrix microarrays representing the gene activity in different types of tumor cells. In short, the data are obtained through the hybridization property of complementary nucleic acids. Single RNA nucleotide strands, 25 bases long, are extracted from the tumor cell and labeled with a fluorescent tag. In the hybridization procedure, the single RNA strand from the tumor will connect to its complementary sequence built in a small chip or array, if such sequence is considered in the array. Next, the array is washed to remove unmatched material and a laser is applied to activate the microarray fluorescence. The light intensity is scanned and translated into a gene expression measurement which is proportional to gene activity of that sequence in the tumor. It is important

to clarify that each sequence of 25 bases is called a sequence probe representing a fraction of the gene; in this study, we use the term "gene" to refer to a probe set containing 11–20 probes in the microarray.

Before fitting the models it is necessary to preprocess the data to remove noise effects; this is a standard procedure in microarray analysis. The brightness of a chip is subject to distortions as the amount of RNA in the sample and changes in the camera exposure time. Also, the raw data intensities are skewed and have a long tail. We assume that the intensities of each probe follows the normal distribution; the mean and variance may vary from probe to probe. The scanner calibration can also generate unwanted effects, which occurs in all microarrays. To deal with those issues, we used the following steps: (i) divide the probe intensities on each array by the mean intensity of the corresponding array, (ii) transform data using the natural logarithm, (iii) normalize the rows, and (iv) calculate the first principal component of the whole data set and subtract it from the expressions in each row of X, i.e., $X_i - X_i \times pc_1 \times pc_1'$. The data matrix used to fit the models is the result of this procedure.

15.3.2 Simulated Data Analysis I

In order to study the performed of the model, we performance a simulated study. In this case, real values of the target parameters are known and were chosen to generate the data. Performance is assessed by comparison of estimated and real values. Data was generated via the following steps: (i) independently sample 251 λ_j's from $N(0,1)$, (ii) independently sample 40 α_i's from $N(0,1)$, (iii) randomly set 5 α_i's as zeros, (iv) sample 40 σ_i^2 from $IG(2.1, 1.1)$, (v) generate the ϵ_{ij}'s from $N(0,\sigma^2)$, and (vi) calculate $X = \alpha\lambda' + \epsilon$. The data matrix X has dimensions 40×251, the same size of a real matrix considered ahead. Preprocessing is not required in this simulated case.

Figure 15.1 shows the real values of the parameters, the estimated posterior mean, and the HPD intervals (95 %). As it can be seen, the model was effective in estimating the parameters; most of the HPD intervals contain the real values. The SLFM incorrectly classifies 7 α_i's out of 40 loadings, i.e., the success rate of classification was 82.5 %.

15.3.3 Simulated Data Analysis II

This simulated study is concerned with the classification of the whole matrix representing a probeset. The procedure to generate the data follows the steps: (i) compute, for all 22,283 probesets, the correlation matrix expressing the association between rows of a real matrix X; this gives us an idea of how strong would be the pattern in each data matrix; (ii) select 500 matrices (200 with low correlation and 300 with

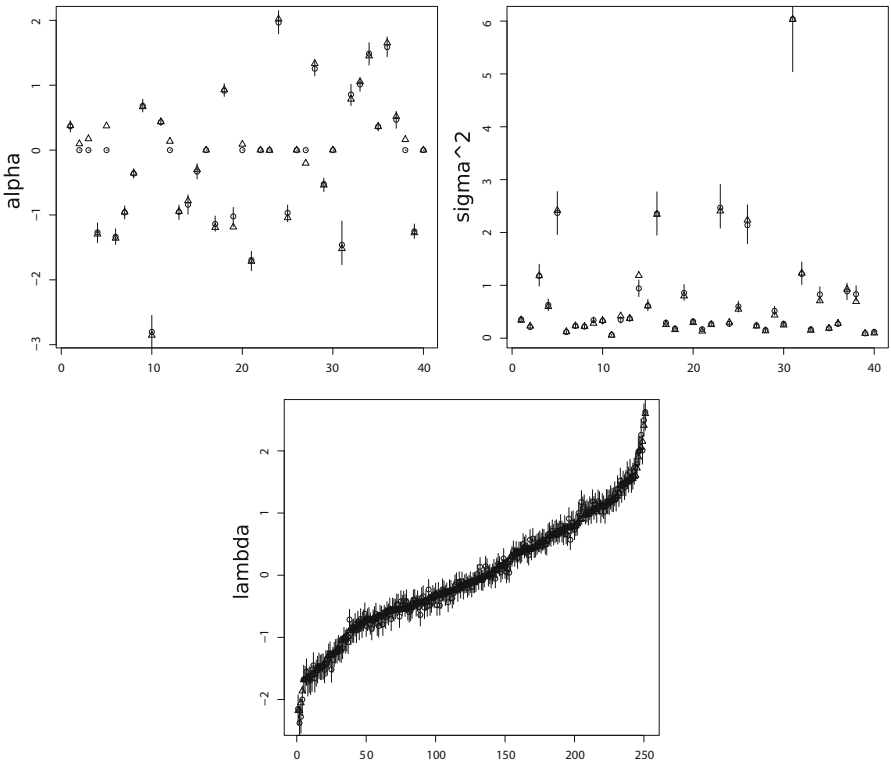

Fig. 15.1 Performance of the model: real values (*triangles*), posterior mean values (*circles*), and the HPD interval represented by the *line segment*. Parameters: α (*left*); λ (*center*) σ^2 (*right*)

Table 15.1 Comparison of false positive rate (FPR), false negative rate (FNR), true positive rate (TPR), and true negative rate (TNR) for MAS 5.0 P/A and SLFM-U

Model	FPR (%)	FNR (%)	TPR (%)	TNR (%)
MAS 5.0 P/A	45.5	8.6	91.3	54.5
SLFM-U	0	0	100	100

intermediate to strong correlation); (iii) fit the SLFM-U, for those 500 matrices, and save the parameter estimates $\{\hat{\alpha}, \hat{\lambda}, \hat{\sigma^2}\}$ as reasonable choices of real values; and (iv) generate 500 simulated matrices using $X = \hat{\alpha}\hat{\lambda} + \epsilon$; for the 200 low pattern cases, set $\alpha = \mathbf{0}$. A good model should be able to correctly identify the status: "absent" (200 cases with $\alpha = 0$) or "present" (300 cases with intermediate to strong pattern).

Table 15.1 shows that SLFM-U was able to classify 100 % of the matrices correctly, i.e., 0 % of false positives and 100 % of true positives. It also shows that MAS 5.0 P/A had 8.6 % of false negative and 45.5 % of false positives, suggesting that SLFM-U outperforms the MAS 5.0 P/A. Figure 15.2 shows a graph representing the proportions of "presence" for each matrix. The first 200 points in the level 0 are

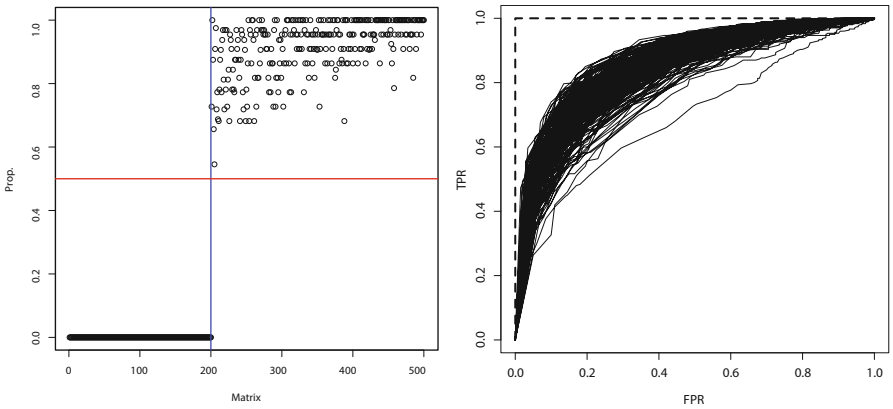

Fig. 15.2 *Left*: Proportion of "presence" in each matrix; *right*: ROC curve comparing MAS 5.0 P/A (*continuous*) and SLFM-U (*dashed*)

related to low pattern matrices (in these cases, 0 is the proportion of rows displaying significant patterns). The following 300 cases are related to intermediate-strong pattern matrices where the proportion of significant rows are above 0.5 for all cases. Figure 15.2 also presents ROC curves indicating that SLFM-U was able to achieve higher TPR and lower FPR than MAS 5.0 P/A.

15.3.4 Real Data Analysis I

In this study, we compare the SLFM-U to other two detection methods (MAS 5.0 P/A and PANP) proposed in the literature. The MAS 5.0 P/A is a method developed by *Affymetrix* and it is built within the MAS 5.0 preprocessing software. This method uses information from perfect match (PM) and mismatch (MM) probes to calculate the score $R = (PM - MM)/(PM + MM)$. The Wilcoxon signed rank test is applied to the score data and, based on the p-value, the matrices are classified as "present," "marginal," or "absent." The PANP method, proposed by [15], uses information from the negative strand matching probes NSMPs) to build an empirical cumulative density function; based on an arbitrary cut point, the data matrices (probe sets) are classified into one of those three categories.

The data set used in this analysis was developed by *Affymetrix* and can be found at: http://www.affymetrix.com/support/technical/sample_data/datasets.affx. This data set consists of 3 replicates of 14 separate hybridizations of 42 DNA transcripts from the human genome. Solutions with different concentrations (0–512 pm) of 42 target sequences were used in the hybridization. In this context, a good detection method should be able to identify the 42 target sequences as present in the DNA transcription.

This study has also been performed by [9], but using a Bayesian factor model with a single multivariate mixture prior to the whole vector α (SLFM-M, the M stands

for the multivariate configuration of the sparsity prior). Their results are very similar to the ones obtained here: the PANP was able to identify 16 sequences as "present" while the MAS 5.0 P/A indicates 14 cases as "present"; the SLFM-U proposed here correctly identifies all 42 target genes as "present". Therefore, SLFM-U outperforms PANP and MAS 5.0 P/A in this particular data analysis. PANP and MAS 5.0 P/A are very simple methods whose classification depends on an arbitrary choice of cutoff points. In addition, these methods provide a P/M/A call for each array or sample, whereas, the SLFM-U takes into account the coherent patterns displayed along all arrays to generate its call; the most frequent call generated by PANP and MAS 5.0 P/A across all samples is used to summarized their detection results for each gene. Given their simplicity, PANP and MAS 5.0 P/A have the advantage of being computationally faster to run; however, the SLFM-U has a better performance and may be implemented in a more effective way reduce its computational cost.

We also have used SLFM-U to analyze all the 22,300 probesets from this data set and compare the results with the other methods. The MAS 5.0 P/A identifies 46.9 % of the genes as "present," the PANP method identifies 99.8 % of presence while SLFM-U recognized 1.9 % of the genes as "present." Since the correct classification is not known, it is not possible to evaluate the performance of the methods for the whole data set, but here we can see that the SLFM-U is more restrained in a sense that it indicates the "presence" status only if the the data provides strong evidence for this.

15.3.5 Real Data Analysis II

In this study, we evaluate the influence of copy number alteration (CNA) on the gene expression of four different types of cancer. Previous studies [7, 11] have identified for breast cancer, chromosome regions associated with the occurrence of CNA. This type of abnormality may have important effects on the evolution of cancer. We will check if the genes located in a chromosomal region with CNA, identified in [7] for breast cancer, also indicate the occurrence of CNA in other types of cancer (lung, ovarian, and brain tumors).

Figure 15.3 shows that for each data set approximately half of the genes were classified as "present." We have found that genes 5, 11, 15, and 22 were classified as "present" in all seven data sets. We have analyzed a second region with CNA, and the results are similar: three genes of region 2 were classified as "present" in all seven data sets.

15.4 Conclusions

In this study, we have discussed an SLFM-U to investigate how strong are the patterns exhibited across samples/arrays. The coherent pattern across samples is an important source of information to evaluate the gene activity. The probes (fractions of the

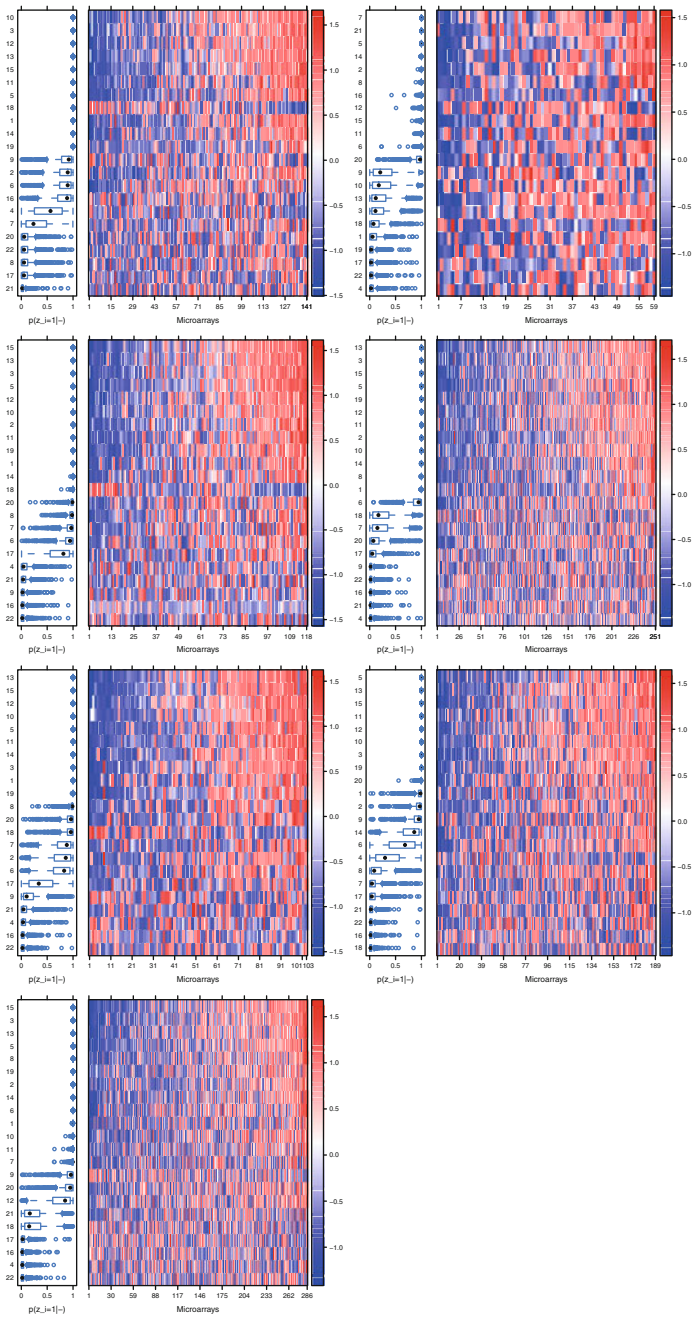

Fig. 15.3 CNA analysis of genes in a chromosomal region for seven data sets representing four types of cancer. The *boxplots* represent the posterior distribution of the probabilities q_i^*, used in our criterion to evaluate the significance of α_i

gene sequence) tend to behave similarly, in terms of gene expression, within the same sample when the gene is present; combining their expressions along samples would display a coordinated pattern. The SLFM-U takes advantage of this type of information to generate a P/M/A detection call.

We have implemented the Gibbs sampling algorithm for the posterior inference related to the SLFM-U. Our main goal was to use the SLFM-U in the gene expression classification problem. This study have considered simulated and real data sets to explore the model results. In the first simulated study, we have generated data to evaluate how accurate is the parameter estimation for the Bayesian model. Looking at the plots comparing the real and estimated values, we can conclude that the model performs well, since most of the HPD intervals contain the true values. In the second simulated study, we have generated data to assess the matrix classification regarding the presence of patterns. The SLFM-U outperformed the MAS 5.0 P/A, achieving 100 % of true positives with 0 % of false positives in our data analysis.

In the first real data analysis, we have used data previously studied in [9]; our results agree with those presented in this paper. In brief, we have 42 arrays hybridized with solutions containing different concentrations of 42 target sequences; thus, these 42 sequences are supposed to be classified as "present." The SLFM-U has correctly identified all 42 cases as "present," while other methods (MAS 5.0 P/A and PANP) provide wrong detections. We have also compared the classification of 22,300 probe-sets using SLFM-U, MAS 5.0 P/A and PANP; the three methods indicate different results with the percentage of presence cases being much lower for the SLFM-U.

In the second real data analysis, we have fitted the model to study CNA in different types of cancer. In the presence of CNA, the expression of neighbour genes tend to behave coherently leading to a strong pattern in our data matrix; here, each row represents a whole probe set. In the literature, a group of genes was identified in a chromosomal region with CNA for breast cancer. The purpose of this study was to determine whether this CNA is also observed for the same group of genes in other type of cancer. Our results indicate that some genes are affected by CNA in all four types of cancer. This conclusion is based on the analysis of coherent patterns displayed in each data matrix; the boxplots located above 0.5 provide evidence of CNA.

References

1. Bild, A.H., Yao, G., Chang, J.T., Wang, Q., Potti, A., Chasse, D., Joshi, M.B., Harpole, D., Lancaster, J.M., Berchuck, A., Jr, J.A.O., Marks, J.R., Dressman, H.K., West, M., Nevins, J.R.: Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature **439**(19), 353–357 (2006)
2. Carvalho, C., Chang, J., Lucas, J.E., Nevins, J.R., Wang, Q., West, M.: High-dimensional sparse factor modeling: applications in gene expression genomics. J. Am. Stat. Assoc. **103**(484), 1438–1456 (2008)
3. Chin, K., DeVries, S., Fridlyand, J., Spellman, P.T., Roydasgupta, R., Kuo, W.L., Lapuk, A., Neve, R.M., Qian, Z., Ryder, T., Chen, F., Feiler, H., Tokuyasu, T., Kingsley, C., Dairkee, S., Meng, Z., Chew, K., Pinkel, D., Jain, A., Ljung, B.M., Esserman, L., Albertson, D.G.,

Waldman, F.M., Gray, J.W.: Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. Cancer Cell **10**, 1149–1158 (2006)
4. Freije, W.A., Castro-Vargas, F.E., Fang, Z., Horvath, S., Cloughesy, T., Liau, L.M., Mischel, P.S., Nelson, S.F.: Gene expression profiling of gliomas strongly predicts survival. Cancer Res. **64**, 6503–6510 (2004)
5. Lopes, H.F., West, M.: Bayesian model assessment in factor analysis. Stat. Sin. **14**, 41–67 (2004)
6. Lucas, J.E., Carvalho, C., Wang, Q., Bild, A., Nevins, J.R., West, M.: Sparse statistical modelling in gene expression genomics. In: Muller P., Do K., Vannucci M. (eds.) Bayesian Inference for Gene Expression and Proteomics, pp. 155–176. Cambridge University Press, Cambridge (2006)
7. Lucas, J.E., Carvalho, C.M., Chen, J.L.-Y., Chi, J.-T., West, M.: Cross-study projections of genomic biomarkers: an evaluation in cancer genomics. PLoS ONE. **4**(2), e4523. (2009). doi:10.1371/journal.pone.0004523
8. Marks, J.R., Davidoff, A.M., Kerns, B.J., Humphrey, P.A., Pence, J.C., Dodge, R.K., Clarke-Pearson, D.L., Iglehart, J.D., Bast, R.C., Berchuck, A.: Overexpression and mutation of p53 in epithelial ovarian cancer. Cancer Res. **51**, 2979–2984 (1991)
9. Mayrink, V.D., Lucas, J.E.: Bayesian factor models for the detection of coherent patterns in gene expression data. Braz J Probab Statistic. **29**(1), 1–33 (2015)
10. Miller, L.D., Smeds, J., George, J., Vega, V.B., Vergara, L., Ploner, A., Pawitan, Y., Hall, P., Klaar, S., Liu, E.T., Bergh, J.: An oncogenic signature for p53 status in human breast cancer predicts mutation status, transcriptional effects and patient survival. Proc. Natl. Acad. Sci. U S A **102**(38), 13550–13555 (2005)
11. Pollack, J.R., Sorlie, T., Perou, C.M., Rees, C.A., Jeffrey, S.S., Lonning, P.E., Tibshirani, R., Botstein, D., Dale, A.L.B., Brown, P.O.: Microarray analysis reveals a major direct role of DNA copy number alteration in transcriptional program of human breasts tumors. Proc. Natl. Acad. Sci. U S A **99**(20), 12963–12968 (2002)
12. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna (2014). http://www.R-project.org
13. Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Fox, S., Smeds, J., Nordgren, H., Farmer, P., Praz, V., Kains, B.H., Desmedt, C., Larsimont, D., Cardoso, F., Peterse, H., Nuyten, D., Buyse, M., Vijver, M.J.V.D., Bergh, J., Piccart, M., Delorenzi, M.: Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. J. Natl. Cancer Inst. **98**(4), 262–272 (2006)
14. Wang, Y., Klijn, J.G.M., Zhang, Y., Sieuwerts, A.M., Look, M.P., Yang, F., Talantov, D., Timmermans, M., Gelder, M.E.M.V., Yu, J., Jatkoe, T., Berns, E.M.J.J., Atkins, D., Foekens, J.A.: Gene expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet **365**, 671–679 (2005)
15. Warren, P.D., Taylor, P.G.V., Martini, J.J., Bienkowska, J.: Panp—a new method of gene detection on oligonucleotide expression arrays. Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering pp. 108–115 (2007)
16. West, M.: Bayesian factor regression models in the large p, small n paradigm. Bayesian Statistics, Oxford University Press 7 (2003)

Chapter 16
Bayesian Hypothesis Testing in Finite Populations: Bernoulli Multivariate Variables

Brian Alvarez R. de Melo and Luis Gustavo Esteves

Abstract Bayesian hypothesis testing for the (operational) parameter of interest in a Bernoulli (multivariate) process observed in a finite population is the focus of this study. We introduce statistical test procedures for the relevant parameter under the predictivistic perspective of Bruno de Finetti in contrast with the usual superpopulation models. The comparison between these approaches, exemplified in a simple scenario of majority elections, shows considerable differences between the corresponding results for the case of observed large sampling fractions.

16.1 Introduction

Finite population models are commonly used when the size of the population under investigation, N, is known. When we do not know the value of N, superpopulation models are used instead. Applications of finite population models usually take place in situations where the population is small (e.g., when we want to study characteristics of a number of industries or banks of a certain town) or the sampling fraction [10] is large (for instance, when forecast about the final result of an election must be performed based on the partial counting of votes at advanced stage). In such conditions, it is expected that inferences drawn from these models are more accurate than those from superpopulation models.

In this chapter, we consider a population composed of N units, $P = \{1, 2, \ldots, N\}$. A real number (or vector) is associated to each unit in such a way that we have the vector of population values $\mathbf{X} = (X_1, X_2, \ldots, X_N)$, where X_i denotes the characteristic of interest of the ith individual, $X_i \in \mathbb{R}^k$, $k \geq 1$. We assume that the vector $\mathbf{X}$ is unknown, its distribution is exchangeable, that is, the order of the population units does not change the distribution of $\mathbf{X}$ [5], and that the population being studied is closed, in the sense that the only random quantities under consideration are $X_1, \ldots X_N$ [4].

B. A. R. de Melo (✉) · L. G. Esteves
Institute of Mathematic and Statistics, University of Sao Paulo, Sao Paulo, Brazil
e-mail: brian@ime.usp.br

L. G. Esteves
e-mail: lesteves@ime.usp.br

A. Polpo et al. (eds.), *Interdisciplinary Bayesian Statistics,*
Springer Proceedings in Mathematics & Statistics 118, DOI 10.1007/978-3-319-12454-4_16

In finite population models, inferences are made about operational parameters, which are functions of the vector $\mathbf{X}$. For instance, if $\mathbf{X} = (X_1, X_2, \ldots, X_N) \in \mathbb{R}^N$, $max\{X_1, \ldots, X_N\}$ and $X_1 + \ldots + X_N$ are operational parameters. The inference about the operational parameter is carried out by determining its posterior distribution (the prior is derived directly from the uncertainty on $\mathbf{X}$) given the observed values from the sample (in this chapter, we then test certain hypotheses of interest). The advantages of working with the predictivistic approach [4], and therefore with operational parameters, are that the parameters are observable quantities, as opposed to abstract quantities, such as limits of relative frequencies, and there is no need to define a parametric space [9]. The predictivistic perspective was introduced by Bruno de Finetti and is considered a more "pure" Bayesian approach [5].

We consider the scenario of a small majority election with three candidates, as described by Fossaluza [3]. We regard the population total(s) as the operational parameter of interest and suppose that a sample of $n < N$ units, $(X_1, \ldots, X_n)$ (without loss of generality as $\mathbf{X}$ is exchangeable), is available. The aim is to test hypotheses of interest concerning the operational parameter under the predictivistic approach and compare the results obtained with those found by adopting the usual Bayesian superpopulation model.

This chapter is organized in the following way: In Sect. 16.2, we examine the case where $X_1, \ldots, X_N$ are Bernoulli random variables indicating if the electors will (or will not) vote for a given candidate. One-sided and simple hypotheses tests are discussed. Section 16.3 deals with the case where $X_1, \ldots, X_N$ represent the preferences of the voters among all the candidates. The possibility of a second round election (if none of the candidates attains a simple majority in the first round) is tested on the basis of the preferences of n electors. Finally, in Sect. 16.4, we make our final comments and point a few questions for future investigation.

16.2 Univariate Case

Consider a situation in which we desire to estimate the amount (proportion) of voters that will vote for a given candidate. We may consider, for each elector, a Bernoulli variable that takes the value 1 if he votes for the specified candidate, and 0 otherwise. Thus, our interest is to make inference about the population total $T(\mathbf{X}) = \sum_{i=1}^{N} X_i$ (or, equivalently, about the population proportion $\frac{T(\mathbf{X})}{N}$). The posterior distribution of the operational parameter $T(\mathbf{X})$, given the sample information $X_1 = x_1, \ldots, X_n = x_n$, $n < N$, under the finite population predictivistic standpoint is presented in the following result.

Result 15.1. *Let* $\mathbf{X} = (X_1, \ldots, X_N)$ *be an exchangeable random vector taking values on* $\{0, 1\}^N$ *according to the probability measure* $\mathbb{P}$ *and define* $p_t = \mathbb{P}(T(\mathbf{X}) = t)$, $t = 0, 1, \ldots, N$, *as the prior uncertainty about the parameter* $T(\mathbf{X})$. *The posterior*

distribution for $T(\mathbf{X})$*, given a sample* $(x_1, \ldots, x_n)$*, is:*

$$P\big(T(\mathbf{X}) = t \big| X_1 = x_1, \ldots, X_n = x_n\big) \propto \frac{\binom{N-n}{t-\sum_{i=1}^{n} x_i}}{\binom{N}{t}} p_t \,,$$

if $t \in \{\sum_{i=1}^{n} x_i, \ldots, N-n+\sum_{i=1}^{n} x_i\}$*, and* $P(T(\mathbf{X}) = t \,|X_1 = x_1, \ldots, X_n = x_n) = 0$*, otherwise.*

The proof of Result 15.1 follows immediately from Bayes theorem and can be found in [1].

On the other hand, when N is large, we usually model the proportion of votes that a candidate will receive instead of the total, taking into account the superpopulation model. Therefore, we may consider a Beta(α,β), $\alpha, \beta > 0$, distribution as the prior uncertainty for the proportion π of interest and assume that the (infinite) random variables $(X_i)_{i\geq 1}$, given π, are conditionally independent Bernoulli random variables with parameter π. Thus, the posterior distribution for the proportion π, given a sample $(x_1, \ldots, x_n)$, is $Beta(\alpha + \sum_{i=1}^{n} x_i, \beta + n - \sum_{i=1}^{n} x_i)$.

Next, we perform hypothesis testing concerning the population total $T(\mathbf{X})$ (proportion π), taking into account N known (unknown), that is, the finite population (superpopulation) model.

16.2.1 One-Sided Hypotheses

First, the interest is to know whether the total (proportion) of votes that a candidate will receive will exceed a specified value. In other words, our goal here is to test the hypothesis $H_0 : T(\mathbf{X}) \leq t_0$ ($H_0^{'} : \pi \leq \pi_0$) against the alternative $H_1 : T(\mathbf{X}) > t_0$ ($H_1^{'} : \pi > \pi_0$). Taking $t_0 = \frac{N}{2}$ ($\pi_0 = 0.5$), then rejecting H_0 ($H_0^{'}$) means that the candidate will receive more than 50 % of the votes and win the election in the first round. In order to test these hypotheses, we consider the Bayes test that basically consists of calculating the posterior probability of the null hypothesis and rejecting it when its posterior probability is smaller than a specified threshold [2].

The null hypotheses posterior probabilities, given the sampling total (which is a sufficient statistic) $\sum_{i=1}^{n} X_i = s$, $s \in \{0, 1, \ldots, n\}$, under the superpopulation and finite population approaches are, respectively, given by

$$P\left(H_0^{'} \middle| \sum_{i=1}^{n} X_i = s\right) = \int_0^{0.5} \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+s)\Gamma(\beta+n-s)} \pi^{\alpha+s-1}(1-\pi)^{\beta+n-s-1} d\pi$$

and

$$P\left(H_0 \middle| \sum_{i=1}^{n} X_i = s\right) = \frac{1}{c} \sum_{t=s}^{\lfloor \frac{N}{2} \rfloor} \frac{\binom{N-n}{t-s}}{\binom{N}{t}} p_t \,,$$

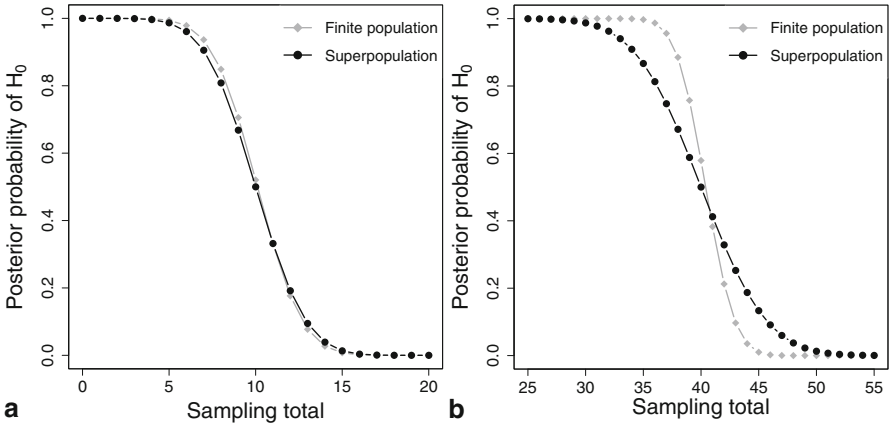

Fig. 16.1 Values of the probabilities of $H_0 : T(\mathbf{X}) \leq \frac{N}{2}$ ($H_0' : \pi \leq 0.5$) as functions of the sampling total for a population with $N = 100$ units

where c is the normalizing constant, $c = \sum_{t=s}^{N} P(X_1{=}x_1, \ldots, X_n{=}x_n | T(\mathbf{X}){=}t)\, p_t$, and $\lfloor b \rfloor$ is the largest integer less than or equal to b.

Figure 16.1 shows the posterior probabilities of the null hypotheses as functions of the sampling total assuming uniform prior distributions (Beta(1,1) in the superpopulation model and Uniform$\{0, 1, \ldots, \mathrm{N}\}$ in the finite population one). Considering a population of size $N = 100$ and samples of sizes $n = 20$ and $n = 80$, we observe that there is almost no difference between such probabilities when the sample is small ($n = 20$), but when the sample size increases the differences are greater, exceeding 20 % in some cases (for instance, when $\sum_{i=1}^{n} X_i = 38$ the difference is 0.214).

16.2.2 Simple Hypotheses

The researcher may also be interested in predicting the exact amount (proportion) of votes that a candidate will receive. In this case, we can test $H_0^* : T(\mathbf{X}) = t_0$ ($H_0^{**} : \pi = \pi_0$) against $H_1^* : T(\mathbf{X}) \neq t_0$ ($H_1^{**} : \pi \neq \pi_0$). Note that, under the superpopulation model of Sect. 16.2.1, H_0^{**} is a sharp hypothesis and its probability will be zero for all possible sample outcomes. Thus, the Bayesian procedure based on the posterior probability of the null hypothesis (Sect 16.2.1) should now be abandoned. Hence, we consider the *full Bayesian significance test* (FBST) [11] which is a Bayesian procedure suitable for testing sharp null hypotheses without the need to assign positive prior probabilities to them (for details, see [6, 7, 12]). Under the finite population point of view, the probability of H_0^* may be positive, and we could consider the Bayes test based on the posterior probability of the null hypothesis. Nonetheless, we still use the FBST in this case for two reasons: (1) when N is large, the prior distribution is so widely spread over the set $\{0, 1, \ldots, N\}$ that any simple

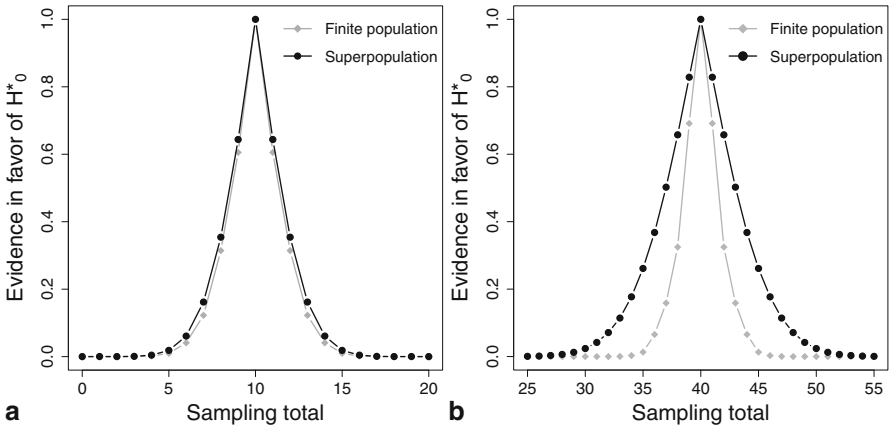

Fig. 16.2 Values of the evidence in favor of $H_0^* : T(\mathbf{X}) = \lfloor \frac{N}{2} \rfloor$ ($H_0^{**} : \pi = 0.5$) as functions of the sampling total for a population with $N = 100$ units

null hypothesis (even the mode of the posterior distribution) may be rejected because its posterior probability is too small (when compared with a threshold fixed beforehand) and (2) the comparison between evidence values of null hypotheses generated by the FBST under both approaches seems to be more suitable than the comparison between the FBST evidence and the posterior probability of a null hypothesis.

For $\pi_0 = 0.5$, the tangent region and evidence value [11] for the superpopulation model are given by

$$T_x^s = \{\pi \in (0,1) \ : \ f(\pi|s) > f(\pi_0|s)\} \text{ and}$$

$$ev(\{\pi_0\}; \mathbf{x}) = 1 - \int_{T_x^s} \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+s)\Gamma(\beta+n-s)} \pi^{\alpha+s-1}(1-\pi)^{\beta+n-s-1} d\pi,$$

whereas under the predictivistic model, for $t_0 = \lfloor \frac{N}{2} \rfloor$, they are given by

$$T_x^f = \left\{ k \in \{s, \ldots, N\} : P\left(T(\mathbf{X}) = k \,\middle|\, \sum_{i=1}^n X_i = s\right) > P\left(T(\mathbf{X}) = \left\lfloor \frac{N}{2} \right\rfloor \,\middle|\, \sum_{i=1}^n X_i = s\right) \right\}$$

and $\quad ev(\{\lfloor \frac{N}{2} \rfloor\}; \mathbf{x}) = 1 - \sum_{T_x^f} P\left(T(\mathbf{X}) = k \,\middle|\, \sum_{i=1}^n X_i = s\right),$

where $f(.|s)$ denotes the posterior Beta density with parameters $\alpha + s$ and $\beta + n - s$.

Assuming again uniform prior distributions, Fig. 16.2 shows the evidences in favor of the null hypotheses for all the possible values of the sampling total. As in the previous section, the greater the sampling fraction, the greater are the differences between the results under the finite population and the superpopulation models.

Simulations with different prior distributions have also been conducted, considering both one-sided and simple hypotheses, to evaluate the effect of the prior on

the results of the tests. To do this, under the superpopulation model, we have given different values to the hyperparameters α and β of the prior Beta distribution of π and, for the finite population model, we have considered the distribution of $\lfloor N\pi \rfloor$ as a prior for $T(\mathbf{X})$, so as to minimize (in a sense) the differences between the shapes of the priors. We have observed that the finite population method is less sensitive to prior specification than the superpopulation model. We have also performed simulations for other values of t_0 (or π_0) and have obtained similar results (curves are only moved to right (to left) as we increase (decrease) its value).

16.3 Multivariate Case

Now, consider the situation in which we aim to learn about the population proportions (totals) of votes of M candidates $a_1, \ldots, a_M$. For each elector, consider an M-dimensional vector $\mathbf{X}_j$ that takes the value $\mathbb{I}_k^{(M)} = (0, \ldots, 0, 1, 0, \ldots, 0)$ (the M-tuple composed of zeros except for the value 1 at the kth position) if the elector intends to vote for the candidate a_k, $k = 1, \ldots, M$. In this way, we want to infer about the population totals $\mathbf{T}(\mathbf{X}) = \sum_{j=1}^{N} \mathbf{X_j} = (T_1(\mathbf{X}), \ldots, T_M(\mathbf{X}))$, where $T_k(\mathbf{X})$ denotes the number of votes for candidate a_k, $k = 1, \ldots, M$.

Another way to think about this model is to imagine an urn containing N balls and that each ball is labeled with one of the different numbers $a_1, \ldots, a_M$. The interest lies in discovering how many balls of each type exist in the urn.

The following result shows the posterior distribution of $\mathbf{T}(\mathbf{X})$ under the predictivistic approach, given the observation $(\mathbf{x_1}, \ldots, \mathbf{x_n})$, $n < N$.

Result 15.2. *Let $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_N)$ be an exchangeable random vector taking values on $\left\{\mathbb{I}_1^{(M)}, \ldots, \mathbb{I}_M^{(M)}\right\}^N$, where $\mathbb{I}_k^{(M)}$ is an M-dimensional vector of zeros except for the value 1 at the k-th position, $k = 1, \ldots, M$, according to the probability measure $\mathbb{P}$ and consider $p_t = \mathbb{P}(\mathbf{T}(\mathbf{X}) = \mathbf{t})$, $\mathbf{t} \in \{(t_1, \ldots, t_M) \in \mathbb{N}^M : \sum_{i=1}^{M} t_i = N\}$, as the prior uncertainty about the vector $\mathbf{T}(\mathbf{X})$. Then, its posterior probability function given $(\mathbf{x_1}, \ldots, \mathbf{x_n})$, is:*

$$P\left(\mathbf{T}(\mathbf{X}) = \mathbf{t} \middle| \mathbf{X_1} = \mathbf{x_1}, \ldots, \mathbf{X_n} = \mathbf{x_n}\right) \propto$$

$$\frac{\binom{N-n}{t_1 - t_{1s} \ldots t_M - t_{Ms}}}{\binom{N}{t_1 \ldots t_M}} p_t \prod_{i=1}^{M} I_{\{t_{is}, \ldots, N - \sum_{j \neq i} t_{js}\}}(t_i)\,,$$

where $\mathbf{t} = (t_1, \ldots, t_M) \in \mathbb{N}^M$ with $\sum_{i=1}^{M} t_i = N$ and $\mathbf{t_s} = \sum_{i=1}^{n} \mathbf{x_i} = (t_{1s}, \ldots, t_{Ms}) \in \mathbb{N}^M$ with $\sum_{i=1}^{M} t_{is} = n$. (The proof of this result can be found in [8])

From a Bayesian superpopulation point of view, we model the vector of proportions $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_{M-1})$ where π_i represents the proportion of votes for candidate a_i, $i = 1, \ldots, M-1$, so that $\pi_i \geq 0$, $\pi_1 + \ldots + \pi_{M-1} \leq 1$ (we define $\pi_M = 1 - \sum_{i=1}^{M-1} \pi_i$). We suppose that the prior uncertainty on $\boldsymbol{\pi}$ can be

modeled by a Dirichlet distribution of order $M-1$ [2], that is, $\boldsymbol{\pi} \sim Dir(\boldsymbol{\alpha})$, with $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_M)$, $\alpha_i > 0, i = 1, \ldots, M$. We also assume that the (infinite) random vectors $(\boldsymbol{X}_i)_{i \geq 1}$, given $\boldsymbol{\pi}$, are conditionally independent $Multinomial(1, \boldsymbol{\pi})$. Thus, the posterior distribution of $\boldsymbol{\pi}$, given $\mathbf{X_1} = \mathbf{x_1}, \ldots, \mathbf{X_n} = \mathbf{x_n}$, is also Dirichlet, with updated parameter $\boldsymbol{\alpha} + \mathbf{t_s}$.

16.3.1 Testing the Possibility of a Second Round

In this section, we consider the application of the multivariate model, described in Sect. 16.3, to the scenario of majority elections with $M = 3$ candidates. In this simplified scenario, we are interested in the possibility of a second round in a majority election. This question may be answered by means of hypothesis testing, in which we must choose one of the hypotheses:

$$\begin{cases} H_0 : \bigcap_{i=1}^{3} \{T_i(\mathbf{X}) \leq \frac{N}{2}\} & (H_0^{'} : \bigcap_{i=1}^{3} \{\pi_i \leq 0.5\}) \\ H_1 : \bigcup_{i=1}^{3} \{T_i(\mathbf{X}) > \frac{N}{2}\} & (H_1^{'} : \bigcup_{i=1}^{3} \{\pi_i > 0.5\}) \end{cases}. \tag{16.1}$$

Hypothesis H_0 (or $H_0^{'}$) states that none of the candidates have more than 50 % of the votes so that there will be a runoff. On the other hand, the alternative hypothesis H_1 ($H_1^{'}$) states that (at least) one of the candidates gets more than half of the votes and, therefore, there will be no need to conduct the second round.

To test the hypotheses in (16.3.1), we use the Bayesian approach discussed in Sect. 16.2.1. Considering the superpopulation approach, we can compute the probability of $H_0^{'}$ using the marginal distributions of the vector $\boldsymbol{\pi} = (\pi_1, \pi_2)$ (recall $\pi_3 = 1 - \pi_1 - \pi_2$), as follows (note that $H_1^{'}$ is a union of disjoint sets):

$$P\left(H_0^{'} \middle| \mathbf{X_1} = \mathbf{x_1}, \ldots, \mathbf{X_n} = \mathbf{x_n}\right) =$$

$$1 - \sum_{i=1}^{3} \int_{0.5}^{1} \frac{\Gamma(\sum_{k=1}^{3} \{\alpha_k + t_{ks}\})}{\Gamma(\alpha_i + t_{is})\Gamma(\sum_{j \neq i} \{\alpha_j + t_{js}\})} \pi_i^{\alpha_i + t_{is} - 1} (1 - \pi_i)^{\sum_{j \neq i} \{\alpha_j + t_{js}\} - 1} d\pi_i.$$

Under the predictivistic approach for finite population, the posterior probability of the null hypothesis H_0 is:

$$P(H_0 | \mathbf{X_1} = \mathbf{x_1}, \ldots, \mathbf{X_n} = \mathbf{x_n}) = \sum_{\mathbf{t} \in D} P(\mathbf{T}(\mathbf{X}) = \mathbf{t} | \mathbf{X_1} = \mathbf{x_1}, \ldots, \mathbf{X_n} = \mathbf{x_n}),$$

where $D = \left\{(t_1, t_2, t_3) \in \mathbb{N}^3 : t_1 \leq \frac{N}{2},\ t_2 \leq \frac{N}{2},\ t_1 + t_2 \geq \frac{N}{2} \text{ and } t_1 + t_2 + t_3 = N\right\}$.

To compare the results obtained from the superpopulation and predictivistic models, we consider a population of size $N = 100$, samples with $n = 20$ and $n = 80$ units and uniform priors: In the finite population model, a uniform distribution on the set $\{(t_1, t_2, t_3) \in \mathbb{N}^3 : t_1 + t_2 + t_3 = N\}$ and, in the superpopulation model, we

consider $\boldsymbol{\pi} \sim Dir(1, 1, 1)$ as the prior distribution. Calculating the probabilities of the null hypotheses, we observe that when the sampling fraction is small ($n = 20$), the results produced by both methods are very similar, although some differences may still occur, for example, if the sampling vector of totals is $\mathbf{t_s} = (9, 2, 9)$, then the probability of H_0 calculated under the predictivistic approach for finite population is 0.53, while, under the superpopulation model, this value drops to 0.48. Due to the exchangeability and uniformity of prior distributions, the same values occur when the sampling totals are $\mathbf{t_s} = (2, 9, 9)$ or $\mathbf{t_s} = (9, 9, 2)$.

On the other hand, when the sample size increases ($n = 80$), the differences between the results obtained from these methods stand out, reaching more than 36 %. For example, when $\mathbf{t_s} = (37, 5, 38)$ (or any permutation of this vector), the probability of the null hypothesis in the superpopulation model is 0.49 and it increases to 0.85 in the finite population model. If $\mathbf{t_s} = (38, 4, 38)$, the probabilities are 0.79 for finite population and 0.42 under the superpopulation model.

16.4 Final Remarks

In this chapter, we explore the differences between the results of hypothesis testing obtained from the usual Bayesian superpopulation model and the corresponding results under the predictivistic approach considering a simple scenario of majority elections.

It is well-known that such methods yield similar results in cases where the sampling fraction is small. On the other hand, little emphasis has been given in the literature to the situations of small populations or of large sampling fractions. The latter usually occurs, for instance, in electoral processes when statistical forecasts on the final result of the election must be conducted on the basis of the partial counting of votes at advanced stage. In such circumstances, the finite population approach should be preferred over the superpopulation model.

In short, we consider, in a simplified scenario of elections, tests for univariate and multivariate operational parameters. In the univariate case, the FBST and the procedure based on the posterior probability of the null hypothesis are applied, respectively, to simple and one-sided hypotheses, under both the finite population and superpopulation models. In the multivariate example, procedures are developed to test the possibility of a second round in elections.

For small sampling fractions, the approaches yield similar results, while for large sampling fractions, the differences between the posterior probabilities (evidences) of the hypotheses of interest are prominent, exceeding 20 % in some cases. When the sample size is close to the population size, this occurs because, under the predictivistic model, the amount of values of $T(\mathbf{X})$ with positive posterior probability is much lower than prior to sampling, while in the superpopulation model, the supports of the prior and posterior distributions are the same, although there is a considerable decrease in the variance. For the same reason, the finite population model is less sensitive to changes in the prior distributions. In addition, the fluctuations of the null hypothesis

posterior probabilities (and evidences) due to variations of the sufficient statistic (sampling totals) are smoother and less sensitive in superpopulation models.

The comparative study of the aforementioned approaches for other operational parameters, such as the population maximum and minimum, is the goal of future inquiries. Also open to investigation is the derivation of theoretical bounds for the differences between probabilities of hypotheses of interest determined from these approaches.

References

1. De Finetti, B.: La prévision: ses lois logiques, ses sources subjectives. Annales de l'institut Henri Poincaré, **7**, 1–68 (1937)
2. DeGroot, M.H.: Optimal Statistical Decisions. McGraw-Hill, New York (1970)
3. Fossaluza, V.: Testes de hipóteses em eleições majoritárias. Master's thesis, Instituto de Matemática e Estatística, Universidade de São Paulo (2008)
4. Iglesias, P., Loschi, R., Pereira, C., Wechsler, S.: A note on extendibility and predictivistic inference in finite populations. Braz. J. Probab. Stat. **23**(2), 216–226 (2009)
5. Iglesias, P.L.: Formas finitas do teorema de de finetti: a visão preditivista da inferência estatística em populações finitas. Ph.D. thesis, Instituto de Matemática e Estatística, Universidade de São Paulo (1993)
6. Madruga, M., Esteves, L., Wechsler, S.: On the Bayesianity of Pereira-Stern tests. Test **10**(2), 291–299 (2001)
7. Madruga, M., Pereira, C.A.B.: Power of fbst: standard examples. Instituto Interamericano de Estadística, Estadística 57, 1–9 (2005)
8. Melo, B.A.R.: Um estudo comparativo entre abordagens bayesianas à testes de hipóteses. Master's thesis, Instituto de Matemática e Estatística, Universidade de São Paulo (2013)
9. Mendel, M.B.: Operational parameters in Bayesian models. Test **3**, 195–206 (1994)
10. Montaquila, J.M., Kalton, G.: Sampling from finite populations. In: International Encyclopedia of Statistical Science, pp. 1277–1281. Springer (2011)
11. Pereira, C.A.B., Stern, J.M.: Evidence and credibility: full Bayesian significance test for precise hypotheses. Entropy J. **1**, 104–115 (1999)
12. Pereira, C.A.B., Stern, J.M., Wechsler, S.: Can a significance test be genuinely Bayesian? Bayesian Anal. **3**, 79–100 (2008)

Chapter 17
Bayesian Ridge-Regularized Covariance Selection with Community Behavior in Latent Gaussian Graphical Models

Lijun Peng and Luis E. Carvalho

Abstract Gaussian graphical models have been extensively used to model conditional independence via the concentration matrix of a random vector. They are particularly relevant to incorporate structure when the length of the vector is large and naive methods lead to unstable estimation of the concentration matrix. In covariance selection, we have a latent network among vector components such that two components are not connected if they are conditionally independent, that is, if their corresponding entry in the concentration matrix is zero. In this work, we expect that, in addition, vector components show a block dependency structure that represents community behavior in the context of biological and social applications, that is, connections between nodes from different blocks are sparse while connections within nodes of the same block are dense. Thus, to identify the latent network and detect communities, we propose a Bayesian approach with a hierarchical prior in two levels: a spike-and-slab prior on each off-diagonal entry of the concentration matrix for variable selection; and a degree-corrected stochastic blockmodel (SBM) to capture the community behavior. To conduct inference, we develop an efficient routine based on ridge regularization and maximum a posteriori (MAP) estimation. Finally, we demonstrate the proposed approach in a meta-genomic dataset of complex microbial biofilms from dental plaque and show how bacterial communities can be identified.

17.1 Introduction

In a network, objects are represented by *nodes* and their interactions by *edges*. A cluster of nodes with dense connections within nodes but with sparse interactions with nodes in other clusters is thought of as *community*. In recent years, there has been tremendous interest in applying network analysis in a number of fields. In biology, for

L. Peng (✉) · L. E. Carvalho
Department of Mathematics and Statistics, Boston University, 111 Cummington Mall, Boston, MA 02215, USA
e-mail: ljpeng@bu.edu

L. E. Carvalho
e-mail: lecarval@math.bu.edu

A. Polpo et al. (eds.), *Interdisciplinary Bayesian Statistics,*
Springer Proceedings in Mathematics & Statistics 118, DOI 10.1007/978-3-319-12454-4_17

instance, network analysis has been employed to describe interdependency among genes and gene products as well as discovering modules that function alike in an attempt to gain a better understanding of a working cell [1, 3, 7, 18]. In other words, it is of great significance to identify and detect the community structure underlying biological networks such as protein interaction, association, and metabolic networks.

Gaussian graphical models [4] have been widely used to describe conditional independence between components of a random vector [20]. In a Gaussian graphical model, we associate to each component X_i of a Gaussian random vector $\vec{X}$ a node in a graph $G = (V, E)$, and two nodes i and j are *not* connected in G if and only if their corresponding components are conditionally independent given all the other components, that is, if the partial correlation between X_i and X_j is zero. In other words, $(i, i) \notin E$ if and only if $\rho_{X_i, X_j \mid X_{V \setminus \{i,j\}}} = 0$. Equivalently, if C is the *concentration* matrix of X, that is, the inverse variance of X, then, since the partial correlation $\rho_{X_i, X_j \mid X_{V \setminus \{i,j\}}} \propto C_{ij}$, $(i, j) \notin E$ if and only if $C_{ij} = 0$. Thus, inferring conditional independence is equivalent to estimating null entries in a concentration matrix [14].

There are several difficulties in identifying the conditional independence of Gaussian variables. The first challenge is the "large p, small n" regimen that stems from inferring a large concentration matrix based on relatively few observations. Under this regimen, the sample covariance matrix is singular and thus cannot be inverted to directly compute the concentration matrix. A commonly used alternative is to include some form of regularization, such as graph lasso [4, 8, 16]. The second challenge is developing an effective procedure to identify the corresponding network based on the estimated concentration matrix. Traditional variable selection approaches such as stepwise regression and those especially adapted to Gaussian graphical models [6] offer a way to check the significance of estimated partial correlations. However, determining edges of networks and inferring concentration matrices are performed separately. A potential drawback of these methods is that once a "bad" model is selected, the estimation of concentration matrices is unreliable as a result. To remedy this problem, efficient approaches to jointly perform model selection and graph estimation have been proposed [22, 23]. However, none of the methods mentioned above has taken community structure into account when estimating the conditional independence.

Among the many community detection approaches, we focus on *stochastic blockmodels* (SBM) where the probability of an association between two nodes depends on the communities to which they belong [5, 12]. We employ a hierarchical Bayesian SBM that regards probabilities of association as random and group membership as latent variables which allows for *degree-correction*, that is, models where the degree distribution of nodes within each group can be heterogeneous [17].

Our main contribution in this chapter is to jointly estimate concentration matrices and latent networks while taking community structure into account. To this end, we propose a Bayesian approach with a hierarchical prior with two levels in Sect. 17.2:

1. We develop a Bayesian ridge-regularized covariance selection that specifies a spike-and-slab prior on each off-diagonal entry of the concentration matrix. With this approach, we are able to obtain a positive-definite estimate of the concentration matrix and determine the underlying network simultaneously. We relate covariance selection and variable selection for Gaussian graphical models through an efficient algorithm in Sect. 17.3.
2. We offer a Bayesian approach for community detection that explicitly characterizes community behavior and a *maximum a posteriori* (MAP) estimator to efficiently conduct inference in Sect. 17.3.

Results from a simulation study comparing our ridge-regularized covariance selection to other methods are reported in Sect. 17.4. We show that our proposed method is efficient and as reliable as other commonly used methods. A real-world meta-genomic dataset of complex microbial biofilms is used to demonstrate the covariance selection as well as community detection in Sect. 17.5. Finally, we offer some concluding remarks and directions for future work in Sect. 17.6.

17.2 Model Framework

We develop a hierarchical model to (i) perform covariance selection on a latent network of associations between individuals and (ii) identify the set of communities to which these individuals belong. We start by assuming that the data $\vec{X} = (\vec{X}_1, \ldots, \vec{X}_n)$ for each sample follows

$$\vec{X}_i \mid \vec{\mu}, \vec{C} \overset{\text{iid}}{\sim} \mathsf{N}(\vec{\mu}, \vec{C}^{-1}), \quad i = 1, \ldots, n, \tag{17.1}$$

where each $\vec{X}_i$ is p-dimensional. The mean μ can have more structure, as we will see in Sect. 17.5. We set an non-informative prior on μ, $\mathbb{P}(\mu) \propto 1$.

The likelihood in (17.1) implicitly defines a Gaussian graphical model on a undirected graph with p nodes and adjacency matrix $\vec{A}$. Recall that in a Gaussian graphical model, node i and node j are conditionally independent ($C_{ij} = 0$) if and only if there is no edge between them ($A_{ij} = 0$). To select which off-diagonal entries in $\vec{C}$ are zero we adopt a spike-and-slab prior [9, 11] with $\vec{A}$ as indicators:

$$C_{ij} \mid A_{ij} \overset{\text{ind}}{\sim} \mathsf{N}(0, \rho^2 A_{ij} + \rho^2 \nu_0 (1 - A_{ij})), \quad i, j = 1, \ldots, p, i < j, \tag{17.2}$$

where ρ^2 is chosen to be large (the "slab") while ν_0 is small (the "spike.") For the diagonal entries we set

$$C_{ii} \mid \lambda \overset{\text{ind}}{\sim} \mathsf{Exp}(\lambda/2), \quad i = 1, \ldots, p, \tag{17.3}$$

for computational convenience. In addition, we settle on a non-informative prior for λ, $\mathbb{P}(\lambda) \propto 1$.

Finally, to model the adjacency matrix $\vec{A}$, we adopt a degree-corrected SBM which specifies that the probability of an edge between node i and j depends on their labels (σ_i, σ_j) and their expected degrees, and that σ follows a product multinomial distribution [17]:

$$A_{ij}|\vec{\sigma}, \vec{\gamma}, \vec{\eta} \overset{\text{ind}}{\sim} \mathsf{Bern}(\text{logit}^{-1}(\gamma_{\sigma_i\sigma_j} + \eta_i + \eta_j)), \quad i, j = 1, \ldots, p, i < j, \tag{17.4}$$
$$\sigma_i \overset{\text{ind}}{\sim} \mathsf{MN}(1; \vec{\pi}), \quad i = 1, \ldots, p.$$

Hyper-parameters $\vec{\gamma}$ capture within and between community probabilities of association (in logit scale) and node intercepts $\vec{\eta} = (\eta_1, \ldots, \eta_p)$ capture the expected degrees of the nodes. A more realistic model is attained by further setting a hyper-prior distribution on γ and η,

$$(\gamma, \eta) \sim I(\gamma \leq 0) \cdot \mathsf{N}\big(0, \tau^2 I\big), \tag{17.5}$$

where τ^2 controls how informative the prior is. The constraint $\gamma \leq 0$ in this SBM is essential to community detection since we should expect as many as or fewer edges between communities than within communities on average, and thus that the log-odds of between and within probabilities is non-positive.

To summarize, in the likelihood, we adopt a Gaussian graphical model; in the next level, we select the covariance structure in $\vec{C}^{-1}$ with a spike-and-slab prior; and finally, we capture community behavior in the components of $\vec{X}$ via a SBM on $\vec{A}$.

17.3 Inference

To develop the MAP estimator for $\vec{C}$, $\vec{A}$, σ, γ, and η, we follow a cyclic gradient descent approach where each parameter is obtained by optimizing

$$[\vec{C}, \vec{A} \,|\, \sigma, \gamma, \eta, \mu, \lambda, \vec{X}], \qquad [\sigma \,|\, \gamma, \eta, \vec{C}, \vec{A}, \mu, \lambda, \vec{X}], [\gamma, \eta \,|\, \sigma, \vec{C}, \vec{A}, \mu, \lambda, \vec{X}],$$
$$[\mu \,|\, \sigma, \vec{C}, \vec{A}, \sigma, \gamma, \eta, \lambda, \vec{X}], \quad [\lambda \,|\, \sigma, \vec{C}, \vec{A}, \sigma, \gamma, \eta, \mu, \vec{X}]$$

in turn. While we have a step using μ, in general we have $\hat{\mu} = \sum_{i=1}^{n} \vec{X}_i/n$ and so we often consider $\vec{X}_i \,|\, \vec{C} \overset{\text{iid}}{\sim} \mathsf{N}(\vec{0}, \vec{C}^{-1})$ by precentering $\vec{X}$. Similarly, the MAP estimator for λ is straightforward: $\hat{\lambda} = 2/\sum_{i=1}^{p} C_{ii}$.

Now, we want to find a concentration matrix $\vec{C}$ and latent network $\vec{A}$ that maximize $\log \mathbb{P}(\vec{C}, \vec{A}|\sigma, \gamma, \eta, \vec{X})$, or equivalently,

$$\log \mathbb{P}(\vec{C}, \vec{A}, \vec{X}|\sigma, \gamma, \eta) = \frac{n}{2} \log|\vec{C}| - \frac{1}{2} \sum_{i=1}^{n} (\vec{X}_i - \mu)^\top \vec{C} (\vec{X}_i - \mu)$$
$$- \frac{1}{2\rho^2} \sum_{1 \leq i < j \leq p} \frac{C_{ij}^2}{A_{ij} + \nu_0(1 - A_{ij})} - \frac{\lambda}{2} \sum_{i=1}^{p} C_{ii} + \sum_{1 \leq i < j \leq p} A_{ij}(\gamma_{\sigma_i\sigma_j} + \eta_i + \eta_j). \tag{17.6}$$

To find the conditional MAP estimator for $\vec{C}$ and $\vec{A}$, we focus on each of their rows (or columns) at a time. For the ith row and column, we consider the log-likelihood as a function of $\vec{C}_{i,\cdot}$, that is,

$$\log \mathbb{P}(\vec{C}_{i,\cdot}, \vec{X}, \vec{A}) = \frac{n}{2} \log |C_{ii} - \vec{C}_{i,-i}\vec{C}_{-i,-i}^{-1}\vec{C}_{-i,i}| - \frac{1}{2}(S_{ii}C_{ii} + 2\vec{S}_{i,-i}\vec{C}_{-i,i}) - \frac{1}{2\rho^2}\sum_{j\neq i}\frac{C_{ij}^2}{A_{ij} + \nu_0(1-A_{ij})} - \frac{\lambda}{2}C_{ii},$$

up to terms that do not involve $\vec{C}_{i,\cdot}$. Here $\vec{S} = \sum_{i=1}^{n}(\vec{X}_i - \hat{\mu})(\vec{X}_i - \hat{\mu})^\top$ is a sufficient statistic. Then, if $\vec{V}_i = \mathrm{Diag}_{j\neq i}\{1/[\rho^2 A_{ij} + \rho^2\nu_0(1-A_{ij})]\}$, $\vec{C}_{i,-i}$ is the ith row with ith column removed and $\vec{C}_{-i,-i}^{-1}$ is the sub-matrix of $\vec{C}^{-1}$ with ith row and column removed, the ridge-regularized estimator for $\vec{C}_{i,\cdot}$ is given by

$$\begin{aligned}\hat{\vec{C}}_{i,-i} &= -[(S_{ii} + \lambda)\vec{C}_{-i,-i}^{-1} + \vec{V}_i]^{-1}\vec{S}_{i,-i},\\ \hat{\vec{C}}_{i,i} &= \frac{n}{S_{ii}+\lambda} + \hat{\vec{C}}_{i,-i}\vec{C}_{-i,-i}^{-1}\hat{\vec{C}}_{i,-i}^{\top}.\end{aligned} \tag{17.7}$$

We note that $\vec{C}$ is kept positive definite along the whole procedure. Since $\vec{C} \succ 0$, $\vec{C}_{-i,-i} \succ 0$ for any i; after the update $\vec{C}_{i,\cdot}$, $\vec{C}$ is still positive definite since

$$C_{ii} - \vec{C}_{i,-i}\vec{C}_{-i,-i}^{-1}\vec{C}_{-i,i} = \frac{n}{S_{ii}+\lambda} > 0,$$

as noted in [22]. We can obtain the estimate by solving (17.7) for each row and iterating until convergence. Thus, if the initial value of the concentration matrix is symmetric and positive definite, then the estimate based on (17.7) is also symmetric and positive definite throughout the iterative procedure.

This series of updates is conditional on $\vec{A}$, as seen in Eq. (17.6). However, we further propose an approach where the adjacency matrix is estimated along with the concentration matrix in the covariance selection procedure, that is, we update $\vec{C}$ and $\vec{A}$ jointly. Moreover, to avoid the risk of getting stuck if we update the whole row $\vec{A}_{i,\cdot}$ each time, we propose to update at most one entry in $\vec{A}$ at each iteration. That is, we move to $\vec{A}_{i,\cdot}^{(t+1)}$ where the candidate move set for $\vec{A}_{i\cdot}$ is $\{\vec{A}_{i,\cdot} : H(\vec{A}_{i,\cdot}, \vec{A}_{i,\cdot}^{(t)}) \leq 1\}$ and $H(\cdot,\cdot)$ is the Hamming distance. In practice, we adopt the SWEEP operator [10] and Cholesky up/down-dates to make the iterative algorithm more efficient.

Once we are given the adjacency matrix $\vec{A}$ representing relationships between "actors" i and j in a network, we adopt a Bayesian degree-corrected SBM given in Eq. (17.4) to detect the community structure in the network [17], that is, to find a conditional MAP estimator for $[\sigma|\gamma, \eta, \vec{C}, \vec{A}, \vec{X}]$ and $[\gamma, \eta|\sigma, \vec{C}, \vec{A}, \vec{X}]$. First, we take σ_i to be the mode of $\sigma_i|\sigma_{[-i]}, \beta, \vec{A}$. Next, a regularized iterative reweighted least-squares (IRLS) is carried out. IRLS is usual when fitting logistic regression models [15]. To guarantee that the community constraints $\gamma \leq 0$ are met, we use an active-set method [21]. More details can be found in [17].

To summarize, the Bayesian ridge-regularized graph estimate is obtained by iterating until convergence the following steps:

Algorithm 1 Bayesian ridge-regularized covariance selection

Set initial $\mathbf{C}, \mathbf{A}$; obtain initial λ, σ, η and γ based on $\mathbf{C}, \mathbf{A}$.
repeat
  for $i = 1, \ldots, p$ **do**
    Set $\text{lhood}_{\max} = -\infty$, $\mathbf{W} = \text{SWEEP}(\mathbf{C}^{-1}, i)$
    for $j \neq i$ **do**
      $\left\{ \begin{array}{lll} A_{ij} = 0 & \Longrightarrow & \widehat{\mathbf{C}}_{i,\cdot}^{(0)}; \text{ compute lhood}_0 \\ A_{ij} = 1 & \Longrightarrow & \widehat{\mathbf{C}}_{i,\cdot}^{(1)}; \text{ compute lhood}_1 \end{array} \right\}$, lhood=$\max_{k \in \{0,1\}} \text{lhood}_k$
      if $\text{lhood} > \text{lhood}_{\max}$
        $\text{lhood}_{\max} = \text{lhood}, j^\star = j, k^\star = \arg\max_{k \in \{0,1\}} \text{lhood}_k, \widehat{\mathbf{C}}_{i,\cdot} = \widehat{\mathbf{C}}_{i,\cdot}^{(k^\star)}$
      end if
    end for
    Update $\mathbf{A}_{ij^\star} = \mathbf{A}_{j^\star i} \leftarrow k^\star$
    Update $\mathbf{W}_{-i,i} \leftarrow \mathbf{W}_{-i,-i} \widehat{\mathbf{C}}_{-i,i}, \mathbf{W}_{i,-i} \leftarrow -\mathbf{W}_{-i,i}^\top, \mathbf{W}_{i,i} = \frac{n}{S_{ii} + \lambda}$
    Update $\mathbf{C}^{-1} \leftarrow \text{SWEEP}(\mathbf{W}, i)$
  end for
  Update λ
  Update σ, η and γ from the community detection procedure in Section 1.3
until the change in the log-likelihood is within certain tolerance

17.4 Experimental Results

In this section, we evaluate the performance of our proposed Bayesian ridge-regularized estimator in identifying latent networks. For comparison, we also estimate the concentration using sample estimates and a lasso-regularized estimator [22].

Our simulation study generates networks from a popular benchmark suite due to Fortunato [13] that accounts for heterogeneities in node degree distributions and community sizes. The model used in the simulation considers the following parameters: both degree distribution and the community sizes are assumed to follow power law distributions with exponents $a = 2$ and $b = 1$, respectively; each network consists of $p = 50$ nodes and has average degree $\langle k \rangle = 10$. Mixing parameter μ captures the proportion of between-community edges. We highlight two community behaviors: gregarious, with $\mu = 0.1$, or non-assortative, with $\mu = 0.4$.

We further generated concentration matrices based on the networks as ground truth according to Eq. (17.2) with fixed $\rho^2 = 100$ and $v_0 = 10^{-6}$ for simplicity. The value $\rho^2 = 100$ is large enough to distinguish the differences in the concentration matrix when edges in the latent network are present or absent. The data $\vec{X} = \{\vec{X}_1, \ldots, \vec{X}_n\}$ for $n = (10, 25, 50, 100, 200)$ was generated as in Eq. (17.1). We estimated concentration

matrices based on $\vec{X}$ by sample concentration, our approach with $\vec{A}$ known as well as unknown (latent), and Lasso estimates with different tuning parameters ranging from 0.001 to 10. The comparison in terms of the log relative Frobenius norm of estimated concentration matrices, $\log(\|\widehat{\vec{C}} - \vec{C}\|_F / \|\vec{C}\|_F)$, is shown in Fig. 17.1. Our proposed approach outperforms the sample and Lasso estimates in terms of the log relative Frobenius norm. In addition, the error we made in estimating latent networks is mainly due to false negatives (failing to detect an edge when there is one), especially when we have fewer observations.

17.5 Case Study

Periodontitis is the inflammation of the tissues that surround the teeth, and is caused by specific bacteria. These bacteria often explore symbiotic relations, and are thus expected to be found in communities. In this case study, we take a dataset that measures 16S ribonucleosomal expression levels using the Human Oral Microbe Identification Microarray (HOMIM) for 276 bacteria and contrasts 90 sites in healthy individuals to 514 sites in patients with varying degrees of periodontitis [7]. We assume, as before, that individual samples are independent, but now we exploit a decomposable mean model where the mean response for bacteria i and sample j is given by:

$$\mu_{ijc} = \theta_{ic} + \phi_j, \quad i = 1, \ldots, p, \quad j = 1, \ldots, n, \tag{17.8}$$

where c is the condition, either healthy or diseased. Parameters θ_{ic} capture the expression effect of each bacteria per condition, while parameters ϕ_j represent the baseline expression level per sample and are considered nuisance.

After running our proposed procedure for $K = 2, \ldots, 10$ communities, we select $K = 3$ based on the Bayesian information criterion (BIC). The two first panels in Fig. 17.2 depict the inferred networks and communities. As can be seen, the "diseased" bacterial community is more connected and has a stronger community effect. To compare the joint effect of expression via θ and connectivity, we compute alpha-centralities [2] using $\hat{\theta}_{\cdot c}$ as weights. The rightmost panel in Fig. 17.2 contrasts alpha-centrality between the two conditions; for comparison, we mark bacterial species according to [19] complexes. Interestingly, bacteria from the red and orange complexes—usually associated to more severe forms of periodontitis—tend to have higher alpha-centralities in the diseased sample group relative to the healthy group.

17.6 Discussion

In this chapter, we developed a Bayesian ridge-regularized covariance selection model that incorporates community behavior through a latent network. This class of models has many practical applications in social sciences and systems biology. Good

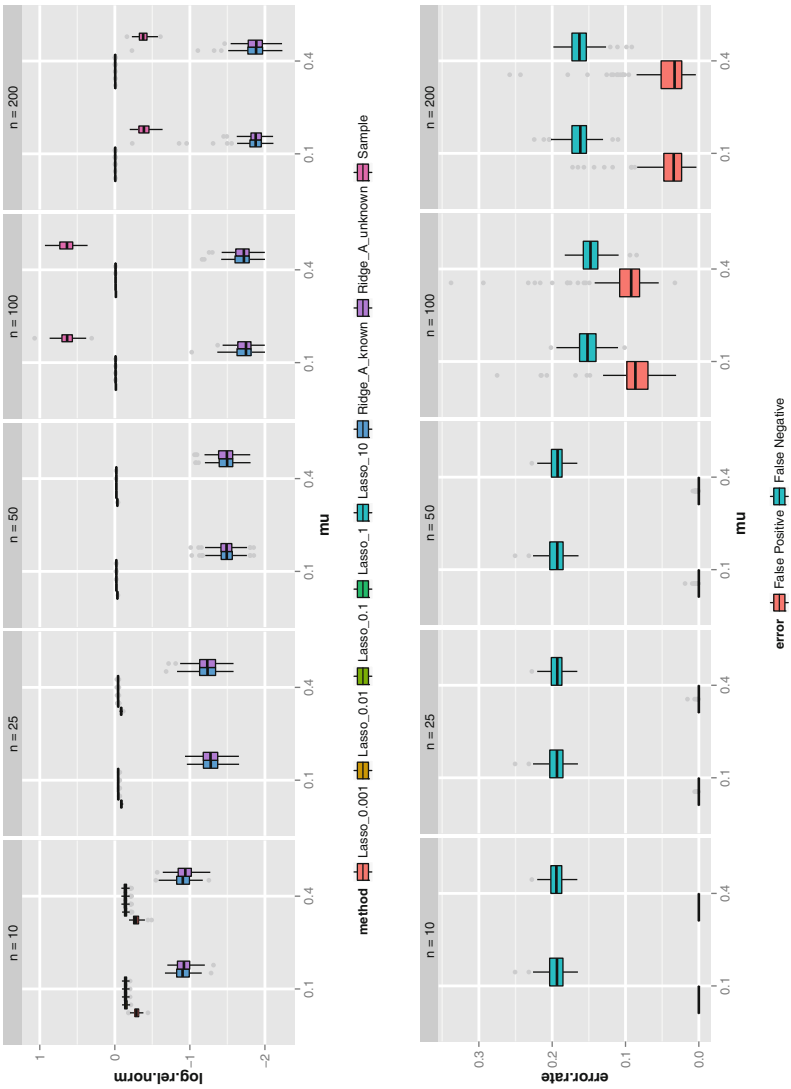

Fig. 17.1 (*Top*) The log relative Frobenius norm of estimated concentration matrices under different approaches. The sample estimates when $n < p$ have relatively large norms are not shown to maintain a short scale. (*Bottom*) The false positive and negative rates of estimated adjacency matrices under our proposed model

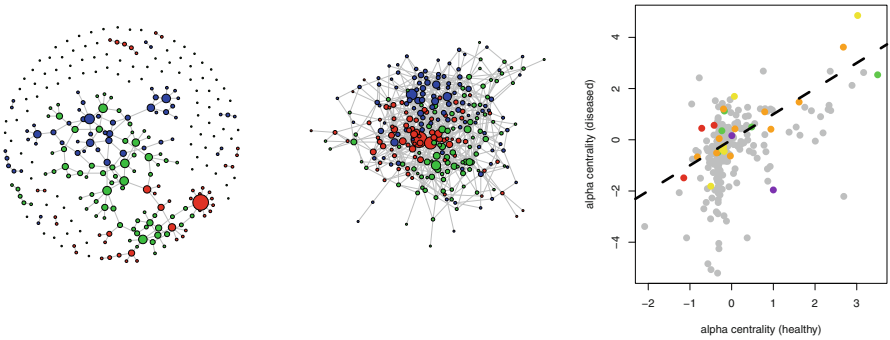

Fig. 17.2 Inferred networks for healthy (*left*) and diseased (*right*) samples. *Colors mark* inferred communities. In the *right*, alpha-centrality with $\alpha = 0.5$ with θ estimates as exterior weights; *colors mark* Socransky complex classification [19]

results based on our simulation study indicate that the proposed approach is a serious contender for covariance selection when compared to Lasso-based estimators. Moreover, as the case study shows, our estimator reliably captures biological assortativity in bacterial communities, and is able to classify bacteria with respect to their different responses in expression and connectivity under two scenarios. Moreover, since most of the bacteria in dental biofilms are not cultivable, the proposed model gives insight into which partnerships are needed for these bacteria under different conditions. For future work, we plan to extend the proposed model in two main ways: to accommodate different types of data and add more flexibility, we intend to adopt other generalized linear model families in the latent network layer of the model, that is, we also want to consider valued and weighted networks; many applications have time series data, and so we plan to develop dynamic models.

Acknowledgments L. Peng and L. E. Carvalho were partially supported by NSF grant DMS-1107067.

References

1. Andrade, R.F., Rocha-Neto, I.C., Santos, L.B., de Santana, C.N., Diniz, M.V., Lobão, T.P., Goés-Neto, A., Pinho, S.T., El-Hani, C.N.: Detecting network communities: an application to phylogenetic analysis. PLoS Comput. Biol. **7**(5), e1001131 (2011)
2. Bonacich, P., Lloyd, P.: Eigenvector-like measures of centrality for asymmetric relations. Soc. Netw. **23**(3), 191–201 (2001)
3. de Silva, E., Stumpf, M.P.: Complex networks and simple models in biology. J. R. Soc. Interface **2**(5), 419–430 (2005)
4. Dempster, A.P. Covariance selection. Biometrics **28**, 157–175 (1972)
5. Doreian, P., Batagelj, V., Ferligoj, A.. Generalized Blockmodeling. Cambridge University Press, Cambridge (2005)

6. Drton, M., Perlman, M.D.: Model selection for Gaussian concentration graphs. Biometrika **91**(3), 591–602 (2004)
7. Duran-Pinedo, A.E., Paster, B., Teles, R., Frias-Lopez, J.: Correlation network analysis applied to complex biofilm communities. PloS ONE **6**(12), e28438 (2011)
8. Friedman, J., Hastie, T., Tibshirani, R.: Sparse inverse covariance estimation with the graphical lasso. Biostatistics **9**(3), 432–441 (2008)
9. George, E.I., McCulloch, R.E.: Variable selection via Gibbs sampling. J. Am. Stat. Assoc. **88**(423), 881–889 (1993)
10. Goodnight, J.H.: A tutorial on the SWEEP operator. Am. Stat. **33**(3), 149–158 (1979)
11. Ishwaran, H., Rao, J.S.: Spike and slab variable selection: frequentist and Bayesian strategies. Ann. Stat., **33**, 730–773 (2005)
12. Karrer, B., Newman, M.E.: Stochastic blockmodels and community structure in networks. Phys. Rev. E **83**(1), 016107 (2011)
13. Lancichinetti, A., Fortunato, S., Radicchi, F.: Benchmark graphs for testing community detection algorithms. Phys. Rev. E **78**(4), 046110 (2008)
14. Lauritzen, S.L.: Graphical Models. Oxford University Press, Oxford (1996)
15. McCullagh, P., Nelder, J.A.: Generalized Linear Models. Chapman and Hall, London (1983/1989)
16. Meinshausen, N., Bühlmann, P. High-dimensional graphs and variable selection with the lasso. Ann.Stat., **34**, 1436–1462 (2006)
17. Peng, L., Carvalho, L. Bayesian degree-corrected stochastic block models for community detection. arXiv:1309.4796v1 (2013)
18. Scutari, M., Strimmer, K.: Introduction to graphical modelling. arXiv:1005.1036 (2010)
19. Socransky, S., Haffajee, A., Cugini, M., Smith, C., Kent, R.: Microbial complexes in subgingival plaque. J. Clin. Periodontol. **25**(2), 134–144 (1998)
20. Whittaker, J.: Graphical Models in Applied Multivariate Statistics. Wiley, Chichester (2009)
21. Wright, S., Nocedal, J.: Numerical Optimization, vol. 2. Springer, New York (1999)
22. Yuan, M.: Efficient computation of ℓ_1 regularized estimates in Gaussian graphical models. J. Comput. Graph. Stat. **17**(4), 809–826 (2008)
23. Yuan, M., Lin, Y.: Model selection and estimation in the Gaussian graphical model. Biometrika **94**(1), 19–35 (2007)

Chapter 18
Bayesian Inference of Deterministic Population Growth Models

Luiz Max Carvalho, Claudio J. Struchiner and Leonardo S. Bastos

Abstract Deterministic mathematical models play an important role in our understanding of population growth dynamics. In particular, the effect of temperature on the growth of disease-carrying insects is of great interest. In this chapter we propose a modified Verhulst—logistic growth—equation with temperature-dependent parameters. Namely, the growth rate r and the carrying capacity K are given by thermodynamic functions of temperature T, $r(T)$ and $K(T)$. Our main concern is with the problem of learning about unknown parameters of these deterministic functions from observations of population time series $P(t,T)$. We propose a strategy to estimate the parameters of $r(T)$ and $K(T)$ by treating the model output $P(t,T)$ as a realization of a Gaussian process (GP) with fixed variance and mean function given by the analytic solution to the modified Verhulst equation. We use Hamiltonian Monte Carlo (HMC), implemented using the recently developed **rstan** package of the R statistical computing environment, to approximate the posterior distribution of the parameters of interest. In order to evaluate the performance of our algorithm, we perform a Monte Carlo study on a simulated example, calculating bias and nominal coverage of credibility intervals. We then proceed to apply this approach to laboratory data on the temperature-dependent growth of a Chagas disease arthropod vector, *Rhodnius prolixus*. Analysis of this data shows that the growth rate for the insect population under study achieves its maximum around 26 °C and the carrying capacity is maximum around 25 °C, suggesting that *R. prolixus* populations may thrive even in non-tropical climates.

L. M. Carvalho (✉) · C. J. Struchiner · L. S. Bastos
Program for Scientific Computing (PROCC), Oswaldo Cruz Foundation,
Rio de Janeiro, RJ, Brazil
e-mail: lmax.procc@gmail.com

C. J. Struchiner
e-mail: stru@fiocruz.br

L. S. Bastos
e-mail: lsbastos@fiocruz.br

A. Polpo et al. (eds.), *Interdisciplinary Bayesian Statistics,*
Springer Proceedings in Mathematics & Statistics 118, DOI 10.1007/978-3-319-12454-4_18

18.1 Introduction

Deterministic models of population growth have played a major role in our understanding of biological populations [1, 2, 4, 7]. These mechanistic models allow us to draw a pittoresque picture of the real world and capture the main features of the system(s) under study.

In many applications it is of interest to estimate parameters of these models using observed (output) data. This chapter proposes a Bayesian approach to this problem in the context of models for insect population growth. In what follows, we introduce the necessary notation and concepts to be used hereafter.

Consider a deterministic model $M(\cdot)$. Let $y \in \mathcal{Y} \subset \mathbb{R}^n$ be the set of model inputs and $x \in \mathcal{X} \subset \mathbb{R}^p$ be the model outputs. The deterministic model $M(x;\boldsymbol{\theta}) = y$, where $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^q$ is a q-dimensional parameter vector, completely specifies the relationship between x and y.

Here we are concerned with the problem of, having observed x and y, draw inference about $\boldsymbol{\theta}$. From a Bayesian perspective, we are concerned with obtaining the posterior distribution [7]:

$$p(\boldsymbol{\theta}|x,y) \propto p(y,x|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \tag{18.1}$$

$$\propto p(y|x,\boldsymbol{\theta})\pi(x|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \tag{18.2}$$

$$\propto p(y|x,\boldsymbol{\theta})\pi(x)\pi(\boldsymbol{\theta}), \tag{18.3}$$

where (18.3) follows from the assumption of a priori independence of the inputs and parameters.

We discuss several aspects of the uncertainty in such models and illustrate with a temperature-dependent Verhulst model, proposed in [11].

This chapter is organized as follows: in Sect. 18.1 we outline the necessary theory and notation. Sect. 18.2 details the model, likelihood, priors, and posteriors. A simulation study is presented in Sect. 18.2.4 and a discussion with our closing remarks is given in Sect. 18.3.

18.2 Logistic Growth with Temperature-Dependent Parameters

Global temperature change may be an important factor on infectious diseases emergence [6]. Arthropod-borne diseases are particularly influenced by climate change; their vectors are very sensitive to temperature variation. In this chapter we are interested in studying the temperature dependency of population growth of a Chagas disease vector, the insect *Rhodnius prolixus*.

To this end, we introduce a modified logistic growth equation also known as the Verhulst equation [10, 11], with temperature-dependent parameters. The ordinary non-linear differential equation

$$\frac{dP}{dt} = r\left(1 - \frac{P}{K}\right)P, \tag{18.4}$$

takes two parameters, the growth rate r and the carrying capacity K. For a given initial population condition N_0, an analytic solution is available for the number $P(t)$ of individuals at time t:

$$P(t) = \frac{K}{1 + \left(\frac{K - N_0}{N_0}\right) e^{-rt}}. \tag{18.5}$$

In order to incorporate temperature-dependent behavior, we introduce temperature-dependent parameters, $r(T)$ and $K(T)$. The analytic solution in (18.5) is slightly modified to yield $P(t, T)$ for time t and temperature T:

$$P(t, T) = \frac{K(T)}{1 + \left(\frac{K(T) - N_0}{N_0}\right) e^{-r(T)t}}. \tag{18.6}$$

To complete the model we must specify $K(T)$ and $r(T)$ as smooth functions of T. We model $K(T)$ and $r(T)$ as Gaussian kernels over T:

$$K(T) = c_K exp\left(-\frac{(T - a_K)^2}{b_K}\right) \tag{18.7}$$

$$r(T) = c_r exp\left(-\frac{(T - a_r)^2}{b_r}\right). \tag{18.8}$$

18.2.1 Likelihood

Assume $P(t, T)$ to be a Gaussian process (GP) with fixed variance τ^2. Recalling the notation in Sect. 18.1, we have $x = \{T, t, N_0\}$, $y = \{P(t, T)\}$ and $\boldsymbol{\theta} = \{a_K, b_K, c_K, a_r, b_r, c_r, \tau^2\}$. Additionally, note that the parameter vector $\boldsymbol{\theta}$ can be split into the disjoint sets $\boldsymbol{\theta}_K = \{a_K, b_K, c_K\}$ and $\boldsymbol{\theta}_r = \{a_r, b_r, c_r\}$, which parameterize $K(T)$ and $r(T)$, respectively.

Let $\mathbf{y} = \{y_1, y_2, \ldots, y_N\}$ be an output vector with N measurements, which we observe directly. Moreover, let $\mathbf{t} = \{t_1, t_2, ..., t_N\}$ be the vector which contains the observed times of the observations and $\mathbf{T}$ the analogous vector for experimental temperatures. We then specify

$$y_i | t_i, T_i, N_0, \boldsymbol{\theta} \sim \mathcal{N}(\mu(t_i, T_i, N_0; \boldsymbol{\theta}), \tau^2) \tag{18.9}$$

$$\mu(t_i, T_i, \boldsymbol{\theta}) = \frac{K(T_i; \boldsymbol{\theta}_K)}{1 + \left(\frac{K(T_i; \boldsymbol{\theta}_K) - N_0}{N_0}\right) e^{-r(T_i; \boldsymbol{\theta}_r) t_i}}, \quad \forall i = 1, 2, \ldots, N \tag{18.10}$$

which is equivalent to writing $y_i = M(t_i, T_i, N_0; \boldsymbol{\theta}) + \epsilon, \epsilon \sim \mathcal{N}(0, \tau^2)$.

18.2.2 Priors

For $i = 1, 2, .., K$, let $\boldsymbol{\theta}_i$ denote each of the $K = 7$ model parameters. Assuming a priori independence, we have

$$\pi(\boldsymbol{\theta}) = \prod_{i=1}^{K} \pi(\boldsymbol{\theta}_i). \tag{18.11}$$

We then specify prior distributions for the $\boldsymbol{\theta}$ as follows[1]:

$$a_K, a_r \sim Normal(20, 10) \tag{18.12}$$

$$b_K, b_r \sim Gamma(4, 5) \tag{18.13}$$

$$c_K \sim Gamma(1, 1000) \tag{18.14}$$

$$c_r \sim Normal(1/2, 2) \tag{18.15}$$

$$\tau^2 \sim Gamma(1/10, 10). \tag{18.16}$$

These priors were formulated to reflect both model restriction (e.g., positivity and concavity) and biological knowledge about model parameters.

The a_K and a_r parameters mimic the mean parameter in a Gaussian distribution, and control where the functions will achieve their maximum. We assume a normal distribution with moderate variance to provide a relatively uninformative accounting of placement in the positive side of the real line.

For b_K and b_r it is also necessary to ensure positive-definiteness, which we achieve using Gamma priors. Our prior for b allows very "stretched" forms of $K(T)$ and $r(T)$ (see Fig. 18.3), while ensuring concavity. Finally, c_K and c_r control curve height for $K(T)$ and $r(T)$ respectively.

The choice for c_K is essentially arbitrary, since little is known about natural populations of *R. prolixus* in terms of carrying capacity. We thus parameterize $\pi(b_K)$ to reflect rough projections taking into account the number of laid eggs. Since r in Eq. 18.4 can theoretically assume negative values, we used a prior for c_r that allows negative values.

Now, let $\boldsymbol{\theta}^r$ and $\boldsymbol{\theta}^K$ be partitions of the parameter vector, which in combination with the deterministic forms given in (18.7) and (18.8) induce prior distributions on r and K for each fixed temperature T, $\pi^*(r) = r(T; \pi(\boldsymbol{\theta}^r))$, and $\pi^*(K) = K(T; \pi(\boldsymbol{\theta}^K))$ respectively. These prior distributions can be easily approximated using Monte Carlo sampling and are presented in Fig. 18.3 (dashed lines).

[1] Please note that throughout this text we assume gamma distributions are parameterised in terms of shape and scale—as opposed to rate— and normal distributions are parameterised in terms of mean and standard deviation.

18.2.3 Posterior

Hence, from (18.3) and assuming a priori independence of $\mathbf{t}$ and $\mathbf{T}$

$$p(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{t},\boldsymbol{T}) \propto p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{t},\boldsymbol{T})\pi(\boldsymbol{t})\pi(\boldsymbol{T})\pi(\boldsymbol{\theta}), \tag{18.17}$$

is the desired posterior.

This distribution can not be obtained in closed form, therefore we have resort to MCMC techniques to obtain an approximation. Hamiltonian Monte Carlo (HMC)[5] is an MCMC algorithm that replaces the random walk (RW) proposal of the Metropolis-Hastings algorithm by a "leapfrog" proposal based on the gradient of the log-posterior distribution. We then used the **stan** [9] package of the R Statistical Computing Environment [8] to approximate (18.17) through Hamiltonian Monte Carlo (HMC). The HMC implementation of **rstan** relies on the No-U-turn sampler (NUTS) [3] which automatically sets the step size (ϵ) and the number of steps, i.e., trajectory length (L), dropping the need for hand-tuning and making the algorithm efficient for a broad class of target distributions. Among the the advantages of HMC we would like to highlight the ability of making moves far away from the current value with higher acceptance probability than RW, thus providing quicker convergence even for highly correlated target posteriors. R code implementing the posterior approximation presented here is available at https://github.com/maxbiostat/CODE/tree/master/BIDPGM.

The MCMC was run for 50,000 iterations until convergence which was assessed by inspecting the trace- and autocorrelation plots and potential scale reduction factor. After discarding the burn-in we approximate the posterior distribution of $P(t,T)$ using Monte Carlo sampling as follows. Let Q be the number of samples.

1. Construct a grid of values for t and T
2. For each $q = 1, 2, ..., Q$ draw a vector $\boldsymbol{\theta}^{(q)} = \{a_K^{(q)}, b_K^{(q)}, c_K^{(q)}, a_r^{(q)}, b_r^{(q)}, c_r^{(q)}\}$ from the posterior distribution of the parameters
3. Evaluate $M(t,T,N_0;\boldsymbol{\theta}^{(q)})$ to get $P(t,T)^{(q)}$

From these Q samples, we can compute several quantities of interest, for instance the posterior mean of the population for a particular temperature at a given time. Let t_j and T_j represent a particular pair of temperature and time. Then the posterior mean of $P(t_j,T_j)$ is

$$\mathbb{E}\left[P(t_j,T_j)\right] = \frac{1}{Q}\sum_{q=1}^{Q} P(t_j,T_j)^{(q)} = \frac{1}{Q}\sum_{q=1}^{Q} M(t_j,T_j,N_0;\boldsymbol{\theta}^{(q)}). \tag{18.18}$$

The posterior median and quantiles can be obtained in a similar fashion.

Table 18.1 Bias assessment using the simulated data set— posterior mean

Parameter	Value[a]	Posterior mean	Bias[b]	MSE	Coverage[c]
a_K	30.00	29.44	0.01	7.99	0.93
a_r	23.00	22.71	0.00	3.31	0.86
b_K	10.00	13.08	0.95	450.26	0.94
b_r	15.00	16.82	0.22	16.77	0.88
c_K	700.00	692.17	0.09	7203.06	0.96
c_r	0.40	0.43	0.00	0.04	0.85
τ	3.16	4.89	0.94	67.02	0.88

[a] "True" value used for simulation
[b] Bias divided by the true parameter value
[c] Coverage of the 95 % credibility intervals

18.2.4 Simulation Study

We conducted a Monte Carlo simulation study in order to evaluate the approach proposed here. Simulation was carried out as follows: for each m in a total of M simulation steps,

1. Fix $\boldsymbol{\theta}$, τ and N_0
2. Construct a grid of values for t and T
3. For each point in the grid, sample from a normal distribution with mean $M(T, t, N_0; \boldsymbol{\theta})$ and variance τ, generating a data set $P^{(m)}$
4. Using 50,000 iterations of HMC, obtain an estimate $\hat{\boldsymbol{\theta}}^{(m)}$ of model parameters

In this chapter, we use both the *a posteriori* mean and median as point estimates for $\boldsymbol{\theta}$. With these results at hand, we are able to compute the normalized squared bias and mean squared error (MSE) for each parameter, as well as nominal coverage for the 95 % credibility intervals. The normalized squared bias for each parameter is defined as

$$B(\theta_i) = \theta_i^{-1}\mathbb{E}\left[\hat{\theta}_i - \theta_i\right]^2. \tag{18.19}$$

Let $Z_i^{(m)}$ be the indicator variable that assumes 1 if the mth 95 % credibility interval contains the true value of θ_i and zero otherwise. Coverage is defined as

$$C(\boldsymbol{\theta}_i) = \mathbb{E}\left[Z_i^{(m)}\right] = \frac{1}{M}\sum_{j=1}^{M} Z_{ij}. \tag{18.20}$$

Code for this step is also available at https://github.com/maxbiostat/CODE/tree/master/BIDPGM.

Table 18.1 shows the bias, MSE and coverage for each parameter using the posterior mean as a point estimate. Parameter values were chosen so as to reflect a biologically sound behavior for $P(t, T)$.

The results for the posterior median as point estimator were largely in agreement with those for the posterior mean and thus are omitted here.

In Fig. 18.1 we present prior and posterior distributions for model parameters used in this simulation study where we can notice the posteriors are substantially less diffuse than the priors.

18.2.5 R. prolixus data

We now turn our attention to a real-world data set. Chagas disease is an important tropical disease, transmitted by a blood sucking bug, *R. prolixus*. Temperature is key factor for both insect development and vector competence. We are thus interested in drawing inference on model parameters to understand the role of temperature in the insect's population dynamics.

In a laboratory experiment, $N_0 = 30$ females were observed in several temperatures, and the number of eggs laid was recorded for 35 days. We take the cumulative number of ecloded eggs for each temperature condition as a good approximation of $P(t, T)$, since all conditions (light, water, food, etc.) were optimal. The data thus consisted of $N = 350$ observations, from 10 temperature conditions for 35 days each.

In Fig. 18.2 we show prior and posterior distributions for each parameter using this data. Results are similar to that of the simulated data, with substantially concentrated posteriors in comparison to the priors. Interestingly, while the prior expectation for b_K was 25, we obtained a posterior mean around 106 (Table 18.2), indicating that the variability of the response of *Rhodnius* to temperature is much greater than we anticipated. See more on this at Sect.18.3.

Figure 18.3 shows the prior and posterior (induced) distributions of the thermodynamic functions. These are easily obtained using Monte Carlo sampling from the prior and posterior distributions of $\boldsymbol{\theta}$, respectively.

Finally, we present the posterior distribution for $P(t, T)$ obtained using the approach described in Sect. 18.2.4. We also present the heat map of the original laboratory data described above and the posterior distribution obtained by Monte Carlo sampling of the marginal posteriors from MCMC (Fig. 18.4). It can be noticed that the posterior distribution allows a region of optimality for populational growth around $20 - 25\,^{\circ}\mathrm{C}$. Each thermodynamic function, $r(T)$ and $K(T)$ can have its own point of maximum, however.

In Table 18.2 we provide posterior estimates obtained for the real data along with prior expectations and 95 % credibility intervals.

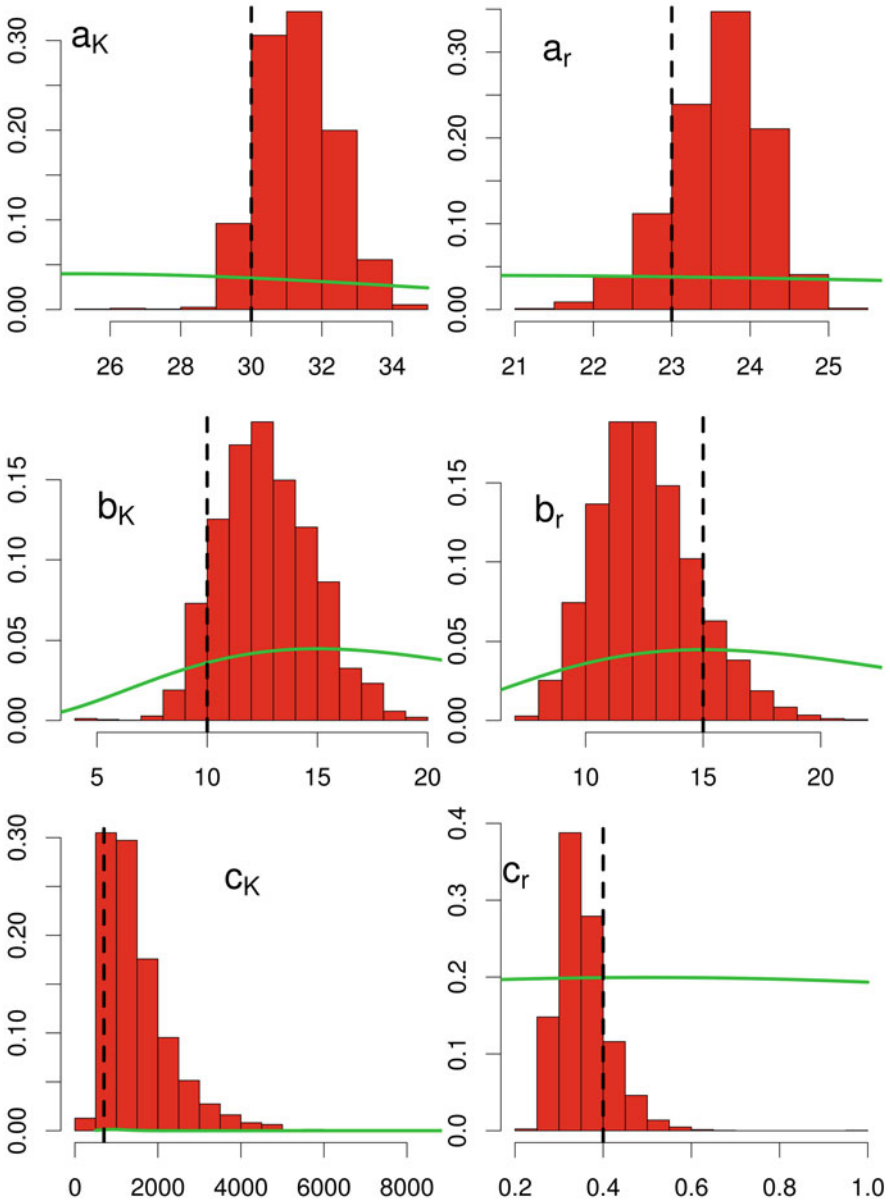

Fig. 18.1 Prior and posterior distributions of parameter for the simulated data. *Red bars* (histogram) show the marginal posterior distributions and *green lines* depict prior densities. *Dashed vertical lines* mark true parameter values. Please note that vertical axes differ

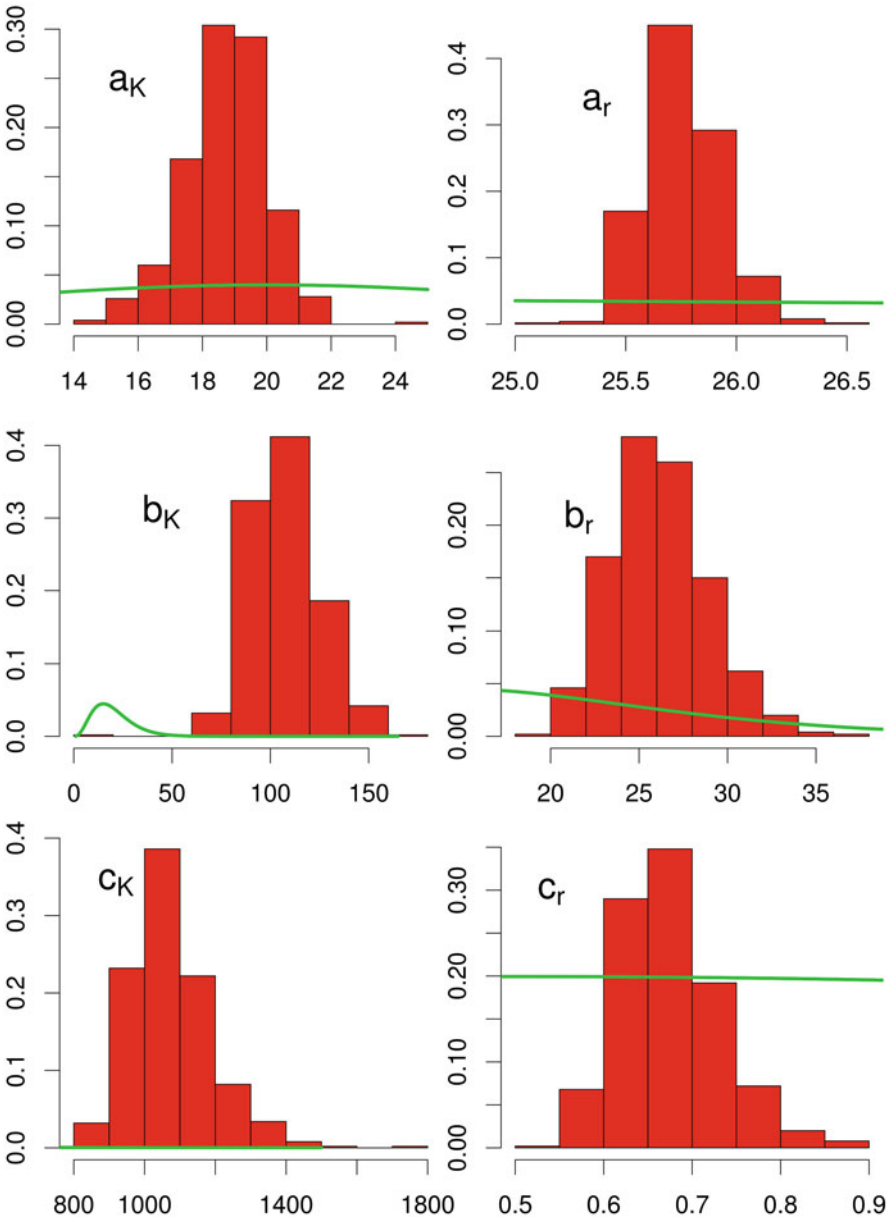

Fig. 18.2 Prior vs posterior distributions of parameter for the real data set. Please note that vertical axes differ

Table 18.2 Posterior inference results for the real set. We report posterior and prior expectations, along with the appropriate 95 % credibility intervals. Five independent chains were run for 50,000 iterations each with the first 25,000 discarded as burn-in. Convergence was assessed using trace plots and the potential scale reduction factor

	Posterior mean (95 % C.I.)	Prior mean (95 % C.I.)
a_K	19.23 (17.56 – 21.09)	25.00 (5.40–44.60)
a_r	25.73 (25.44–26.10)	25.00 (5.40–44.60)
b_K	106.17 (75.25–137.31)	20.00 (5.44–43.84)
b_r	26.77 (22.59–32.19)	20.00 (5.44–43.84)
c_K	1023.32 (898.28–1165.40)	1000.00 (25.31–3688.87)
c_r	0.66 (0.58–0.76)	0.50 (−3.41–4.41)
τ	177.33 (166.10–191.78)	1.00 (0.00–9.78)

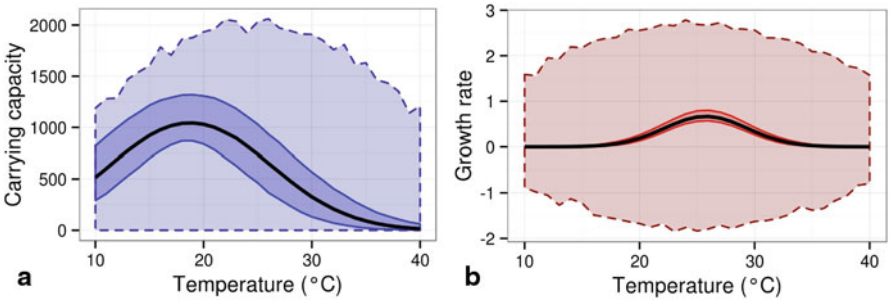

Fig. 18.3 Prior and posterior credibility intervals for the thermodynamic functions under for the real data set. We show $K(T)$ in (**a**) and $r(T)$ in (**b**). *Dashed lines* and *lighter tones* depict prior and *solid lines* and *darker tones* represent the posterior 95 % credibility intervals. *Thick solid lines* mark the medians

18.3 Discussion

Deterministic, differential-equation-based models are an important tool in Theoretical Biology, allowing us to study the behavior of biological systems using simple and easily interpretable equations. In this chapter we adapt a classical population growth model, the Verhulst logistic equation, initially proposed in 1838, to accommodate temperature-dependent parameters. We then proceed to develop a Bayesian approach to learn about model parameters when population time series are available.

Our approach is based on the idea that both the growth rate r and the carrying capacity K are smooth functions of temperature, which we model as Gaussian kernels of the form presented in (18.7). This parameterisation is very flexible, allowing us to model a variety of temperature-response patterns. It is biologically motivated, since the response of the insect populations to temperature change should have a maximum point and substantially decrease at extremely high and extremely low temperatures [11].

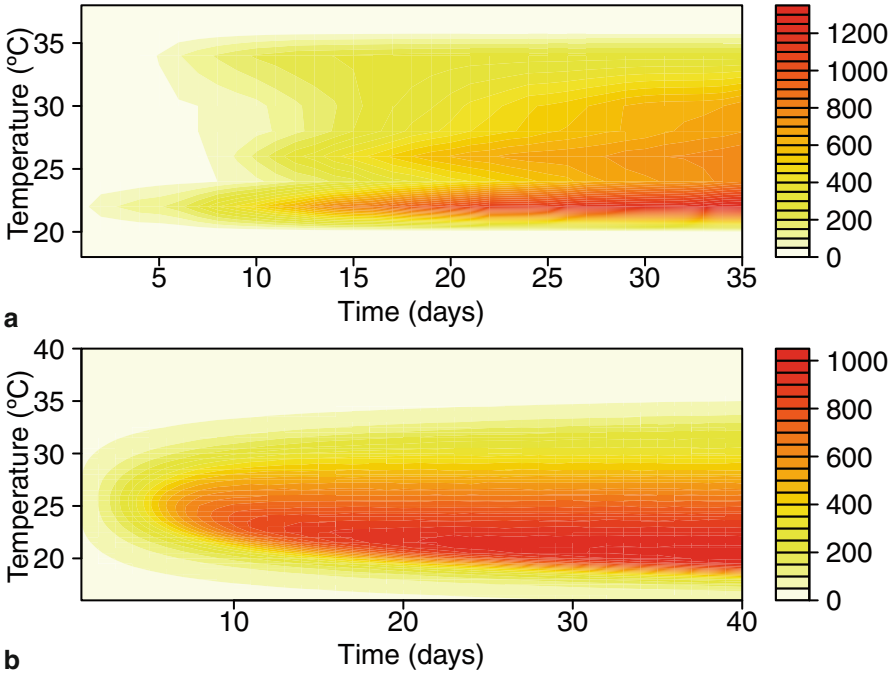

Fig. 18.4 Heat maps showing population through time and temperature. In **a** we show the laboratory data collected for *Rhodnius* and in **b** the posterior mean, obtained using Monte Carlo sampling from the posterior distribution of the parameters (Sect. 18.2.4). Please note that legend ranges differ slightly

The main idea is to model population through time as a GP with a deterministic mean function $M(\cdot)$ which is given by the solution to the differential equation (Eq. 18.6). The posteriors outlined in (18.3) and (18.17) allow for the incorporation of uncertainty regarding model inputs. Although this source of uncertainty is negligible in our setting due to carefully controlled experimental conditions, it could play an important role in studies dealing with field data.

It should also be noted that in this chapter we assume independence between model inputs. This assumption may be irrealistic, since temperature is likely to depend on time in general. In our particular conditions, all experiments were performed in controlled environments, where temperature was kept constant throughout time. In the case of population time series obtained from field data, this assumption is likely not to hold.

From Table 18.1 it can be noticed that our approach presents good frequentist properties, with high coverage probabilities of the credibility intervals for most parameters. The only exception is the GP variance, τ^2, for which we could not recover the true value with good accuracy, albeit achieving good coverage. This result most likely stems from the restricting assumption of fixed variance (over time). Replacing

the fixed variance by a smooth function of time $\tau^2(t)$ could greatly improve model fit and is an important future direction of research.

In conclusion, in this chapter we provide insight on how to perform inference on the parameters of a complex, non-linear deterministic model of population growth when population time series are available. These parameters are directly interpretable and provide important information about the underlying biological dynamics. The framework proposed here can be adapted to a broad class of models, for example, to learn about the parameters of epidemiological models using data on disease incidence. Moreover, the Bayesian approach allows for a complete treatment of uncertainty on model inputs and outputs, thus providing more realistic predictions when data is subject to measurement uncertainty. The issue of how to incorporate and accommodate uncertainty about model structure is also an important one and constitutes an interesting avenue for future research.

Acknowledgements The authors would like to thank Professor Angela H. Lopes, Dr. Luciana Zimmermann and Luiz R. Vasconcellos (UFRJ) for providing the data set analysed in this chapter and for fruitful discussions. CJS was partially funded by CNPq and FAPERJ.

References

1. Costello, W.G., Taylor, H.M.: Deterministic population growth models. Am. Math. Mon. **78**(8), 841–855 (1971)
2. Gillespie, C.S., Golightly, A.: Bayesian inference for generalized stochastic population growth models with application to aphids. J. R. Stat. Soc.: Ser. C (Appl. Stat.) **59**(2), 341–357 (2010)
3. Hoffman, M.D., Gelman, A.: The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. J. Mach. Learn. Res. **15**, 1593–1623 (2014)
4. May, R.M., et al.: Simple mathematical models with very complicated dynamics. Nature **261**(5560), 459–467 (1976)
5. Neal, R.M.: MCMC using Hamiltonian dynamics. In: Brooks, S., Gelman, A., Jones, G., Meng, X.-L. (eds.) Handbook of Markov Chain Monte Carlo. Chapman & Hall/CRC Press (2010)
6. Patz, J.A., Epstein, P.R., Burke, T.A., Balbus, J.M.: Global climate change and emerging infectious diseases. J. Am. Med. Assoc. **275**(3), 217–223 (1996)
7. Poole, D., Raftery, A.E.: Inference for deterministic simulation models: the Bayesian melding approach. J. Am. Stat. Assoc. **95**(452), 1244–1255 (2000)
8. R Core Team : R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2013). http://www.R-project.org/
9. Stan Development Team: Stan: a C++ library for probability and sampling, version 2.2 (2014). http://mc-stan.org/
10. Verhulst, P.F.: Notice sur la loi que la population suit dans son accroissement. Correspondance mathématique et physique publiée par a. Quetelet **10**, 113–121 (1838)
11. Zimmermann, L.T., Carvalho, L.M.F., Vasconcellos, L.R., Bastos, L.S., Struchiner, C.J., Lopes, A.H.: Temperature-dependent oviposition and egg eclosion of Chagas disease vector *Rhodnius prolixus* . Submitted (2014)

Chapter 19
A Weibull Mixture Model for the Votes of a Brazilian Political Party

Rosineide F. da Paz, Ricardo S. Ehlers and Jorge L. Bazán

Abstract Statistical modeling in political analysis is used recently to describe electoral behaviour of political party. In this chapter we propose a Weibull mixture model that describes the votes obtained by a political party in Brazilian presidential elections. We considered the votes obtained by the Workers' Party in five presidential elections from 1994 to 2010. A Bayesian approach was considered and a random walk Metropolis algorithm within Gibbs sampling was implemented. Next, Bayes factor was considered to the choice of the number of components in the mixture. In addition the probability of obtain 50 % of the votes in the first round was estimated. The results show that only few components are needed to describe the votes obtained in this election. Finally, we found that the probability of obtaining 50 % of the votes in the first ballot is increasing along time. Future developments are discussed.

19.1 Introduction

Statistical modelling in political analysis has been used recently to describe the electoral behaviour of a political party and examples of study of voting behaviour are [12]. In Brazil the electoral behaviour underwent a process of change since 1994 to the recent days. In 1994, Brazilians voted in one of the most important elections held since 1945. This was the second election held since the end of military rule from 1964 to 1985. In terms of the presidential vote, in 1994 the candidate of The Brazilian Social Democracy Party (in Portuguese: Partido da Social Democracia Brasileira, PSDB) won the majority of votes on the first ballot (54.3 %) and the candidate of Workers' Party (in Portuguese: Partido dos Trabalhadores, PT) obtained 38.4 % of

R. F. da Paz (✉)
Universidade Federal de São Carlos, São Carlos, Brazil

Instituto de Ciências Matemáticas e de Computação. USP. São Carlos, SP. Brazil
e-mail: rfpaz@icmc.usp.br

R. S. Ehlers · J. L. Bazán
Instituto de Ciências Matemáticas e de Computação. USP. São Carlos, SP. Brazil
e-mail: ehlers@icmc.usp.br

J. L. Bazán
e-mail: jlbazan@icmc.usp.br

A. Polpo et al. (eds.), *Interdisciplinary Bayesian Statistics,*
Springer Proceedings in Mathematics & Statistics 118, DOI 10.1007/978-3-319-12454-4_19

the total of votes. For more information about this election see for example [16] or the Superior Electoral Court (TSE) website `http://english.tse.jus.br`. However the Workers' Party elected its candidates for president in the last three elections occurring in 2002, 2006 and 2010. Results on the presidential elections in Brazil are available on the TSE website.

In order to investigate the probabilistic behaviour of votes obtained by a political party in Brazilian general elections we propose a Weibull mixture model. In particular, we considered the presidential votes obtained by Workers' Party in the five elections from 1994 to 2010 using data obtained from TSE website. As seen in [4], analysts have argued that the social policies that President Luis Inacio Lula da Silva implemented enabled the number of voters of the Workers' Party to expand from middle-class and highly educated people to low-income and poorly educated individuals from the Northeast of Brazil. From the nine states in the Northeast region we chose to analyse the data from Sergipe State (SE) for illustration purposes because this is the state with the smallest number of electoral districts.

For estimation purposes, a Bayesian approach was considered and a random walk Metropolis algorithm within Gibbs sampling was implemented. Next, a Bayes factor approach was considered to the choice of the number of components in the mixture. Finally the probability of obtaining 50 % of the votes in the first ballot was estimated. The results show that only a few components are needed in the mixture to describe the votes obtained in this election. In addition we found that the probability of obtaining 50 % of the votes in the first ballot is increasing along time.

The rest of the chapter is organized as follows. In Sect. 19.2, we describe the finite mixture model and a finite mixture of Weibull distributions is proposed to model the data of votes of a Brazilian political party. In Sect.19.3 we present the main results and the future developments are discussed in Sect.19.4.

19.2 Statistical Model

19.2.1 The Votes of a Political Party

Percentages of votes obtained by a political party in an election in different cities of a region or country can be assumed as a random variable $X > 0$. As usually observed in the histogram of this type of data, they present a positive asymmetric distribution, that is the votes are concentred in lower percents and occasionally are observed higher values and the mean is greater than the median. As suggested by Durtschi et al. [8], these are some characteristics of Benford's Law which has been invoked as evidence of elections data for example by Bërdufi [2]. Recently, Cuff et al. [7] have established the relation between the Weibull distribution and Benford's Law.

In this chapter the percentage of votes to each city are assumed to follow a Weibull distribution which is governed by two parameters, that is $X \sim W(\delta, \eta)$; being zero the lower end of its support. The parameter δ is a shape parameter and η is a scale

parameter. η determines the scale along its support of votes and the parameter δ, determines the concentration of the distribution of votes. High values of η correspond to a high degree of concentration (low dispersion) of votes. The Weibull distribution can be seen as a generalisation of the exponential distribution and commonly describes the time we have to wait for one event to occur, if that event becomes more or less likely with time. Here the η parameter describes how quickly the probability ramps up (proportional to $x^{\eta-1}$). For $0 < \eta < 1$, the density function tends to ∞ if x approaches zero from above and is strictly decreasing. For $\eta = 1$, the density function of votes tends to $1/\delta$ to lower votes x approaches zero from above and is strictly decreasing. For $\eta > 1$, the density function of votes tends to zero as the votes x approaches zero from above, increases until its mode $\delta\left(\frac{\eta-1}{\eta}\right)^{1/\eta}$ and decreases after it.

Additionally by observing the histogram of percent of votes by cities multimodality is identified, that is it is possible to identify different populations (clusters of votos) because a different electoral behaviour between cities is expected. Consequently a Weibull mixture distribution can be assumed in order to identify these sub populations. In [22] this type of distribution has been considered in different areas but similar situations.

19.2.2 Mixture Model

Finite mixture of distributions is a flexible method of modelling. Its more direct role in data analysis and inference is to provide a convenient and flexible family of distributions to estimate or approximate distributions which are not well modelled by any standard parametric family. This type of model is useful in the modelling of data from a heterogeneous population, that is a population which can be divided in clusters or components. In this sense, the components in the data can be modelled for unimodal distributions belonging to a same parametric distribution family. For more details about modelling and applications of finite mixture models, see for example [15].

By observation of the data in Fig. 19.1 we propose to model these data as a k-component mixture of distributions.
This finite mixture approach is flexible enough for modelling the sort of data that appears in such an experiment where the data clearly exhibits multimodality.

A random variable X is said to follow a finite mixture of distributions if its density function is given by,

$$f(x|\boldsymbol{\theta}, w) = \sum_{j=1}^{M} w_j f_j(x|\boldsymbol{\theta}_j), \tag{19.1}$$

where each $f_j(x|\boldsymbol{\theta}_j)$ is a density function indexed by a parameter vector $\boldsymbol{\theta}_j$ and the weights $w_1, \ldots, w_M$ are such that $0 < w_j < 1$ and $\sum_{j=1}^{M} w_j = 1$ and M is the

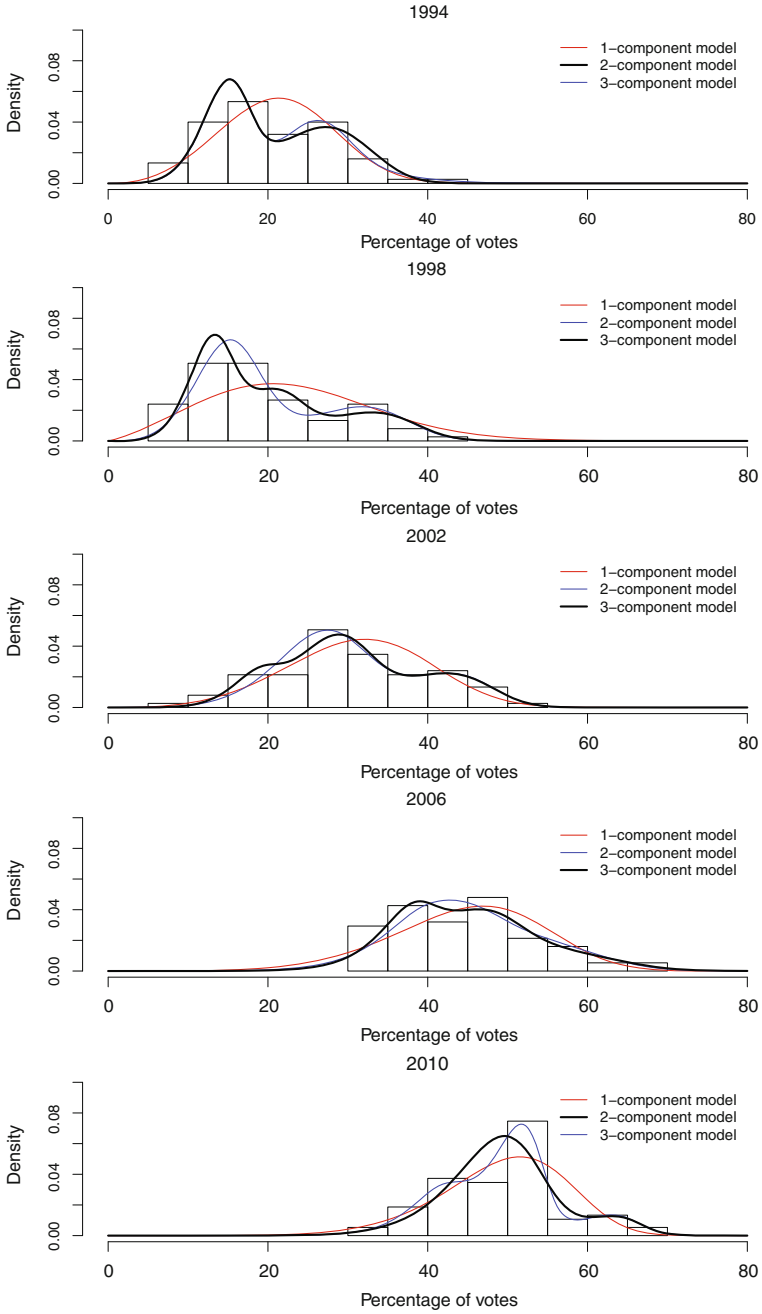

Fig. 19.1 Histograms of the data of voting percentage obtained by Workers' Party in presidential elections, in the cities of Sergipe State, from year 1994 and 1998, when the PT lost the presidential election, to 2002, 2006 and 2010, when the PT candidate was Presidential winner, and its estimated densities based on the posterior predictive distribution for 1, 2 and 3 components

number of components of mixture. See for example [13], for a review on existing techniques for Bayesian modelling and inference on mixtures of distributions.

Suppose that $x = (x_1, ..., x_n)$ are independent and identically distributed coming from a distribution with the probability density function 19.1. The likelihood function is given by,

$$L(\boldsymbol{\theta}, w) = \prod_{i=1}^{n} \sum_{j=1}^{M} w_j f(x_i|\theta_j).$$

To eliminate the summation in this last expression, it is convenient to introduce the auxiliary vectors $Z_i = (Z_{i1}, ..., Z_{iM})$ where $Z_{ij} = 1$ if the ith observation belongs to the jth mixture component, $Z_{ij} = 0$ otherwise and $P(Z_{ij} = 1) = w_j$. The distribution of each X_i given Z_i has density function given by,

$$f(x_i|Z_i, \boldsymbol{\theta}) = \prod_{j=1}^{M} [f(x_i|\theta_j)]^{Z_{ij}},$$

and the joint distribution of (X_i, Z_i) can be written as

$$f(x_i, Z_i|\boldsymbol{\theta}, w) = \prod_{j=1}^{M} [w_j f(x_i|\theta_j)]^{Z_{ij}}.$$

Therefore, the complete data likelihood function is given by,

$$L(\boldsymbol{\theta}, w) = \prod_{i=1}^{n} \prod_{j=1}^{M} [w_j f(x_i|\theta_j)]^{Z_{ij}}.$$

This hierarchical representation of the model facilitates the Bayesian analysis.

In this chapter we assume a finite mixture of Weibull distributions for each X_i where the jth component has scale and shape parameters η_j and δ_j respectively. We prefer Weibull distribution since that gives a distribution for which the failure rate is proportional to a power of time and the parameters of the model are easily interpretable. Consequently other distributions were discarded.

Considering Weibull distribution the complete data likelihood function is then given by,

$$\prod_{i=1}^{n} \prod_{j=1}^{M} \left[w_j \frac{\delta_j}{\eta_j} \exp\left(- \left(\frac{x_i}{\eta_j} \right)^{\delta_j} \right) \left(\frac{x_i}{\eta_j} \right)^{\delta_j - 1} \right]^{Z_{ij}}. \tag{19.2}$$

19.2.3 Prior Specification

Following the Bayesian paradigm, we need to complete the model specification by assigning prior distributions to the parameters. Then, by applying the Bayes theorem

the posterior density is proportional to the product the likelihood function 19.2 by the prior density.

We shall assume that all the parameters are a priori independent. Then, within each component, gamma prior distributions are assigned to the Weibull parameters, that is $\eta_j \sim Gamma(a_j, b_j)$ and $\delta_j \sim Gamma(c_j, d_j)$, $j = 1, \ldots, M$, here the notation $Y \sim Gamma(c_j, d_j)$ means that the random variable Y follows a gamma distribution with parameters c_j and d_j. Also, since the vector of weights w is defined on the simplex $\{w \in \mathbb{R}^M : 0 < w_j < 1, j = 1, ..., M, \sum_{j=1}^{M} w_j = 1\}$ we consider a Dirichlet prior distribution for w. Its prior density function is then given by,

$$p(w|\nu_1, \ldots, \nu_M) = \frac{\Gamma(\nu_1 + \cdots + \nu_M)}{\Gamma(\nu_1) \cdots \Gamma(\nu_M)} \prod_{j=1}^{M} w_j^{\nu_j - 1}$$

where $\nu_1 > 0, \ldots, \nu_M > 0$ are the hyperparameters. In this chapter, the hyperparameters η_j, δ_j and ν_j, $j = 1, \ldots, M$ are held fixed.

Finally, we need to impose identifiability constraints since the labelling of the mixing components is arbitrary and we need some rule to discriminate among the components (see for example [11]). A typical solution, also adopted here, is to impose an ordering constraint $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_M$ where μ_j is the mean of the jth component in the mixture.

19.2.4 Inference

Since the posterior density cannot be fully obtained in closed form we use a Markov chain Monte Carlo (MCMC) approach to simulate parameter values and obtain parameter estimates. Details of MCMC methods can be found for example in [21]. In order to obtain a sample from the joint posterior distribution of the parameters we first obtain the complete conditional distributions. First note that

$$\begin{aligned} P(Z_{ij} = 1|x, \boldsymbol{\eta}, \boldsymbol{\delta}, w) &\propto f(x_i|Z_{ij} = 1, \eta_j, \delta_j) P(Z_{ij} = 1|w_j) \\ &\propto w_j \frac{\delta_j}{\eta_j} \exp\left(-\left(\frac{x_i}{\eta_j}\right)^{\delta_j}\right)\left(\frac{x_i}{\eta_j}\right)^{\delta_j - 1} \end{aligned}$$

for $i = 1, \ldots, n$. So, for each observation we just need to sample $j \in \{1, \ldots, M\}$ with probability proportional to $w_j \frac{\delta_j}{\eta_j} \exp\left(-\left(\frac{x_i}{\eta_j}\right)^{\delta_j}\right)\left(\frac{x_i}{\eta_j}\right)^{\delta_j - 1}$. Now, combining the likelihood function 19.2 with the prior densities of δ_j and η_j it follows that,

$$p(\eta_j|x, Z, \boldsymbol{\delta}, \boldsymbol{\eta}_{-j}) \propto \eta_j^{a_j - n_j\delta - 1} \exp\left\{-\sum_{i:Z_{ij}=1}\left(\frac{x_i}{\eta}\right)^{\delta_j} - \eta_j b_j\right\}$$

$$p(\delta_j|x, Z, \boldsymbol{\eta}, \boldsymbol{\delta}_{-j}) \propto \delta_j^{n_j + c_j - 1} \eta^{-n_j\delta} \exp\left\{-\sum_{i:Z_{ij}=1}\left(\frac{x_i}{\eta}\right)^{\delta_j} - d_j\delta_j\right\} \prod_{i:Z_{ij}=1} x_i^{\delta_j - 1}$$

where $n_j = \sum_{i=1}^{n} Z_{ij}$ denotes the number of observations in the jth mixture component.

The complete conditional density of each δ_j and η_j is not of any standard form and we use a Metropolis-Hastings algorithm. We adopt a random walk Metropolis algorithm by proposing values of $\log(\delta_j)$ and $\log(\eta_j)$ from a normal distribution centered about its current value and fixing the variance to tune the acceptance rates.

Finally, the complete conditional density of w is given by,

$$p(w|x, Z, \boldsymbol{\delta}, \boldsymbol{\eta}) \propto \prod_{j=1}^{M} w_j^{\nu_j + n_j - 1},$$

which represents a Dirichlet distribution with parameters $\nu_1 + n_1, \ldots, \nu_M + n_M$. Sampling from this complete conditional distribution is then accomplished by drawing independent gamma variables and scaling them to sum to 1.

In this work, the Gibbs sampling method is used combined with Metropolis-Hasting algorithm to obtain sample of the posterior distribution of parameters $\boldsymbol{\delta}, \boldsymbol{\eta}, w$ and Z_i, for $i = 1, ..., N$, see [13]. The Gibbs sampling algorithm can be written as follows,

1. Initialize choosing $w^{(0)}$, $\delta_j^{(0)}$ and $\eta_j^{(0)}$, for $j = 1, ..., M$
2. For $t = 1, 2, \ldots$ repeat
   a) For $i = 1, ..., N$ generate $Z_i^{(t+1)} \sim Multinomial(1, \hat{\pi}_{i1}^{(t)}, ..., \hat{\pi}_{iM}^{(t)})$, wherein

$$\hat{\pi}_{ij}^{(t)} = p(Z_{ij}^{(t)} = 1|x_i, \delta_j^{(t-1)}, \eta_j^{(t-1)}) = \frac{p(x_i|\delta_j^{(t-1)}, \eta_j^{(t-1)}, M) w_j^{(t-1)}}{\sum_{j=1}^{M} w_j^{(t-1)} p(x_i|\delta_j^{(t-1)}, \eta_j^{(t-1)}, M)} \tag{19.3}$$

   b) Generate $w^{(t)}$ from the $p(w|Z^{(t)})$
   c) For $j = 1, ..., M$ do
      (i) Generate $\left(\delta_j^{'}, \eta_j^{'}\right) \sim Lognormal\left(\left(\log(\delta_j^{(t-1)}), \log(\eta_j^{(t-1)})\right), \sigma_j^2 I\right)$ with $\sigma_j^2 = 0.05$.
      (ii) Generate $u \sim Uniform(0, 1)$
      (iii) Compute
      $\alpha\left(\left(\delta_j^{(t-1)}, \eta_j^{(t-1)}\right), \left(\delta_j^{'}, \eta_j^{'}\right)\right) =$
      $\min\left\{1, \frac{p\left(\left(\delta_j^{'}, \eta_j^{'}\right)|x, Z\right)}{p\left(\left(\delta_j^{(t-1)}, \eta_j^{(t-1)}\right)|x, Z\right)} \frac{LN\left(\left(\delta_j^{(t-1)}, \eta_j^{(t-1)}\right)|\left(\log(\delta_j^{'}), \log(\eta_j^{'})\right), \sigma_j^2 I\right)}{LN\left(\left(\delta_j^{'}, \eta_j^{'}\right)|\left(\log(\delta_j^{(t-1)}), \log(\eta_j^{(t-1)})\right), \sigma_j^2 I\right)}\right\}$
      where $LN(y|.)$ is the density of log-normal distribution evaluate of y
      (iv) If $\alpha\left(\left(\delta_j^{(t-1)}, \eta_j^{(t-1)}\right), \left(\delta_j^{'}, \eta_j^{'}\right)\right) < u$ then $\left(\delta_j^{(t)}, \eta_j^{(t)}\right) = \left(\delta_j^{'}, \eta_j^{'}\right)$ else $\left(\delta_j^{(t)}, \eta_j^{(t)}\right) = \left(\delta_j^{(t-1)}, \eta_j^{(t-1)}\right)$.

19.2.5 Choosing the Number of Components

The specification of a mixture model involves the determination of the number of components M. Here, instead of endeavouring to apply more complex methods as in [18] and [20] for example, we compare models by means of Bayes factor and marginal likelihood [3]. To describe de Bayes factor suppose two models M_j and M_k with equal prior probabilities $p(M_j)$ and $p(M_k)$. The Bayes factor is obtained as the ratio of marginal likelihood $m_1(x)$ and $m_2(x)$, such that

$$B_{jk} = \frac{p(x|M_j)}{p(x|M_k)} = \frac{p(M_j|x)}{p(M_k|x)} \frac{p(M_k)}{p(M_j)} = \frac{m_j(x)}{m_k(x)}, \tag{19.4}$$

where $p(M_j|x)/p(M_k|x)$ is the posterior odds and $p(M_j)/p(M_k)$ is the prior odds. Since the Bayes factor is higher than 1 then M_j has a higher posterior probability.

In particular, the marginal likelihood for the M-component mixture model is given by,

$$p(x|M) = \int p(x|\boldsymbol{\theta}, M) p(\boldsymbol{\theta}|M) d\boldsymbol{\theta},$$

where $\boldsymbol{\theta}$ denotes the complete set of parameters $(\boldsymbol{\eta}, \boldsymbol{\delta}, w)$. Computation of the marginal likelihood requires proper prior distributions and the analytic evaluation of this integral is not possible in the context treated here (see for example [5] for an extensive description and comparison of available numerical strategies). In this chapter, we compute an approximation to the marginal likelihood based on the MCMC output using the methods described in [6]. The estimator is based on the identity

$$m(x) = \frac{f(x|\boldsymbol{\theta}, M) p(\boldsymbol{\theta}|M)}{p(\boldsymbol{\theta}|x, M)}, \tag{19.5}$$

where the numerator can be directly computed. Thus the calculation of the marginal likelihood is reduced to finding an estimate of the posterior density at a point $\boldsymbol{\theta}^*$. For estimation efficiency we take the point $\boldsymbol{\theta}^* = (\boldsymbol{\eta}^*, \boldsymbol{\delta}^*, w^*)$ as the posterior mean of $\boldsymbol{\theta}$ in the M-component model. We now drop the dependence on M to simplify the notation and note that the posterior density ordinate can be rewritten as,

$$p(\boldsymbol{\theta}^*|x) = p(\boldsymbol{\delta}^*, \boldsymbol{\eta}^*|x) p(w^*|x, \boldsymbol{\eta}^*, \boldsymbol{\delta}^*).$$

Our approach is based on an additional G iterations sampling values of Z from its complete conditional distributions evaluated at (δ_j^*, η_j^*) and sampling values of (δ_j, η_j) from its proposal distribution in the Metropolis-Hastings step also evaluated at (δ_j^*, η_j^*).

Let $q((\delta_j, \eta_j), (\delta_j^{'}, \eta_j^{'}))$ denote the proposal density of (δ_j, η_j) in the random walk Metropolis update. Then, following [6] the marginal density ordinate of each (δ_j^*, η_j^*) is estimated as

$$\hat{p}((\delta_j^*, \eta_j^*)|x) = \frac{L^{-1} \sum_{l=1}^{L} \alpha\left((\delta_j^{(l)}, \eta_j^{(l)}), (\delta_j^*, \eta_j^*)\right) q\left((\delta_j^{(l)}, \eta_j^{(l)}), (\delta_j^*, \eta_j^*)\right)}{G^{-1} \sum_{g=1}^{G} \alpha\left((\delta_j^*, \eta_j^*), (\delta_j^{(g)}, \eta_j^{(g)})\right)},$$

where $\{(\delta_j^{(l)}, \eta_j^{(l)})\}$ are the sampled values from the marginal posterior distribution of (δ_j, η_j), $\{(\delta_j^{(g)}, \eta_j^{(g)})\}$ are drawn from $q\left((\delta_j^*, \eta_j^*), (\delta_j, \eta_j)\right)$ and $\alpha(\cdot, \cdot)$ are the acceptance probabilities. The conditional density ordinates of w_j is estimated by averaging with respect to the sampled values $Z^{(g)}$ their complete conditional densities evaluated at (δ_j^*, η_j^*), that is

$$\hat{p}(w_j|x, \delta_j^*, \eta_j^*) = G^{-1} \sum_{g=1}^{G} p(w_j|x, Z^{(g)}, \delta_j^*, \eta_j^*).$$

Finally, the posterior density ordinate is estimated as,

$$\hat{p}(\boldsymbol{\theta}^*|x) = \prod_{j=1}^{M} \hat{p}((\delta_j^*, \eta_j^*)|x)\hat{p}(w_j^*|x, \eta_j^*, \delta_j^*),$$

which is in turn used in 19.5 to obtain an estimate of the marginal likelihood.

19.2.6 Predictive Distribution

A posterior feature of interest is the predictive distribution for a future observation. As discussed by Escobar and West [9], a density estimation can be obtained by summarizing the unconditional predictive distribution

$$h(x) = p(x_{N+1}|x) = \int p(x_{N+1}|\Theta)dp(\Theta|x) = E_{\theta|x}\left[f(x|\Theta)\right]. \tag{19.6}$$

Thus, the Monte Carlo approximation for $h(x)$ is obtained as

$$\hat{h}(x) = \frac{1}{L} \sum_{l=1}^{L} f(x|\Theta^{(l)}), \tag{19.7}$$

where $\left\{\Theta^{(l)}\right\}_{l=1}^{L}$ are draws from the joint posterior distribution.

In this work, the posterior predictive distribution is used to calculate cumulative probabilities by use of numerical integration such as the Simpson rule, see details of numerical integration methods in [1].

19.3 Results

For each data set of vote percentage, obtained considering elections of 1994, 1998, 2002, 2006 and 2010, as seen above, we have implemented the algorithm shown in Sect.19.2.4 to mixtures of Weibull distributions. In terms of MCMC, we report

Table 19.1 Twice the natural logarithm of the Bayes factor of the data of voting percentage under one model resulting of mixture of Weibull distribution relative to another

Election	$2 \times \log\left(\frac{p(x\|2\text{-component})}{p(x\|1\text{-component})}\right)$	$2 \times \log\left(\frac{p(x\|2\text{-component})}{p(x\|3\text{-component})}\right)$	$2 \times \log\left(\frac{p(x\|3\text{-component})}{p(x\|1\text{-component})}\right)$
1994	560.0	133.3	426.8
1998	753.7	−7.9	761.6
2002	607.9	−3.1	610.9
2006	625.6	−9.4	635.1
2010	562.2	18.3	543.8

results corresponding to 10,000 iterations following a burn-in period also of 10,000 iterations. The convergence of MCMC chain is assessed using separated partial means test proposed by Geweke [10] and all indicate that the chains have converged. The values of hyperparameters in the prior distributions were specified to produce approximately vague priors. Thus, for the five elections and each number of components in the mixture we specified $a_j = (4, 5, 5, 7, 7)$, $c_j = (49, 49, 45, 10, 10)$ and $d_j = (7, 7, 5, 1, 1/2)$. Also, for all elections we set $b_j = 1/10$, and $\nu_j = 1$. The main variation chosen was in the hyperparameter for shape parameter of Weibull distribution as discussed in Sect. 3.2. Thus for elections in 2006 and 2010 smaller values of these hyperparameters were chosen in order to reflect the greater dispersion of the distribution of the data. The acceptance rate in the Metropolis-Hastings algorithm for sampling δ_j and η_j was controlled to lie within the interval 0.20–0.50 which is usually recommended in the MCMC literature. All results presented here were obtained by using the software `R` [17].

The best model among the models with 1, 2 and 3 Weibull components was selected considering twice the natural logarithm of the Bayes factor presented in table 19.1, which interpretation can be seen in [19]. The results in this table show that, for all election, models with two or three components are better than a model with one component. However, when comparing models with two or three components the results may vary. For the 1994 and 2010 elections two components in the mixture are sufficient to fit the distribution of the votes. On the other hand, for the 1998, 2002 and 2006 elections three components are needed in the mixture.

In addition, considering the posterior mean of the parameters for each model we estimate the graph of density for each model and compare with the histogram of votes as showed in Fig. 19.1. We can see that the best model, that is the one for which density estimate is closest to the data, coincides with the choice according to Bayes factor.

Table 19.2 provides posterior means and 95 % HPD credible intervals for the parameters in the model chosen according to Bayes factor for each year. The credible intervals were constructed using the package `MCMCpack` of [14].

Using these parameters we can give some interpretations to the results. For example in 1994 we identified two groups of cities in Sergipe. For the first group, formed by 38 cities and with weight 0.45, we found an expected percentage of votes of 14.4 (posterior mean) with variability of 2.9 (posterior standard deviation), whereas

Table 19.2 Posterior mean and HPD interval of parameters of the best Weibull mixture model chosen by Bayes factor evaluation. Data of voting percentage obtained by Workers' Party in presidential elections in the Sergipe State from year 1994 to 2010 was considered for fitting of the models

Election	M[a]	Posterior mean and HPD Interval (95 %)		
		$w(weight)$	$\delta(shape)$	$\eta(scale)$
1994	2	0.45(0.27, 0.61)	5.81 (4.42, 7.21)	15.57 (14.09, 17.04)
		0.55 (0.39, 0.73)	5.18 (4.00, 6.50)	28.45 (26.97, 31.00)
1998	3	0.47 (0.26, 0.67)	5.38 (3.99, 6.81)	13.91 (12.05, 16.35)
		0.28 (0.09, 0.47)	6.66 (4.72, 8.67)	22.56 (18.13, 29.50)
		0.25 (0.11, 0.42)	6.69 (4.86, 8.52)	34.49 (30.94, 37.92)
2002	3	0.28 (0.05, 0.69)	7.18 (4.70, 9.64)	21.01 (16.52, 29.36)
		0.42 (0.07, 0.66)	8.19 (5.63, 10.88)	31.65 (26.84, 41.88)
		0.30 (0.09, 0.50)	8.57 (6.12, 11.07)	44.01 (40.20, 47.45)
2006	3	0.36 (0.08, 0.72)	11.04 (6.29, 16.50)	39.64 3(5.77, 45.15)
		0.42 (0.06, 0.72)	9.65 (5.36, 14.97)	48.69 (43.40, 55.09)
		0.22 (0.02, 0.48)	8.72 (4.93, 13.65)	58.96 (50.89, 67.03)
2010	2	0.84 (0.63, 0.97)	10.10 (6.16, 12.80)	50.17 (48.39, 52.28)
		0.16 (0.03, 0.37)	18.12 (6.27, 29.01)	62.48 (51.91, 66.81)

[a] Number of components in the mixture

in the second group, formed by 37 cities and with weight 0.55, the corresponding values are higher, 26.2 and 5.8 respectively. Likewise, in 2010, two populations were also identified. For the first population, formed by 67 cities with weight 0.84, we found an expected percentage of votes of 47.8 (posterior mean) with variability of 5.7 (posterior standard deviation), whereas in the second group, formed by 8 cities and with weight 0.16, the corresponding values are higher, 60.7 and 4.1 respectively. Note the significant increment of the percent of votes in both populations between 1994 and 2010. In addition, the first population in 2010 has 33 of the cities in the first group in 1994 indicating specifically that this group of cities had a significant increment over time.

Finally, from the best model for each election, the probability that PT obtains more than 50 % of the votes in the first round was calculated, because if the presidential candidate won the majority of votes in the first ballot the candidate is declared winner of presidential election and the second ballot is not necessary. The probabilities were estimated considering the predictive distribution by numerical integration using the Simpson rule combined with Monte Carlo method as seen in Sect.19.2.6. These corresponding probabilities of winning in the first ballot for PT party considering the Sergipe state for elections in 1994, 1998, 2002, 2006 and 2010 were 3.15×10^{-6}, 9.76×10^{-5}, 0.0175, 0.273 and 0.459 respectively, thus indicating that this probability was increasing over time.

We should note that as suggested by a referee a Mixture Normal model was also implemented considering an algorithm similar to the one defined in Sect. 3.3

without Metropolis-Hasting step. The results showed, that there is a strong evidence in favour of the Weibull mixture model. Additionally as discussed in Sect. 3.1 this model can lead to inferences which can be misleading since the normal is a symmetric distribution and can lead to overfit when additional component need to be included to capture the asymmetry in the data.

19.4 Discussion and Further Development

This chapter proposed a Weibull mixture model to describe the electoral behaviour of a Brazilian political party in different elections. The number of votes obtained by Workers' Party in the five Brazilian presidential elections from 1994 to 2010 were considered for analysis. A fully Bayesian approach was undertaken using MCMC methods.

We note that the results shown in this chapter are purely descriptive. They illustrate how the votes of a particular political party in different elections in Brazil in a given geographic area may exhibit multimodality and how the distribution of votes changes over time. Also, we found that the probability of obtainig 50 % of the votes in the first ballot is increasing along time.

In future developments we consider extending the analysis for all states of Brazil as well as including other parties in the analysis. Also regression models explain the electoral conduct should be considered in future analysis. Since that votes are limited variables, that is votes is between a minimum and maximum value, models for limited distributions as beta distributions also can be explored.

Acknowledgements The first author was supported by CAPES, Brazil. We are grateful to editors and reviewers for valuable comments and suggestions.

References

1. Atkinson, K.: An Introduction to Numerical Analysis, 2nd edn. Wiley, New York (2008)
2. Bërdufi, D.: Statistical Detection of Vote Count Fraud. Albanian Parliamentary Election and Benford's Law Mediterranean Journal of Social Sciences MCSER Publishing, Rome-Italy, **5**, 755–771 (2014)
3. Berkhof, J., van Mechelen, I., Gelman, A.: A Bayesian approach to the selection and testing of mixture models. Stat. Sin. **13**, 423–442 (2003)
4. Bohn, S.R.: Social policy and vote in Brazil Bolsa Familia and the shifts in Lula's electoral base. Lat. Am. Res. Rev. **46**, 54–79 (2011)
5. Chen, M., Shao, Q., Ibrahim, J.: Monte Carlo methods in Bayesian computation. Springer-Verlag, New York (2000)
6. Chib, S., Jeliazkov, E.: Marginal likelihood from the Metropolis- Hastings output. J. Am. Stat. Assoc. **96**, 270–281 (2001)
7. Cuff, V., Lewis, A., Miller, S. J.: The Weibull distribution and Benford's Law. arXiv (2014) `http://arxiv.org/pdf/1402.5854.pdf`

8. Durtschi, C., Hillison, W., Pacini, C.: The effective use of Benford's Law to assist in detecting fraud in accounting data. J Forensic Account. **5**, 17–34. (2004)
9. Escobar, M.D., West, M.: Bayesian density estimation and inference using mixtures. J. Am. Stat. Assoc. **90**, 577–588 (1995)
10. Geweke, J.: Evaluating the accuracy of sampling-based approaches to calculating posterior moments. Bayesian Stat. **4**, 169–193 (1992)
11. Jasra, A., Holmes, C.C., Stephens, D.A.: Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modelling. Stat. Sci. **20**, 50–67 (2005)
12. Jones, K., Johnston, R.J., Pattie, C.J.: People, places and regions: exploring the use of multi-level modelling in the analysis of electoral data. Br. J. Polit. Sci. **22**, 343–380 (1992)
13. Marin, J.M., Mergensen, K., Robert, C.P. : Bayesian modelling and inference on mixtures of distributions. In
: Dey, D., Rao, C. R. (eds.) Handbook of Statistics, vol. 25, pp. 459–507. North-Holland, Amsterdam (2005)
14. Martin, A.D., Quinn, K.M., Park, J.: H.: MCMCpack: Markov chain Monte Carlo in R. J. Stat. Softw. **42**, 0–22 (2011)
15. Mclachlan, G., Peel, D.: Finite Mixture Models. Wiley Series in Probability and Statistics. Wiley-Interscience, United States of America (2000)
16. Meneguello, R.: Electoral behaviour in Brazil; the 1994 presidential elections. Inter. Soc. Sci. J., **47**(4) 627–641 (1995)
17. R Development Core Team: R A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0 (2008)
18. Richardson, S., Green, P.J.: On Bayesian analysis of mixtures with an unknown number of components. J. R. Stat. Soc., Ser. B **59**, 731–792 (1997)
19. Robert, E.K., Adrian, E.: R.: Bayes factors. J. Am. Stat. Assoc. **90**, 773–795 (1995)
20. Stephens, M.: Bayesian analysis of mixture models with an unknown number of components - an Alternative to Reversible Jump Methods. Ann. Stat. **28**, 40–74 (2000)
21. Robert, C.P., Casella, G.: Monte Carlo Statistical Methods, 2nd edn. Springer-Verlag, New York (2004)
22. Tsionas: Bayesian analysis of finiture of Weibull distributions. Communications in Statistics. Theory and. Methods. **31**, 37–48 (2002)

Chapter 20
An Alternative Operational Risk Methodology for Regulatory Capital Calculation

Guaraci Requena, Débora Delbem and Carlos Diniz

Abstract The main objective of this work is to suggest a new method for calculation of regulatory capital required for operational risk as an alternative to the corresponding version advocated by the Basel Committee of Banking Supervision. Our method takes into account genuine dependence among the losses of possible risk units within a financial institution. Our proposal reduces the amount of regulatory capital suggested by Basel Committee, where the risk units are assumed to be perfectly positive-dependent. A simulation study is performed to compare both approaches. Finally, we discuss when Bayesian methods are preferable to the classical ones.

20.1 Introduction

Operational risk, in general, is the risk of loss from an operational failure. This means that financial institutions accept that their employees, processes, and systems are imperfect and losses will arise from possible errors. Therefore, the companies have to be prepared to cover an amount of risk (keeping losses in reasonable tolerance) in pursuit of their objectives. Operational risk management differs from other types of risk (e.g., credit or market risks), because it is used to protect the company but not to generate a profit.

Initially, in the mid-1980s, the operational risk has been defined as a kind of undesirable incident/event, such a fraud or a system error. The operative question to the managers in the risk identification process was "What/where are your risks?", leading to the creation of a huge and unmanageable set of risks.

G. Requena (✉)
Institute of Mathematics and Statistics, University of São Paulo, São Paulo, Brazil
e-mail: guaraci@ime.usp.br

D. Delbem · C. Diniz
Department of Statistics, Federal University of São Carlos, São Carlos, Brazil
e-mail: deboradelbemg@gmail.com

C. Diniz
e-mail: dcad@ufscar.br

A. Polpo et al. (eds.), *Interdisciplinary Bayesian Statistics,*
Springer Proceedings in Mathematics & Statistics 118, DOI 10.1007/978-3-319-12454-4_20

Operational risk became recognized as a major risk class in the mid-1990s following a number of large-scale insolvencies in the banking industry caused by events outside of market and credit risk. Examples are Orange County, 1994; Barings Bank, 1995; and Daiwa Bank, 1995; among others. In these cases, significant losses were incurred due to operational risk failures. In response, the Basel Committee on Banking Supervision released a proposal in June 1999 to replace the 1988 Basel Capital Accord (Basel I), with a new risk-sensitive framework. The initial consultative proposal introduced an operational risk category as a measure of exposure to loss from undesirable nonoverlapping incidents/events (risk classes) and established the corresponding capital requirements.

In the revision of the Basel II regulations (2006), operational risk is defined as "the risk of loss resulting from inadequate or failed internal process, people or systems, or from external events," see [3]. This definition includes the legal risk (exposure to fines, penalties) but excludes strategic and reputation risk. In addition, seven basic event-type categories have been specified as internal fraud; external fraud; employment practices and workplace safety; clients, products, and business practice; damage to physical assets; business disruption and systems failures; execution, delivery, and process management.

The Basel Committee recognizes that operational risk is a term that has a variety of meanings and therefore, for internal purposes, financial institutions are permitted to adopt their own definitions of operational risk, provided that the minimum elements in the Committee's definition are included.

One of the main innovations of the Basel II agreement compared to Basel I has been not only to require allocation of capital to cover operational risk but also to advocate for an operational risk management system. An overview on operational risk measurement techniques is given in [2]. Basel II offers banks three capital calculation methods of increasing complexity.

- The basic indicator approach consists of applying a fixed percentage (15 %) to the average of the positive annual gross income of the financial institution over the previous 3 years;
- The standardized approach (SA) allows to apply a factor (between 12 and 18 %) that depends on the business line. The reason is that some financial activities are more exposed than others to operational risk (at least in relation to gross income). Concretely, the factors for eight basic business lines are: corporate finance 18 %, trading and sales 18 %, retail banking 12 %, commercial banking 15 %, payment and settlement 18 %, agency services 15 %, asset management 12 %, and retail brokerage 12 %. In order to be eligible, this method requires to have figures of losses incurred by each combination of 8 business lines and 7 event types due to operational risks (56 in total).
- Finally, the advanced measurement approach (AMA) allows the bank to build its own method for assessing operational risk. The Basel II Accord allows three alternative approaches: the loss distribution approach (LDA), the scenario-based approach, and the scorecard approach. The three approaches differ only in the emphasis on the information used to calculate regulatory capital. The chosen method as well as the implementation conditions (existence of a centralized risk

control structure, frequency, and relevance of reporting) are then submitted for prior approval to the regulator. In order to be eligible, the AMA method requires the following data to be available: internal loss data (specific to the bank); external loss data (available databases for the whole profession); analysis of potential event scenarios; and business environment and internal control factors.

Basel II Accord implementation, in Brazil, has focused on the SAs to credit, market, and operational risk, see [4]. Brazilian banks use a version of the SA (by reducing several factors of the basic business lines), called the alternative standardized approach (ASA), to calculate capital requirements for operational risk. Data provided by the Brazilian Central Bank (BCB) indicate that if the banks were to apply SA, following Basel II, their capital requirements would, on aggregate, double. The BCB explained that the main reason for making only the ASA available to Brazilian banks relates to anomalous Brazilian interest rate spreads, which make the ASA's asset-based indicator a much better proxy for operational risk in the Brazilian environment than the gross income under the SA.

The method chosen for identification and measurement of operational risk must be consistent within a banking group. The selection of the AMA allows to reduce capital reserves, which is the main objective of this work. We formalize the problem in Sect. 20.2 and propose a new method for reducing the regulatory capital advocated by Basel II. In absence of real data, we provide a simulation study and compare our method with Basel II requirements in Sect. 20.3. We finish with a discussion indicating Bayesian methods as a possible tool for analysis.

20.2 Description of Methods

Operational risk is difficult to identify and assess as the causes are extremely heterogeneous, thus making developing statistical models for operational risk challenging. The most typical example of statistical methods is the "LDA" associated with AMA. It relies on a database of losses collected within the bank, enhanced with data from external sources. The aim is to obtain the distribution of cumulative loss for each business line and each type of loss event (56 risk units in total according to Basel II Accord) and to use it as a base to calculate the regulatory capital.

We will first describe the methodology for calculation of operational risk fixed by Basel II. In the sequel we will propose our alternative.

20.2.1 Basel II Approach

For each risk unit in a financial institution there is a random variable X associated, named "cumulative operational loss," which represents the aggregated loss during 1 year. Following LDA, the cumulative operational loss is defined by $X = \sum_{i=0}^{N} S_i$, where $S_0 = 0$. Its distribution is given by

$$F(x) = P(X \leq x) = \sum_{n=0}^{\infty} P(X \leq x, N = n) = \sum_{n=0}^{\infty} P\left(\sum_{k=0}^{n} S_k \leq x\right) Pr(N = n). \quad (20.1)$$

The relation (20.1) is based on two assumptions: first, severities $S_1, S_2, \ldots$ are independent and identically distributed random variables; and second, observed loss frequency N and $(S_1, S_2, \ldots)$ are independent.

Therefore, the first step of the LDA is to draw for each of 56 risk units, two probability distributions for associated loss. The first represents the frequency of loss events over a time interval (loss frequency distribution), and the other represents the severity of these same events (loss severity distribution). To do so, one sorts loss events by frequency on one hand, and by cost on the other hand, and represents the result graphically (using histograms). For both distributions, the statistician estimates the corresponding parameters that best represent the shape of the curve. In order to validate the choice, one compares the result (frequency or loss) predicted by the corresponding distribution with the output of the curve built from real data: if both curves overlap, the model is considered reliable.

The analytical form of aggregated distribution given by (20.1) is difficult to get, even under independence assumptions. Hence, using a Monte Carlo simulation, selected frequency and severity distributions are combined in order to obtain for each business line and each type of event, an aggregated curve of the loss distribution X for a given time horizon. The LDA is described in [1].

Embrechts and Puccetti [5] argue that the most sensitive methodology for the operational risk is the LDA. The independence assumption between severities $S_1, S_2, \ldots$ is unreasonable because of the common economic influence into a given risk unit. A related discussion can be found in [6] and [7].

The value-at-risk (VaR) of a random variable X with distribution function at level $\alpha \in (0, 1)$ is defined by

$$VaR_\alpha(X) = F^{-1}(\alpha) = \inf\{x | F(x) \geq \alpha\},$$

where F^{-1} is the inverse of distribution function $F(x)$. According to Basel II, the operational VaR for a fixed risk unit (to be denoted by $VaR(X)$), is the 99.9 % quantile of distribution of the cumulative operational loss X, i.e.,

$$VaR(X) = F^{-1}(0.999) = \inf\{x | F(x) \geq 0.999\}.$$

Thus, $VaR(X)$ is a monetary value (the maximum loss incurred with a probability of 99.9 %) that the finance institution needs in order to assign the so-called "marginal unexpected loss" UL_X in the risk unit with loss X. It is given by

$$UL_X = VaR(X) - E(X), \quad (20.2)$$

where $E(X)$ is the expected value of X.

In general, the regulatory capital is defined as the economic capital which have to cover the "total unexpected loss" (TUL) of the finance institution. Following Basel

II Accord, the TUL is calculated as a summation of marginal unexpected losses for all risk units, i.e.,

$$TUL = \sum_{k=1}^{M} UL_{X_k}, \tag{20.3}$$

where UL_{X_k} is given by (20.2) and M denotes the number of risk units (56 in total).

Let $(X_1, \ldots, X_M)$ be a random vector with joint distribution $H(x_1, \ldots, x_M)$, $M \geq 2$. Its upper and lower Fréchet–Hoeffding bounds are given by

$$\max\left\{F_{X_1}(x_1) + \cdots + F_{X_M}(x_M) - M + 1, 0\right\} \leq H(x_1, \ldots, x_M) \leq \min\left\{F_{X_1}(x_1), \ldots, F_{X_M}(x_M)\right\}$$

where F_{X_i} is the distribution function of X_i, $i = 1, \ldots, M$. If the upper bound is attained we say that marginal random variables are comonotonic, i.e, perfectly positive-dependent. The variables are said countermonotonic (perfectly negative-dependent) when the lower bound in last relation is reached.

A careful analysis of relation (20.3) indicates that it is fulfilled only if

$$VaR(X_1 + \cdots + X_M) = VaR(X_1) + \cdots + VaR(X_M).$$

This means that the TUL advocated implicitly assumes that all risk units are comonotonic, i.e., perfectly positive-dependent. In other words, the losses caused by possible combinations of business lines and event types would occur simultaneously in same time during the holding period. Furthermore, (20.3) implies that all losses are driven from the same randomness, which hardly occurs in real situations. In fact, these conclusions are a consequence of the following general result, see Proposition 6.15 in [9].

Proposition *Let $\psi : R^M \to R$ be increasing and left continuous in each argument function and $X_1, \ldots, X_M$ be comonotonic random variables. Then*

$$VaR_\alpha(\psi(X_1, \ldots, X_M)) = \psi[VaR_\alpha(X_1), \ldots, VaR_\alpha(X_M)]$$

for $\alpha \in (0, 1)$.

In our case the function $\psi(x_1, \ldots, x_M) = x_1 + \cdots + x_M$. The interpretation is that VaR calculations can be transported directly through increasing continuous functions of possible comonotonic risk units.

20.2.2 An Alternative Methodology

The identification and the measurement of operational risk is a new issue in the banking sector. As we noted, the Basel II proposal for calculation of TUL by (20.3) is based on an unrealistic assumption of strongest positive dependence structure between all risk units. Thus, the regulatory capital is overestimated believing that all marginal unexpected losses occur jointly. Giacometti et al. [8] argue that a financial institution can effectively reduce its charge of regulatory capital by taking into account the dependence structure among the risk units. In the same sense, Frachot et al. [7] state that the Basel II method is based on contradictory assumptions.

These facts motivate to suggest an alternative methodology for regulatory capital calculation. We will assume a dependence between risk units which is weaker than the perfect positive one. The aim is to recommend aggregate unexpected loss (AUL), which is (20.3).

For simplicity, we will illustrate our approach considering two risk units. Let the cumulative operational losses X and Y be continuous with distributions $F(x)$ and $G(y)$, respectively, and $H(x, y)$ be their continuous joint distribution. The corresponding upper and lower Fréchet–Hoeffding bounds are denoted by

$$H^+(x, y) = \min\{F(x), G(y)\} \quad \text{and} \quad H^-(x, y) = \max\{F(x) + G(y) - 1, 0\}.$$

Since the operational risk have to cover the marginal unexpected losses, we define the probability

$$p = P(E(X) \leq X \leq VaR(X), E(Y) \leq Y \leq VaR(Y)). \tag{20.4}$$

Our proposal is to determine the AUL by

$$AUL = f_k(p) = ap^{1/k} + b, \quad k > 0 \tag{20.5}$$

where the family $f_k(p)$ of power functions depends of positive parameters a and b. We will find the unknown parameters k, a, and b and the domain of p under the following reasonable assumptions:

(A1) Take into account the expert opinion;
(A2) AUL should be nondecreasing in k and p;
(A3) AUL is represented by (20.3) if $H(x, y) = H^+(x, y)$;
(A4) The case of perfect negative dependence to be incorporated in AUL.

The performance of the function $f_k(p)$ depends of the choice of parameter k which can be associated to expert opinion (recommended by Basel II). Adopting a local experience, we will consider a particular bijection $k = tg(\frac{\pi}{2}c)$. If $k \to \infty$, then $c \to 1$ and therefore $c \in [0, 1]$. If the current economic situation is critical, the expert assigns values of c close to 1. In this case of financial institution should apply TUL calculated by (20.3).

The probability p in (20.4) can be equivalently represented by

$$p = H(VaR(X), VaR(Y)) - H(VaR(X), E(Y)) - H(E(Y), VaR(Y)) + H(E(X), E(Y)).$$

Note that when the argument p increases the regulatory capital also must increase and assumption (A2) is satisfied. In the comonotonic case we obtain

$$\begin{aligned} p^+ = \min\{F(VaR(X)), G(VaR(Y))\} - \min\{F(VaR(X)), G(E(Y))\} - \\ - \min\{F(E(X)), G(VaR(Y))\} + \min\{F(E(X)), G(E(Y))\}. \end{aligned}$$

Since $F(VaR(X)) = G(VaR(Y)) = 0.999$, we get the upper bound of p given by

$$p^+ = 0.999 - \max\{F(E(X)), G(E(Y))\}.$$

If at least one of the events $\{X \in [E(X), VaR(X)]\}$ and $\{Y \in [E(Y), VaR(Y)]\}$ do not occur, we have $p = 0$. It is direct to check that such a case is possible only if

Table 20.1 Terms for AUL calculation

$E(X)$	1.13
$VaR(X)$	4.53
$E(Y)$	1.13
$VaR(Y)$	4.69
$min(UL_X, UL_Y)$	3.4
$max(UL_X, UL_Y)$	3.56
p^+	0.4

the lower bound $H^-(x, y)$ is attained, i.e., both risks X and Y are perfectly negative dependent. Substituting $p = 0$ in (20.5) one gets $b = \max\{UL_X, UL_Y\}$, where the unexpected losses of X and Y are computed by (20.2).

Finally, assumption (A3) is fulfilled if

$$a = \frac{(UL_X + UL_Y) - \max\{UL_X, UL_Y\}}{(p^+)^{1/tg(\frac{\pi}{2}c)}}.$$

Thus, we arrive to our proposal.

Theorem *Under assumptions (A1)–(A4), the aggregated unexpected loss is*

$$AUL = \min\{UL_X, UL_Y\}\left(\frac{p}{p^+}\right)^{1/tg(\frac{\pi}{2}c)} + \max\{UL_X, UL_Y\}, \tag{20.6}$$

where $c \in [0, 1]$ *and* $p \in [0, p^+]$.

It is direct to verify that suggested aggregated unexpected loss (20.6) is inferior than the regulatory capital advocated by Basel II through (20.3). In the worst case scenario (perfect positive dependence between X and Y or $c \to 1$) we have $AUL = TUL$.

20.3 Comparison of the Methods

Here we show first the behavior AUL defined by (20.6) for fixed marginal distributions. We assume that the distributions of X and Y already selected and their parameters are estimated using the LDA. Let

$$X \sim Weibull(1.5, 1.25)$$
$$Y \sim Lognormal(0, 0.5).$$

The necessary quantities in (20.6) are summarized in Table 20.1.

Therefore, for the considered example we obtain

$$TUL = UL_X + UL_Y = 6.96 \quad \text{and} \quad AUL = 3.4\left(\frac{p}{0.4}\right)^{1/tg(\frac{\pi}{2}c)} + 3.56,$$

where $p \in [0, 0.4]$ and $c \in [0, 1]$.

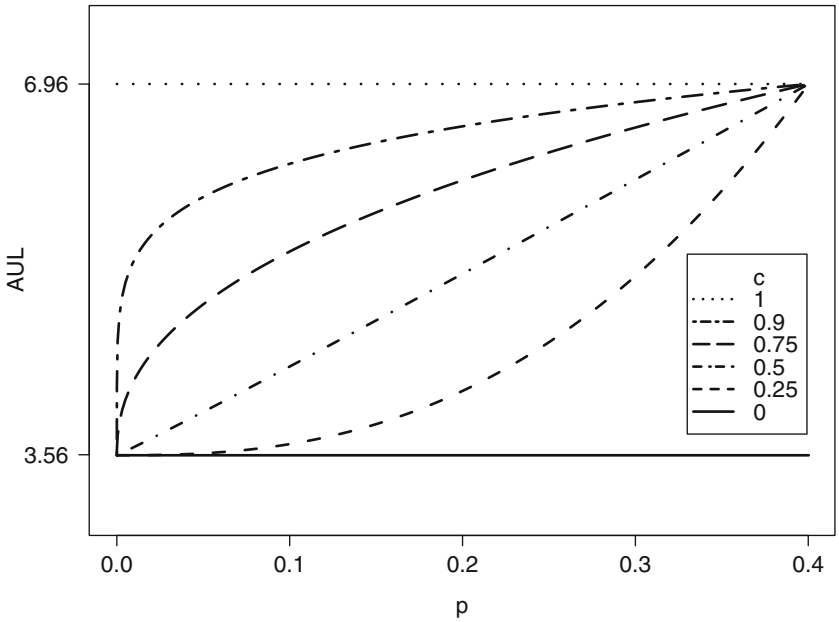

Fig. 20.1 Behavior of proposed AUL

Figure 20.1 shows the performance of AUL for admissible values of p and c. It can be seen that the corresponding AUL curves are below the extreme TUL value 6.96.

Now, let us assume that our risk X and Y are dependent. Their continuous joint distribution $H(x, y)$ can be represented by some copula C as $H(x, y) = C(F(x), G(y))$.

To obtain the joint distribution $H(x, y)$ of X and Y with fixed above marginals, we will use the Gaussian copula with parameter $\rho \in [-1, 1]$. It is given by

$$C_\rho(u, v) = \int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(v)} \phi_{2,\rho}(x, y) dx dy,$$

where $\phi_{2,\rho}(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[x^2 + y^2 - 2\rho xy\right]\right)$, $\rho \in [-1, 1]$ and $a = \Phi^{-1}(u)$ is such that

$$u = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} \exp\left(-\frac{x^2}{2}\right) dx.$$

In fact, we need to calculate the probability p given by (20.4) being a function of $H(x, y)$. The results comparing TUL and AUL for various values of parameters ρ and c are presented in Table 20.2. One can see that AUL values are nondecreasing in each line and column along with increase of c and ρ, respectively. As should be, any calculated AUL is smaller than the corresponding TUL value 6.96.

Table 20.2 AUL-values

ρ	$c = 0.1$	$c = 0.5$	$c = 0.7$	$c = 0.9$
−0.9	3.611271	3.737912	4.241459	5.599157
−0.3	3.642597	4.697959	5.532541	6.510147
0	3.680031	5.040993	5.832651	6.648987
0.3	3.744461	5.560509	6.248796	6.891505
1	6.96	6.96	6.96	6.96

20.4 Discussion and Conclusions

Dealing with operational risk measures is a relatively new field of research. Just after the Basel II has permitted a substantial degree of flexibility within the advanced models in 2004, the interest in calculating the corresponding capital charge has increased. The AMA model is the most risk-sensitive one and leaves the possibility for the bank to develop its own procedures for measuring and assessing the exposure to operational risk. Dispute this freedom, the usage of the AMA is subject of supervisory approval.

The BCB has implemented regulations to permit the use of banks' internal models for regulatory capital calculation. Final rules for the AMA were issued by BCB in March 2013, see [4]. Several Brazilian banks are developing their own AMA models, but remain at relatively early stages of planning. So, it is unlikely that any bank will be approved by BCB to use AMA within the next 3-year period.

Our proposal for regulatory capital calculation through Eq. (20.6) is actual and based on possible dependence between risk units. It reduces the corresponding amount suggested by Basel II, which is reasonable during financial crisis (when risk units are perfectly positive dependent or the parameter c is evaluated by an expert with a value close to 1).

As we noted in Sect. 20.1, the AMA allows to consider the financial institution internal data along with data that could be from an external consortium, or data in the form of risk scores based on opinions from industry experts or the owner of the risk. The internal databases are more objective while the external ones are purely subjective in general. This fact implicitly indicates the use of Bayesian methods for estimation of unknown marginal and copula parameters and corresponding model validation. For example, the external data are used to estimate a "prior density" for the parameter of interest, and the internal data are used to estimate another density for the parameter that is called the "sample likelihood," see [2]. These two densities are then multiplied to obtain the "posterior density" of the corresponding parameter.

For more complex marginal models and/or copula function one may have problems to maximize the likelihood using the conventional statistical techniques. An alternative is to apply the Bayesian methods, which can be more efficient. Smith [10] discusses this option in some detail. For instance, the Bayesian estimation can be extended for nonlinear dependence structures, hierarchical models can be employed for the marginals, etc.

Acknowledgments The authors would like to thank Nikolai Kolev, IME-USP, for numerous helpful comments on earlier drafts.

References

1. Alexander, C.: Operational risk: Regulation, Analysis and Management. Pearson Education, London (2003)
2. Alexander, C.: Statistical models of operational loss. In: Fabozzi, F.G. (ed.) Handbook of Finance, Vol. 1, pp. 129–171 (2008)
3. Basel Committee of Banking Supervision. International convergence of capital measurements and capital standards: a revised framework—comprehensive version (2006). http://www.bis.org/publ/bcbs128.pdf. Accessed 29 Dec. 2014
4. Basel Committee of Banking Supervision. RCAP assessment of Basel regulations in Brazil (December 2013). http://www.bcb.gov.br/pec/appron/apres/RCAP_Brazil_assessment_report.pdf. Accessed 29 Dec. 2014
5. Embrechts, P., Puccetti, G.: Aggregating risk capital, with an application to operational risk. The Geneva Risk and Insurance Review, **31**(2) (2006)
6. Frachot, A., Georges, P., Roncalli, T.: Loss distribution approach for operational risk. Working paper, GRO, Credit Lyonnais (2001). www.thierry-roncalli.com. Accessed 29 Dec. 2014
7. Frachot, A., Roncalli, T., Salomon, E.: The correlation problem in operational risk. Preprint, Credit Agricole (2004)
8. Giacometti, R., Rachev, S., Chernobai, A., Bertocchi, M.: Aggregation issues in operational risk. The Journal of Operational Risk (2008)
9. McNeil, A., Frey, L., Embrechts, P.: Quantitative Risk Management. Princeton University Press, Princeton (2005)
10. Smith, M. S.: Bayesian approaches to copula modelling. eprint arXiv:1112.4204 (2011). http://arxiv.org/pdf/1112.4204v1.pdf. Accessed 29 Dec. 2014

Chapter 21
Bayesian Approach of the Exponential Poisson Logarithmic Model

José Augusto Fioruci, Bao Yiqi, Francisco Louzada and Vicente G. Cancho

Abstract Recently, a new three-parameter lifetime distribution motivated mainly by lifetime issues has been proposed by the authors. In this chapter, we consider the Bayesian analysis for this new distribution and compare its performance with the classic ones. The approximate Bayes estimators obtained by Markov chain Monte Carlo (MCMC) methods under the assumption of noninformative priors are compared with the maximum likelihood estimators by simulation. Finally, the model is fitted to a real data set and it is compared with several models.

21.1 Introduction

Recently several new distributions were proposed under the structure of primary latent causes, such as [7–9]. Fioruci et al. [6] proposed the exponential Poisson logarithmic (EPL) distribution, which is a three-parameter distribution and was constructed under the structure of secondary latent cause activation, i.e, the event of interest will occur when any primary cause is activated, since each primary cause is activated when all of secondary latent causes are activated. The EPL distribution generalizes the exponential Poisson distribution proposed by Cancho et al. [4], and

J. A. Fioruci (✉) · Y. Bao
Department of Statistics, Federal University of São Carlos—UFSCar, Rodovia Washington Luiz, km 235, São Carlos, SP 13565-905, Brazil
e-mail: jafioruci@gmail.com

Y. Bao
e-mail: baoyiqi@gmail.com

F. Louzada · V. G. Cancho
Institute of Mathematics and Computer Science, University of São Paulo—USP, Avenida Trabalhador São-carlense, 400 - Centro, São Carlos, SP 13566-590, Brazil
e-mail: louzada@icmc.usp.br

V. G. Cancho
e-mail: garibay@icmc.usp.br

A. Polpo et al. (eds.), *Interdisciplinary Bayesian Statistics,*
Springer Proceedings in Mathematics & Statistics 118, DOI 10.1007/978-3-319-12454-4_21

its probability density function (*pdf*) is given by

$$f(y) = \frac{\phi\,\theta\,\lambda \exp(-\lambda\, y - \theta\, e^{-\lambda y})}{-\log(1-\phi)\{1 - e^{-\theta} - \phi\,[1 - \exp(-\theta\, e^{-\lambda y})]\}}, \quad y > 0, \tag{21.1}$$

where $\phi \in (0,1)$, $\lambda > 0$ and $\theta > 0$. The parameter λ controls the scale of the distribution and the parameters ϕ and θ control its shape. The corresponding hazard rate function has the expression

$$h(y) = \frac{\phi\theta\lambda \exp(-\lambda y - \theta e^{-\lambda y})}{-\log\left[1 - \phi\frac{1-\exp(-\theta e^{-\lambda y})}{1-e^{-\theta}}\right]\left\{1 - e^{-\theta} - \phi[1 - \exp(-\theta e^{-\lambda y})]\right\}}, \tag{21.2}$$

for $y > 0$, $0 < \phi < 1$, $\theta > 0$, $\lambda > 0$ and it decreases for $\phi > 1 - e^{-\theta}$ and increases for $\phi < 1 - e^{-\theta}$.

The main aim of this chapter is to consider the Bayesian analysis for the EPL distribution and compare its performance with the classical ones. Since the parameters ϕ, λ, and θ have different ranges, we assume the normal truncated priors on them. Although they are not the conjugate priors, the prior information is easy to control through their parameters. In many practical situations, the information about the shape and scale of sampling distribution is available in an independent manner, see [3]. Therefore, here it is assumed that the priors of the parameters ϕ, λ, and θ are independent.

In this chapter we consider the squared error loss function. It is observed that the Bayes estimators cannot be expressed in explicit form, so we compute the approximate Bayes estimators by Markov chain Monte Carlo (MCMC) simulation methods.

The chapter is organized as follows. In Sect. 21.2, we study the inferential procedure based on the Bayesian approach. We conducted a simulation study in order to assess the performance of the Bayesian estimator and compare with the classical estimator. In Sect. 21.4 the application of the new distribution is illustrated considering a real data set, where it is compared with its submodels and also with another models with three parameters. Some final comments in Sect. 21.5 conclude the chapter.

21.2 Bayesian Inference

Let $\mathbf{y} = (y_1, \ldots, y_n)$ be a random sample of the EPL distribution with unknown parameter vector $\boldsymbol{\xi} = (\theta, \lambda, \phi)$. The likelihood function $L(\boldsymbol{\xi}|\mathbf{y})$ is given by

$$L(\boldsymbol{\xi}|\mathcal{D}) = \prod_{i=1}^{n} \frac{\phi\,\theta\,\lambda \exp(-\lambda\, y_i - \theta\, e^{-\lambda y_i})}{-\log(1-\phi)\{1 - e^{-\theta} - \phi\,[1 - \exp(-\theta\, e^{-\lambda y_i})]\}} \tag{21.3}$$

where $\mathcal{D}$ denotes the observed data.

For a Bayesian analysis, we assume the follow prior densities for parameters ϕ, λ, and θ:

- $\phi \sim N(\mu_\phi, \sigma_\phi^2)I(0,1)$, μ_ϕ and σ_ϕ known;
- $\lambda \sim N(\mu_\lambda, \sigma_\lambda^2)I(0,\infty)$, μ_λ and σ_λ known;
- $\theta \sim N(\mu_\theta, \sigma_\theta^2)I(0,\infty)$, μ_θ and σ_θ known;

where $N(\mu, \sigma^2)\mathbf{I}(a,b)$ denote the truncated normal distribution which is the probability distribution of a normally distributed random variable whose value lies within the interval $-\infty \leq a < b \leq \infty$. In several areas, especially in medicine, it is preferable to use the prior information when they are available; moreover, it is worth mentioning that using a truncated normal distribution as prior facilitates the insertion of information in certain regions of the parameter space, since the hyperparameters no longer represent the mean and variance but still control the region of higher probability mass.

Assuming the independence of the parameters, the prior densities for $\boldsymbol{\xi}$ can be written as

$$\pi(\boldsymbol{\xi}) = \pi(\phi)\pi(\lambda)\pi(\theta) \tag{21.4}$$

and for express vague information priors, we consider $\mu_\phi = \mu_\lambda = \mu_\theta = 0$ and $\sigma_\phi^2 = \sigma_\lambda^2 = \sigma_\theta^2 = 100$.

Combining the likelihood function (21.3) and the prior distribution in (21.4), the joint posterior distribution for $\boldsymbol{\xi}$ is obtained as

$$\pi_\xi(\boldsymbol{\xi}|\mathcal{D}) \propto L(\boldsymbol{\xi};\mathcal{D})\pi(\phi)\pi(\lambda)\pi(\theta).$$

This joint posterior density is analytically intractable. So we based our inference on the MCMC method. Moreover, we observed that there is no closed form available for any of the full conditional distributions. Thus, we choose the Metropolis–Hastings algorithm to construct the Markov chain.

To use the random-walk chain with normal distribution as proposal density we will adopt the reparameterization $\boldsymbol{\varphi} = (\varphi_1, \varphi_2, \varphi_3) = \left(\log\left(\frac{\phi}{1-\phi}\right), \log(\lambda), \log(\theta)\right)$, thus the original parameter space is transformed to $\mathbb{R}^3$ space. So the resulting joint posterior density is given by

$$\pi(\boldsymbol{\varphi}|\mathcal{D}) = \pi_\xi(\boldsymbol{\varphi}^{-1}|\mathcal{D}) \times \frac{\exp(\varphi_1 + \varphi_2 + \varphi_3)}{(1+\exp(\varphi_1))^2}.$$

To implement the Metropolis–Hastings algorithm, we proceed as follows:

(1) Start with any point $\boldsymbol{\varphi}_{(0)}$, and stage indicator $j = 0$;
(2) Generate a point $\boldsymbol{\varphi}'$ from the transition kernel distribution $N_3\left(\boldsymbol{\varphi}_j, \Sigma\right)$, where Σ is the covariance matrix of $\boldsymbol{\varphi}$;
(3) Update $\boldsymbol{\varphi}_{(j)}$ to $\boldsymbol{\varphi}'$ with probability $\min\{1, \pi(\boldsymbol{\varphi}'|\mathcal{D})/\pi(\boldsymbol{\varphi}_{(j)}|\mathcal{D})\}$ and make $j = j+1$;
(4) Repeat steps (2) and (3) until the process reaches a stationary distribution.

In practice, it is difficult to obtain analytically the covariance matrix Σ and a numerical approximation for it needs to be used.

21.3 Simulation Study

In this section, we conducted a simulation study to verify and compare the performances of the Bayes estimators to the maximum likelihood estimators obtained by Newton–Raphson method (NR).

A lifetime data was simulated from the quantile function of the distribution (via inversion method) with the parameters $\phi = 0.5$, $\lambda = 1.0$, and $\theta = 2.0$. We took the sample sizes $n = 20, 40, \ldots, 200$ and conducted 1000 replicates for each sample size. The vague information to prior distributions of the parameters was considered in the study. For each generated data set we simulated one chain of size 30,000 for each parameter, and the first 10,000 iterations were disregarded in order to eliminate the effect of the initial values and avoid correlation problems, thus obtaining an effective sample of size 20,000 upon which the posterior is based.

Table 21.1 shows the posterior mean, mean posterior variance, biases, and mean squared errors (MSE) for some sample sizes in the top table and the averages of the maximum likelihood estimates, variances of estimators, biases, and MSEs for some sample sizes in the bottom table.

In the Bayesian estimator simulation study, the following observations were made: the estimators of θ and λ approached the real values of the parameters as n increased, so that their biases and MSEs went to zero as n increased;the estimator of ϕ presented a negative bias, however its biases and MSEs are always near zero. In the classical estimator simulation study the following observations were made: the estimator of λ approached the real values of the parameters as n increased; the estimator of θ and ϕ present a positive and negative bias, respectively, although its biases are always near zero. The MSEs of θ and ϕ kept stable as n increased.

The graphics of biases and MSEs are presented in Figs. 21.1 and 21.2 for Bayesian and classical simulations, respectively. Comparing the results, we can conclude that the Bayesian estimators have better performance than the maximum likelihood estimators in general; indeed the maximum likelihood estimators of ϕ and θ present little biases.

21.4 Applications

In this section, we analyze a data set with 113 observed lifetimes of patients on the waiting list for the Stanford heart transplant program as presented by Escobar [5].

In order to identify the shape of a lifetime data failure rate function we shall consider a graphical method based on the TTT plot [1]. In its empirical version the TTT plot is given by $G(r/n) = \left[\left(\sum_{i=1}^{r} Y_{i:n}\right) + (n-r)Y_{r:n}\right] / \left(\sum_{i=1}^{r} Y_{i:n}\right)$ where

Table 21.1 The averages of estimations, variances of estimators, biases, and MSEs

Bayesian estimator simulation study												
	Bayes estimatives			Var			Bias			MSE		
n	$\hat{\phi}$	$\hat{\lambda}$	$\hat{\theta}$	$\hat{\phi}$	$\hat{\lambda}$	$\hat{\theta}$	$\hat{\phi}$	$\hat{\lambda}$	$\hat{\theta}$	$\hat{\phi}$	$\hat{\lambda}$	$\hat{\theta}$
20	0.498	1.181	2.932	0.080	0.090	2.128	−0.002	0.181	0.932	0.002	0.121	1.801
40	0.487	1.085	2.595	0.078	0.044	1.445	−0.013	0.085	0.595	0.002	0.042	0.909
60	0.487	1.051	2.449	0.078	0.029	1.195	−0.013	0.051	0.449	0.003	0.036	0.765
80	0.489	1.044	2.363	0.078	0.023	1.146	−0.011	0.044	0.363	0.003	0.024	0.473
100	0.482	1.054	2.349	0.077	0.019	1.028	−0.018	0.054	0.349	0.004	0.018	0.417
150	0.473	1.010	2.251	0.076	0.013	0.894	−0.027	0.010	0.251	0.004	0.011	0.309
200	0.471	0.990	2.184	0.075	0.009	0.844	−0.029	−0.010	0.184	0.006	0.009	0.221

Classical estimator simulation study												
	MLE			Var			Bias			MSE		
n	$\hat{\phi}$	$\hat{\lambda}$	$\hat{\theta}$	$\hat{\phi}$	$\hat{\lambda}$	$\hat{\theta}$	$\hat{\phi}$	$\hat{\lambda}$	$\hat{\theta}$	$\hat{\phi}$	$\hat{\lambda}$	$\hat{\theta}$
20	0.427	0.953	1.767	1.020	0.441	14.506	−0.073	−0.047	−0.233	0.146	0.067	1.375
40	0.387	0.965	1.806	0.610	0.577	20.626	−0.113	−0.035	−0.194	0.146	0.036	1.044
60	0.388	0.973	1.886	0.625	0.168	9.364	−0.112	−0.027	−0.114	0.148	0.024	0.906
80	0.405	1.008	2.072	1.247	0.420	24.536	−0.095	0.008	0.072	0.146	0.018	0.831
100	0.393	0.995	2.050	0.252	0.067	3.267	−0.107	−0.005	0.050	0.132	0.014	0.759
150	0.464	0.983	2.233	0.255	0.053	3.091	−0.036	−0.017	0.233	0.139	0.011	0.867
200	0.405	1.009	2.133	0.259	0.052	2.848	−0.095	0.009	0.133	0.133	0.011	0.662

MSE mean squared error, *MLE* maximum likelihood

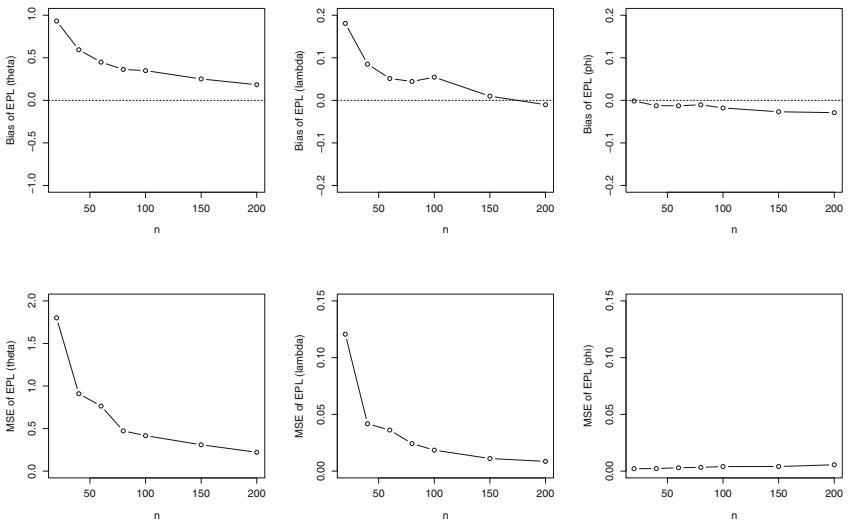

Fig. 21.1 Bias (*left panels*) and mean squared errors (MSE) (*right panels*) of Bayesian estimation

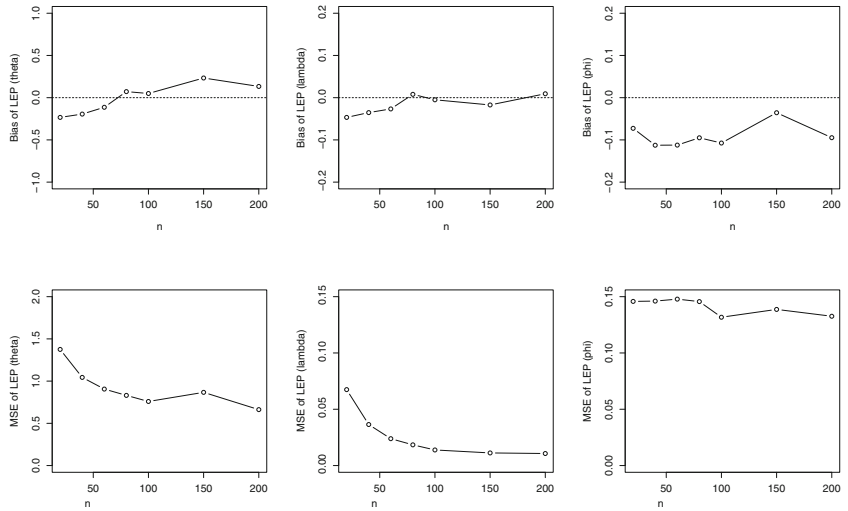

Fig. 21.2 Bias (*left panels*) and mean squared errors (MSE) (*right panels*) of the maximum likelihood estimates

$r = 1, \ldots, n$ and $Y_{i:n}$ represent the order statistics of the sample. It has been shown that the failure rate function is increasing (decreasing) if the TTT plot is concave (convex). Although, the TTT plot is only a sufficient condition, not a necessary one for indicating the failure rate function shape, it is used here as a crude indicator of its shape. Fig. 21.3 shows the TTT plot for the considered data set, which is

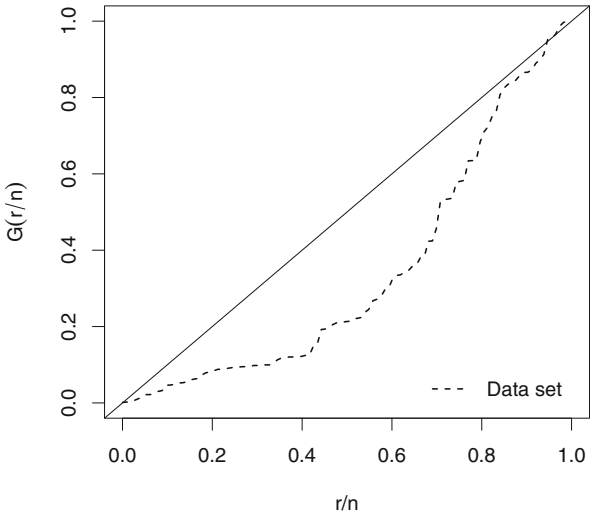

Fig. 21.3 Empirical-scaled TTT-transform for the data sets

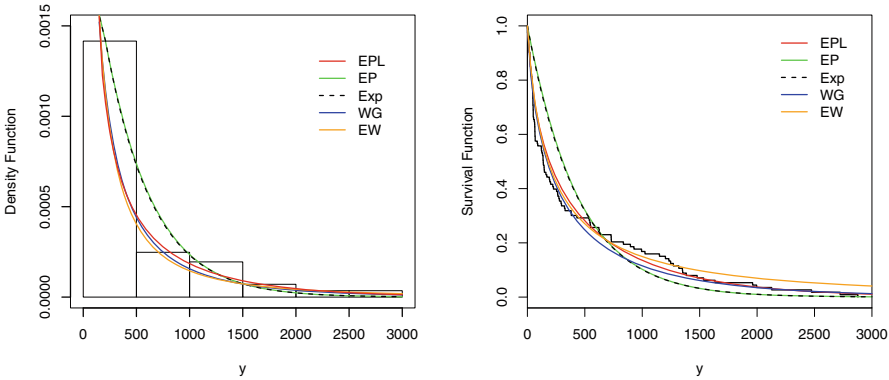

Fig. 21.4 Data set 1. *Left panel*: The density functions of the fitted models superimposed on the histogram. *Right panel*: Kaplan–Meier curve with estimated survival function of the fitted models

convex, indicating a decreasing failure rate function for the first data set, which can be properly accommodated by an EPL model.

We fitted the EPL model, its submodels (EP and exponential (Exp)), and two other three-parameter models, the Weibull-geometric model (WG) proposed by Barreto-Souza et al. [2] and the exponentiated Weibull model (EW) proposed by Mudholkar and Srivastava [10] to the data set. For each model we then ran a total of 50,000 iterations, discarding the first 20,000 realizations as burn-in and thinning to every 5th iteration. Posterior results were then based on 6000 realizations of the Markov chain. The convergence was checked using the Geweke diagnostic, which did not indicate lack of convergence. The models were compared using some more knowledge criteria for this issue—the expected Akaike information criterion (EAIC), expected Bayesian

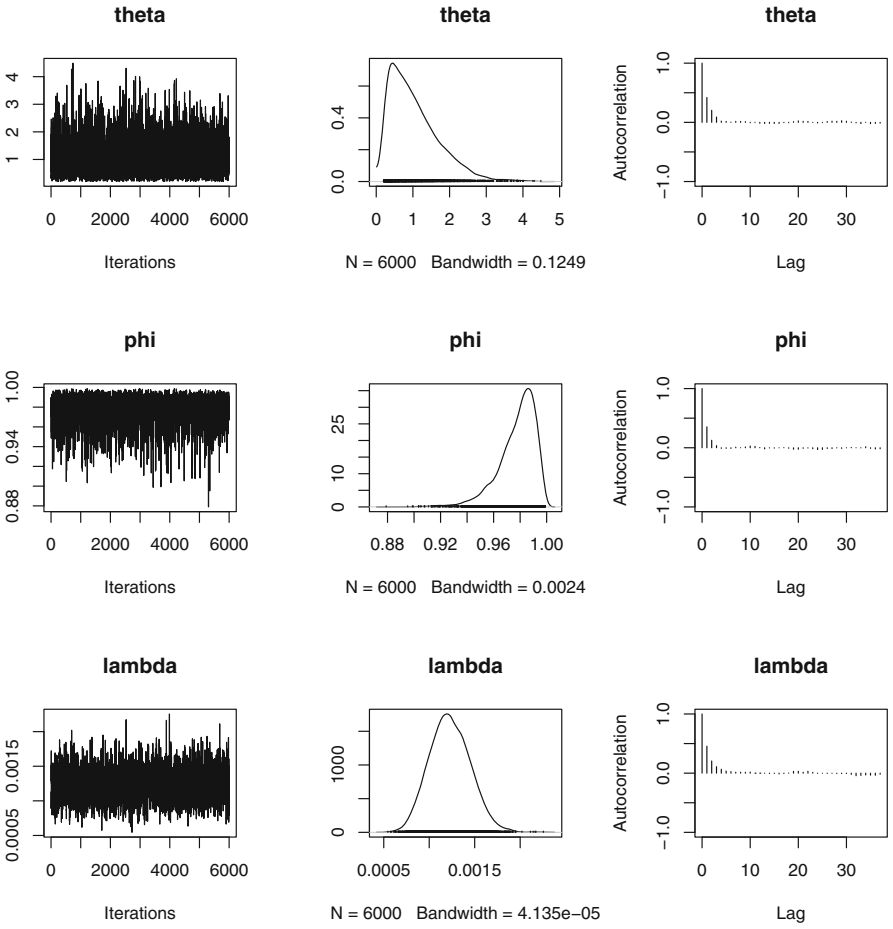

Fig. 21.5 The posterior marginal traces, density estimates, and autocorrelations of ϕ, λ, and θ, respectively

(or Schwarz) information criterion (EBIC), the deviance information criterion (DIC), and the Bayes factor (BF). For a model M with p_M parameters, these statistics are defined as

$$\begin{aligned} EAIC &= E[D_M(\boldsymbol{\theta})] + 2p_M, \\ EBIC &= E[D_M(\boldsymbol{\theta})] + p_M \log(n), \\ DIC &= 2E[D_M(\boldsymbol{\theta})] - D_M[E(\boldsymbol{\theta})], \end{aligned}$$

where $D_M(\boldsymbol{\theta})$ is the deviance function of model M defined as $-2\log L_M(\boldsymbol{\theta})$. The model with the smallest value for any one of these criteria (among all considered models) is commonly taken as the preferred model for describing the given data set.

Table 21.2 Bayesian estimates of the parameters and related statistics

Models	Param.	Mean	CI (95%)	SD	EAIC	EBIC	DIC	BF
EPL	ϕ	0.9772	(0.9402, 0.9965)	0.0152	1561.8	1570.0	1557.3	1.0000
	λ	0.0012	(0.0008, 0.0017)	0.0002				
	θ	1.0775	(0.2312, 2.8761)	0.7255				
EP	λ	0.0023	(0.0019, 0.0027)	0.0002	1608.9	1614.4	1606.0	$<$ 0.0001
	θ	0.0800	(0.0023, 0.2872)	0.0766				
Exp	λ	0.0023	(0.0019, 0.0027)	$<$ 0.0001	1605.0	1607.7	1604.0	$<$ 0.0001
	α	0.7616	(0.6123, 0.9501)	0.0904				
WG	β	0.0019	(0.0006, 0.0033)	0.0008	1565.9	1574.1	1562.2	0.3709
	p	0.4729	(0.0357, 0.8611)	0.2439				
	α	0.2862	(0.2257, 0.3421)	0.0308				
EW	β	10.9110	(1.6989, 25.132)	6.2684	1568.0	1576.2	1562.4	0.0628
	θ	6.0961	(3.7324, 10.778)	1.9116				

CI credible interval, *SD* standard deviation, *EAIC* expected Akaike information criterion, *EBIC* expected Bayesian information criterion, *DIC* deviance information criterion, *BF* Bayes factor, *EPL* exponential Poisson logarithmic, *EP* exponential Poisson submodel, *Exp* exponential submodel, *WG* Weibull geometric model, *EW* exponentiated Weibull model

The Bayes factor evidence of a model M_1 against a model M_2 for a data set $\mathbf{y}$ is defined as

$$BF(M_1; M_2) = \frac{p(\mathbf{y}|M_1)}{p(\mathbf{y}|M_2)},$$

where $p(\mathbf{y}|M)$ is the predictive density of $\mathbf{y}$ through the model M. Low value of $BF(M_1; M_2)$ is considered an evidence against M_1.

Table 21.2 shows the posterior mean, 95 % credible interval, and the standard error for the parameters of the models. The mentioned comparison criteria are present in the Table 21.2 too, where the Bayes factor is computed for all methods against the EPL model. The three-parameter models presented better results than the models EP and Exp, where EPL model was the one that obtained lower values of EAIC, EBIC, and DIC and higher value of BF. Fig. 21.5 shows the density posterior of the EPL parameters.

The graphic of the density functions of the fitted models superimposed on the histogram, and the graphic of the estimated survival functions of fitted models superimposed on the Kaplan–Meier curve are presented in the Fig. 21.4.

21.5 Concluding Remarks

In this chapter, we proposed a Bayesian approach of the EPL model, which is a three-parameter lifetime distribution with strong physical motivation and a possible extension to the EP model. In the simulation study we noted that the bias and MSEs of the Bayesian estimators go to zero quickly as sample size increases, while the MSEs of the classical estimators remained constant for some parameters, independent of the sample size. This suggests that the Bayesian estimator is better than the classical estimator for EPL model. Finally, the application of the model to a real data set was presented and discussed. The EPL model fit was superior to those obtained using its submodels (EP and Exponential models) as well as using other three-parameter models.

Acknowledgments The work of the first and second authors was funded by CAPES - Brazil. The authors thank the editor and two anonymous referees for their valuable comments.

References

1. Aarset, M.: How to identify a bathtub hazard rate. IEEE Trans. Reliability **2**, 106–108 (1987)
2. Barreto-Souza, W., de Morais, A.L., Cordeiro, G.M.: The Weibull-geometric distribution. J. Stat. Comput. Simul. **81**(5), 645–657 (2011) doi:10.1080/00949650903436554. http://www.tandfonline.com/doi/abs/10.1080/00949650903436554
3. Basu, A., Mukhopadhyay, C.: Bayesian analysis for masked system failure data using non-identical Weibull models. J. Stat. Plan. Inference **78**, 255–275 (1999)
4. Cancho, V.G., Louzada-Neto, F., Barriga, G.D.: The Poisson-exponential lifetime distribution. Comput. Stat. Data Anal. **55**, 677–686 (2011)
5. Escobar, L., Meeker, W. Jr: Assessing influence in regression analysis with censored data. Biometrics **48**, 507–528 (1992)
6. Fioruci, J.A., Bao, Y., Louzada, F., Cancho, V.G.: The exponential Poisson logarithmic distribution. Commun. Stat. Theory Methods (2014)
7. Flores, J.D., Borges, P., Cancho, V.G., Louzada, F.: The complementary exponential power series distribution: model, properties, estimation and a comparison with its counterpart. Braz. J. Probab. Stat. **1**, 10 (2012)
8. Louzada, F., Roman, M., Cancho, V.G.: The complementary exponential geometric distribution: model, properties, and a comparison with its counterpart. Comput. Stat. Data Anal. **55**(8), 2516–2524 (2011)
9. Mahmoudi, E., Sepahdar, A.: Exponentiated Weibull-Poisson distribution: model, properties and applications. Math. Comput. Simul. **92**, 76–97 (2013)
10. Mudholkar, G., Srivastava, D.: Exponentiated Weibull family for analyzing bathtub failure-rate data. IEEE Trans. Reliability **42**(2), 299–302 (1993)

Chapter 22
Bayesian Estimation of Birnbaum–Saunders Log-Linear Model

Elizabeth González Patiño

Abstract The Birnbaum–Saunders (BS) distribution was derived to model failure times of materials subjected to fluctuating stresses and strains. Motivated by applications in the characterizations of materials, in 1991 Rieck and Nedelman proposed a log-linear model for the BS distribution. This model has many applications, for instance, to compare the median time life of several populations or to assess the effect of covariates on accelerated life testing. In addition to the model studied under the classical approach, we considered Markov chain Monte Carlo (MCMC) and we made an implementation in `WinBUGS` to get a Bayesian approach under noninformative priori distribution. Similar results for both classical and Bayesian approaches were obtained. This implementation was also adapted for censoring and we assessed the influence of different percentages of censored data.

22.1 Introduction

Motivated by problems in airplanes due to the development and growth of a dominant crack, in 1969 Birnbaum and Saunders proposed the Birnbaum–Saunders (BS) distribution [2]. It describes the failure time T when some kind of accumulating damage $D(t)$ exceeds a threshold ω, i.e.,

$$T = \text{Inf}\{t : D(t) > \omega\}.$$

Let T be the time until the occurrence of the failure, then T is a BS random variable if its distribution is

$$F_T(t) = \Phi\left\{\frac{1}{\alpha}\left[\sqrt{\frac{t}{\beta}} - \sqrt{\frac{\beta}{t}}\right]\right\}, \quad t > 0, \text{ and } \alpha, \beta > 0.$$

E. G. Patiño (✉)
Instituto de Matemática e Estatística, Universidade de São Paulo, Rua do Matão, 1010-Cidade Universitária, São Paulo, São Paulo 05508-090, Brazil
e-mail: lizapat@ime.usp.br

A. Polpo et al. (eds.), *Interdisciplinary Bayesian Statistics,*
Springer Proceedings in Mathematics & Statistics 118, DOI 10.1007/978-3-319-12454-4_22

The probability density function (PDF) is given by

$$f_X(x;\alpha,\beta) = \frac{\sqrt{\frac{x-\mu}{\beta}} + \sqrt{\frac{\beta}{x-\mu}}}{2\alpha(x-\mu)}\phi\left(\frac{\sqrt{\frac{x-\mu}{\beta}} + \sqrt{\frac{\beta}{x-\mu}}}{\alpha}\right), \quad x > \mu \text{ and } \alpha, \beta > 0 \tag{22.1}$$

where μ, α, and β are, respectively, position, shape, and scale parameters. The parameter β also corresponds to the median value of the distribution. The functions $\Phi(x)$ and $\phi(x)$ are the standard normal cumulative distribution function (CDF) and PDF.

If $t = x - \mu$, we can write the PDF (22.1) as

$$f(t;\alpha,\beta) = \frac{t+\beta}{2\alpha\sqrt{2\pi\beta}t^{3/2}} \exp\left\{-\frac{1}{2\alpha^2}\left[\frac{t}{\beta} + \frac{\beta}{t} - 2\right]\right\}, \quad t > 0, \text{ and } \alpha, \beta > 0.$$

In 1991, Rieck and Nedelman [6] were interested in an application in which the main interest was to study the time of failure for a material subjected to different patterns of cycling forces. In order to do so, they proposed a log-linear model for the BS distribution. The model's principle is based on the empirical law

$$\ln(N) = a + bx, \tag{22.2}$$

where N is the number of cycles to failure of the specimen and x is either stress range per cycle, strain range per cycle, or the work per cycle.

According to Rieck and Nedelman (see [6]), under some assumptions, since N can be considered as a random variable, the Eq. (22.2) may be rewritten as

$$N = \mathrm{e}^{a+bx}\delta, \tag{22.3}$$

with $\delta \sim \mathrm{B}S(\alpha, 1)$.

Thereby, a log-linear model with an additive random effect is obtained by taking logarithm in (22.3),

$$\log(N) = a + bx + \log(\delta),$$

where $\log(\delta)$ has sinh-normal (SHN) distribution, SHN$(\alpha, 1)$.

The SHN distribution for a random variable T has distribution function given by

$$F_X(x) = \Phi\left(\frac{2}{\alpha}\sinh\left(\frac{x-\gamma}{\sigma}\right)\right), \quad x \in \mathbf{R}, \text{ and } \alpha, \sigma > 0,$$

where $\Phi(x)$ is the standard normal CDF. This distribution is symmetric around the location parameter γ, is unimodal for $\alpha \leq 2$ and bimodal for $\alpha > 2$, and the mean and variance are given by $\mathrm{E}(Y) = \gamma$ and $\mathrm{var}(Y) = \sigma^2\omega(\alpha)$, where $\omega(\alpha)$ is the variance when $\sigma = 1$. Other properties of SHN distribution can be checked in Rieck [5].

The SHN distribution is also called log-Birnbaum–Sanders with parameters α and γ, denoted as log-BS(α, γ), due to the relationship between the SHN and BS distribution [7], proved by Rieck et al. [6] in the following theorem.

Theorem 1 *Let T be a random variable such as $T \sim BS(\alpha, \beta)$. Then $Y = \log(T)$ has SHN distribution with shape, location, and scale parameter given, respectively, by $\alpha > 0$, $\gamma = \log(\beta)$ and $\sigma = 2$, thus, $Y = \log(T) \sim SHN(\alpha, \gamma, 2)$ with function probability density given by*

$$f(y; \alpha, \gamma) = \frac{1}{\alpha\sqrt{2\pi}} \cosh\left(\frac{y-\gamma}{2}\right) \exp\left\{-\frac{2}{\alpha^2} \sinh^2\left(\frac{y-\gamma}{2}\right)\right\}.$$

Due to the importance of this model to accelerate life testing or to compare the median lives of several populations, our purpose is to review it under a Bayesian perspective.

In order to make inferences we use posterior distribution generated from simulations by MCMC with `WinBUGS`. Since we are working on a Bayesian framework, it does not need large sample properties.

Achcar and Martinez [1] made an exploration of Bayesian methods for this model using a noninformative prior density for the parameters and found expressions for the marginal posterior densities through Laplace's methods for approximation of integrals.

In this work, we use a parametric priori density function and construct the maximum likelihood function to make a simple implementation on `WinBUGS`. This implementation was also adapted for censoring.

A life data set of 46 observations corresponding to the biaxial fatigue test of Brown and Miller, developed in 1978 [3], is used to compare the estimation under classical and Bayesian perspective.

22.1.1 Model

The generalization of Birnbaum–Saunders log-linear model is

$$Y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \qquad i = 1, \cdots, n, \tag{22.4}$$

where

- Y_i is the logarithm of the observed failure time T_i, $\{i = 1, \cdots, n\}$, $T_i \sim \text{BS}(\alpha_i, \beta_i)$ and the distribution of T_i depends on p explanatory variables $\vec{x}_i = (x_{i1}, \cdots, x_{ip})$;
- $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)$ is the vector of unknown parameters associated with the explanatory variables;
- ε_i is the random error of the model with $\varepsilon_i \sim \text{log-BS}(\alpha, 0)$, i.e., $\varepsilon_i \sim \text{SHN}(\alpha, 0, 2)$, $\{i = 1, \cdots, n\}$.

22.2 Estimation

Rieck and Nedelman in [6] proposed point estimation of parameters of the model (22.4) by maximum likelihood and least squares (LS). In this work, we consider MCMC simulations to get posterior densities of parameters of interest.

22.2.1 Maximum Likelihood (ML)

Consider n independent observations $y_1, y_2, \cdots, y_n$ under the model (22.4) , where $\varepsilon_i \sim \text{SHN}(\alpha, 0, 2)$. The likelihood function for $\boldsymbol{\varphi} = (\boldsymbol{\beta}^\top, \alpha)^\top$ is given by

$$L(\boldsymbol{\varphi}; y_i, x_i) = \prod_{i=1}^{n} \frac{1}{\alpha\sqrt{2\pi}} \cosh\left(\frac{y_i - \mathbf{x_i}^\top \boldsymbol{\beta}}{2}\right) \exp\left\{-\frac{2}{\alpha^2} \sinh^2\left(\frac{y_i - \mathbf{x_i}^\top \boldsymbol{\beta}}{2}\right)\right\}. \tag{22.5}$$

The log likelihood function is expressed as

$$l(\boldsymbol{\varphi}; y_i, x_i) \propto -n \ln \alpha + \sum_{i=1}^{n} \ln\left[\cosh\left(\frac{y_i - \mathbf{x_i}^\top \boldsymbol{\beta}}{2}\right)\right] - \frac{2}{\alpha^2} \sum_{i=1}^{n} \sinh^2\left(\frac{y_i - \mathbf{x_i}^\top \boldsymbol{\beta}}{2}\right). \tag{22.6}$$

Considering

$$W_i = \frac{2}{\alpha} \cosh\left(\frac{y_i - \mathbf{x_i}^\top \boldsymbol{\beta}}{2}\right) \quad \text{and} \quad Z_i = \frac{2}{\alpha} \sinh\left(\frac{y_i - \mathbf{x_i}^\top \boldsymbol{\beta}}{2}\right),$$

the expression (22.6) may be rewritten as

$$l(\boldsymbol{\varphi}; y_i, x_i) \propto \sum_{i=1}^{n} \ln W_i - \sum_{i=1}^{n} \frac{Z^2}{2}.$$

The score functions for $\boldsymbol{\beta}$ and α are given respectively by

$$\frac{\partial l(\varphi; y_i, x_i)}{\partial \beta_i} = \frac{1}{2} \sum_{i=1}^{n} x_{ij} \left\{Z_i W_i - \frac{Z_i}{W_i}\right\}, \quad j = 1, \cdots, p \text{ and}$$

$$\frac{\partial l(\varphi; y_i, x_i)}{\partial \alpha} = -\frac{n}{\alpha} + \frac{1}{\alpha} \sum_{i=1}^{n} \sinh\left(\frac{y_i - \mathbf{x_i}^\top \boldsymbol{\beta}}{2}\right). \tag{22.7}$$

From (22.7) it is possible to obtain an expression for the maximum likelihood estimation (MLE) of α^2 in terms of MLE vector $\boldsymbol{\beta}$, given by

$$\hat{\alpha}^2 = \frac{4}{n}\sum_{i=1}^{n} \sinh\left(\frac{y_i - \mathbf{x_i}^\top \hat{\boldsymbol{\beta}}}{2}\right).$$

However, the MLE of $\boldsymbol{\beta}$ must be obtained numerically. The authors propose an iterative procedure to obtain these estimators based on ordinary least squares estimators (LSE).

22.2.2 Least Squares (LS)

According to Rieck and Nedelman in [6], the estimation by ordinary LS produces explicit solutions for $\boldsymbol{\varphi}$ in (22.4). Although LS is not as efficient as ML, the estimates are unbiased. The $\boldsymbol{\beta}$ estimate is highly efficient for small values of α.

In model (22.4), $\mathrm{E}[\varepsilon_i] = 0$ and $\mathrm{Var}[\varepsilon_i] = 4\omega(\alpha)$. Since the observations $y_1, \cdots, y_n$ are independent, $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ $\{i, j = 1, \cdots, n\}$, and so the best linear unbiased estimator is

$$\hat{\varphi} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{Y},$$

with covariance matrix $\mathrm{Cov}(\hat{\varphi}) = 4\omega(\alpha)(\mathbf{X}^\top \mathbf{X})^{-1}$, and an unbiased estimator for $\omega(\alpha)$ is $\hat{\omega}(\alpha) = \sum_{i=1}^{n} \frac{(y_i - \mathbf{X}_i^\top \hat{\boldsymbol{\beta}})^2}{4(n-p)}$.

22.2.3 Bayesian Approach

For the Bayesian approach, we assumed independent priors gamma density function for the shape parameter, $\alpha \sim \mathrm{Gama}(\xi_0, \delta_0)$ and normal density function with mean zero for the parameters of the linear predictor coefficients, $\beta_i \sim \mathrm{N}(0, \sigma_{bj}^2)$, $\{j = 1, \cdots, p\}$. Thus, a priori density of $\boldsymbol{\varphi}$ is given by

$$\pi(\boldsymbol{\varphi}) = \pi(\alpha, \boldsymbol{\beta}) \propto \alpha^{\delta_0 - 1} \exp\{-\alpha\xi_0\} \prod_{j=1}^{p} \exp\left\{-\frac{\boldsymbol{\beta}^2}{2\sigma_{bj}^2}\right\}, \quad j = 1, \cdots, p.$$

Combining this expression with the likelihood function (22.5), we obtain the posterior density

$$\pi(\boldsymbol{\varphi}|y_i, x_i) = \pi(\alpha, \boldsymbol{\beta}|y_i, x_i) \propto \prod_{i=i}^{n} \alpha^{\xi_0 - 2} \cosh\left(\frac{y_i - \mathbf{x}_1^\top \boldsymbol{\beta}}{2}\right)$$

$$\exp\left\{-\frac{2}{\alpha^2}\sinh^2\left(\frac{y_i - \mathbf{x}_1^\top \boldsymbol{\beta}}{2}\right) - \alpha\xi_0\right\} \prod_{j=i}^{p}$$

$$\exp\left\{-\frac{\boldsymbol{\beta}^2}{2\sigma_{bj}}\right\} \propto \prod_{i=i}^{n}\prod_{j=i}^{p} W_i \alpha^{\xi_0-1} \exp\left\{-\frac{Z_i^2}{2} - \alpha\xi_0 - \frac{\boldsymbol{\beta}^2}{2\sigma_{bj}^2}\right\}, \tag{22.8}$$

where

$$W_i = \frac{2}{\alpha}\cosh\left(\frac{y_i - \mathbf{x}_i^\top \boldsymbol{\beta}}{2}\right) \quad \text{and} \quad Z_i = \frac{2}{\alpha}\sinh\left(\frac{y_i - \mathbf{x}_i^\top \boldsymbol{\beta}}{2}\right).$$

From (22.8), it is not simple to find the marginal posterior density for the model's parameters analytically. Notwithstanding, with `WinBUGS`, we may get the posterior density simulated by MCMC.

In the case of one explanatory variable, $\mu = \mathbf{x_i}^\top \boldsymbol{\beta} = \beta_0 + \beta_1 x$, the posterior density has the form

$$\pi(\alpha, \beta_0, \beta_1 | y_i, x_i) \propto \prod_{i=i}^{n} W_i \alpha^{\xi_0-1} \exp\left\{-\frac{Z_i^2}{2} - \alpha\xi_0 - \frac{\beta^2}{2\sigma_{b0}^2} - \frac{\beta^2}{2\sigma_{b1}^2}\right\}.$$

A possible implementation for this with priors $\alpha \sim \text{Gama}(0.001, 0.001)$ and $\beta_i \sim \text{N}(0, 100)$ $\{j = 1, 2\}$ is given below.

```
model
{
c<-10
for(i in 1:n)
{
u[i]=b0+b1*x[i]
logver[i]<--log(a)+log(cosh((y[i]-u[i])/2))-(2/pow (a,2))*
pow(sinh((y[i]-u[i])/2),2)
zeros[i]<-0
aux[i]<--logver[i]+c
zeros[i]~dpois(aux[i])
}
b0~dnorm(0,0.01)
b1~dnorm(0,0.01)
a~dgamma(0.001,0.001)
}
```

Censored Data In the case where random censoring is observed, with δ_i the failure indicator variable ($\delta_i = 1$ for failure and $\delta_i = 0$ for censoring) under the model (22.4), the likelihood function in terms of W_i and Z_i is given by

$$L(\boldsymbol{\varphi}; y_i, x_i) = \propto \prod_{i=i}^{n}\left[\frac{W_i}{2}\exp\left\{-\frac{Z_i^2}{2}\right\}\right]^{\delta_i}[1 - \Phi(Z_i)]^{1-\delta_i},$$

where $\Phi(.)$ is the standard normal CDF. Combining with the prior $\pi(\boldsymbol{\varphi})$, the posterior density can be obtained:

$$\pi(\varphi | y_i, x_i) \propto \prod_{i=i}^{n}\prod_{j=i}^{p}\left[\frac{W_i}{2}\exp\left\{-\frac{Z_i^2}{2}\right\}\right]^{\delta_i}[1 - \Phi(Z_i)]^{1-\delta_i}\alpha^{\xi_0-1}\exp\left\{-\alpha\xi_0 - \frac{\beta^2}{2\sigma_{bj}^2}\right\}.$$

Simulations of marginal posterior densities can be obtained in WingBUGS with the following implementation (considering one explanatory variable).

```
model
{
c<-10
for(i in 1:n)
{
u[i]=b0+b1*x[i]
logver[i]<-delta[i]*(-log(a)+log(cosh((y[i]-u[i])/2))
-(2/pow(a,2))*pow(sinh((y[i]-u[i])/2),2))+
(1-delta[i])*log(1-phi(2/a*sinh((y[i]-u[i])/2)))
zeros[i]<-0
aux[i]<--logver[i]+c
zeros[i]~dpois(aux[i])
}
b0~dnorm(0,0.01)
b1~dnorm(0,0.01)
a~dgamma(0.001,0.001)
}
```

22.3 Application

A data set of 46 observations from Brown and Miller's biaxial fatigue test (1978) [3] was analyzed by Rieck and Nedelman [6] and has been reviewed.

In the test, cylindrical specimens were subjected to axial loads and torsion on constant amplitude cycles to failure . The response variable is the number of cycles to the occurrence of failure N and the explanatory variable is the work per cycle in M_j/m^3. Hence, the interest is to model the number of cycles until failure.

Figure 22.1 shows an asymmetric behavior of response variable indicating that a Birnbaum–Saunders regression model can be appropriate. Let n_i be independent random variables such as $N_i \sim \mathrm{B}S(\alpha, \mu_i), \{i = 1, \cdots, n\}$. From empirical laws, consider the model

$$\ln(\mu_i) = \beta_0 + \beta_1 \ln(W_i), \quad i = 1, \cdots, 46,$$

where $x = \log(W_c)$ and W_c is the work per cycle.

The results of the model fitted under classical perspective are shown in Table 22.1. The second column corresponds to the numeric solution from the analytical derivatives using the package optim from R.

Table 22.2 corresponds to the results under the Bayesian framework by the WinBUGS' implementation, considering distributions Gamma(0.001 ; 0.001) e Unif(0;10) as priori distribution for α and N(0,100) for β_j, $\{j = 1, 2\}$. Chains with 21,000 iterations were considered, with just a spacing of length 10 to minimize the problem of simulated series autocorrelation. To reduce the effect of initial points, the first 1000 iterations were discarded.

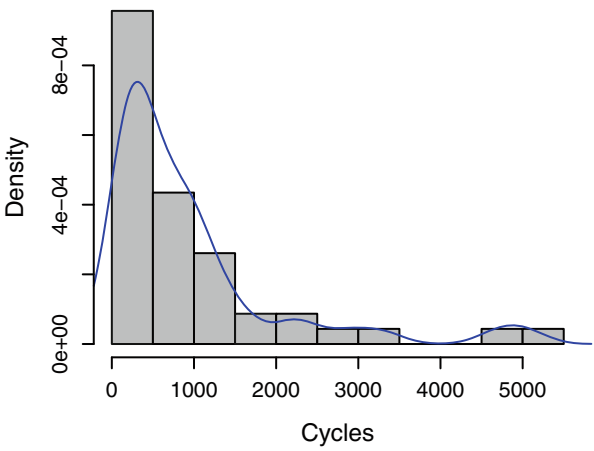

Fig. 22.1 As the histogram of the response variable has an asymmetric behavior, and it is concentrated in the range 0–1000, a Birnbaum–Saunders regression model is appropriate

Table 22.1 Estimates of Birnbaum–Saunders log-linear model under classical approach

Parameter	`optim` (SE)	MLE (SE)	LQE (SE)
α	0.417 (0.043)	0.41	
β_0	12.208 (0.392)	12.280 (0.403)	12.289 (0.406)
β_1	−1.654 (0.109)	−1.671 (0.112)	−1.673 (0.113)

For all situations, it was considered that $\alpha^{(0)} = 0.5$ and LSE for $\beta_0^{(0)} = 12.211$ and $\beta_1^{(0)} = -1.655$ as initial values for simulations.

The convergence of the chains simulates was previously verified. Figures 22.2 and 22.3 correspond to posterior density function and its simulation history, according to Table 22.2.

Based on our results, we note that the prior distribution for α does not appreciably affect the results, the estimates are similar and the Deviance information Criteria (DIC) for the model selection does not change considerably. We can also observe similar estimates from the classical and Bayesian framewok. A residual analysis for classical fit is presented by Dos Santos [4].

Table 22.2 Estimates of Birnbaum–Saunders log-linear model under Bayesian approach

Parameter/priori	Mean	SD	Per. 2.5	Per. 97.5
$\alpha \sim$ G(0.001;0.001)	0.4338	0.0474	0.3529	0.5376
$\beta_0 \sim$ N(0,100)	12.19	0.4094	11.380	13.000
$\beta_1 \sim$ N(0,100)	−1.648	0.1142	−1.872	−1.422
DIC: 889.6	pD: 2.948			
$\alpha \sim$ U(0;10)	0.4388	0.0487	0.3561	0.5482
$\beta_0 \sim$ N(0,100)	12.18	0.4146	11.380	13.010
$\beta_1 \sim$ N(0,100)	−1.648	0.1158	−1.878	−1.421
DIC: 889.7	pD: 2.939			

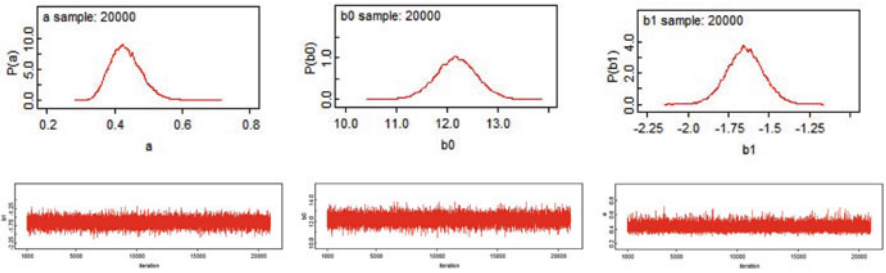

Fig. 22.2 Posterior densities and their simulation history. With prior $\alpha \sim$ Gama(0.001; 0.001), $\beta_0 \sim$ N(0, 100) and $\beta_1 \sim$ N(0, 100)

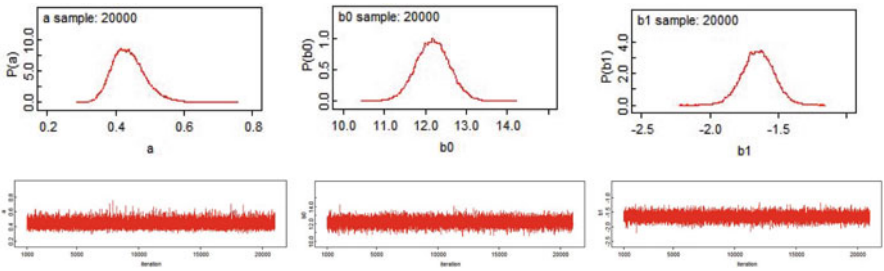

Fig. 22.3 Posterior densities and their simulation history. With prior $\alpha \sim$ U(0; 10), $\beta_0 \sim$ N(0, 100), and $\beta_1 \sim$ N(0, 100)

Censored Data In order to make inference in the presence of censored data, different percentages of random censoring were considered for biaxial fatigue data set. The observations were artificially censored. The estimates are shown in Table 22.3 and marginal posterior densities for 10, 30 and 45 % of censored observation are presented in Figs. 22.4, 22.5, and 22.6, respectively.

We notice that as the censure increases, there is low accuracy due to increase of standard error. We also notice a smaller DIC for low percentage of censoring. From the posterior density for α, the right tail becomes heavier when the percentage of censoring increases.

22.4 Discussion

A motivation for this work was to fit the Birnbaum–Saunders log-linear model proposed in 1991 by Rieck and Nedelman under a Bayesian approach and to compare it with the usual classical fit, which is based on the asymptotical properties for the estimator.

Table 22.3 Results under Bayesian approach of model fitted BS log-linear model for random censoring of biaxial fatigue data set

% Censoring	Parameter	Estim.	SE^a	Lower bound	Upper bound
10 %	α	0.437	0.051	0.351	0.550
	β_0	12.450	0.431	11.600	13.310
	β_1	−1.709	0.119	−1.943	−1.473
DIC: 899.7	pD: 2.938				
30 %	α	0.493	0.067	0.382	0.644
	β_0	12.160	0.490	11.180	13.110
	β_1	−1.597	0.138	−1.864	−1.319
DIC: 924.4	pD: 2.902				
45 %	α	0.550	0.088	0.410	0.752
	β_0	11.890	0.577	10.740	13.010
	β_1	−1.488	0.164	−1.797	−1.151
DIC: 936.5	pD: 2.884				

[a]Standard error

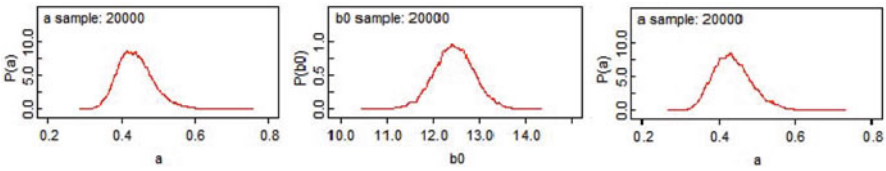

Fig. 22.4 Marginal posterior densities of BS log-linear model with 10 % of censoring

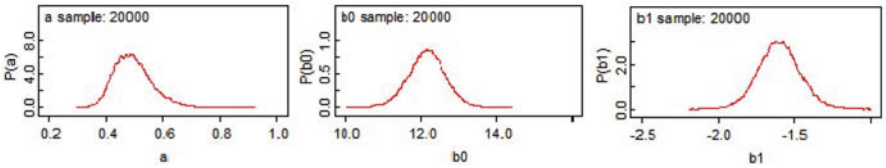

Fig. 22.5 Marginal posterior densities of BS log-linear model with 30 % of censoring

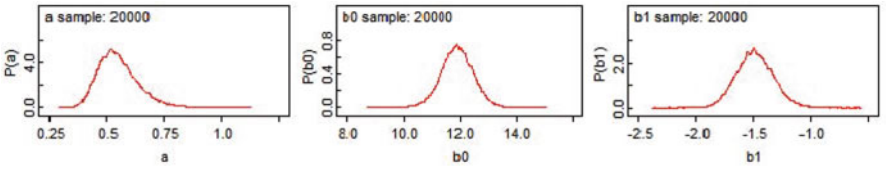

Fig. 22.6 Marginal posterior densities of BS log-linear model with 45 % of censoring

In this study, we show the posterior density distribution assuming independent priors—gamma density function for the shape parameter and normal density function with mean zero for the parameters of the linear predictor coefficients. Also we consider right-censored data and in both situations, it is not easy to obtain analytical expressions for the marginal posterior densities for the parameters of interest. However, we can see that WinBUGS is a useful tool because it allows one to obtain marginal posterior densities considering MCMC with a simple implementation.

Based on the application results, large differences were not observed between the classical and Bayesian framework. Furthermore in all situations, the Markov chains converged quickly and the computational time was short. Notwithstanding, it could be appropriate to conduct a simulation study to determine the optimal values for the parameters of the priori density function.

Since the fit of the Birnbaum–Saunders log-linear model under a Bayesian approach was suitable, it will be a good idea to make a Bayesian residual analysis.

References

1. Achcar, J.A., Martinez, M.: Bayesian methods in accelerated life test considering a log-linear model for the Birnbaum–Saunders distribution. Rev. Bras. Estat. **52**, 47–68 (1991)
2. Birnbaum, Z.W., Saunders, S.C.: A new family of life distributions. J. Appl. Probab. **6**, 319–327 (1969)
3. Brown, M.W., Miller, K.J.: Biaxial fatigue data. Report CEMR1/78, University of Sheffield, Department of Mechanical Engineering (1978)
4. Dos Santos, M.F.: Estimação e modelagem com a distribuição Birnbaum-Saunders: uma nova reparametrização. Dissertation (Master: Estatística Matemática), Universidade Federal de Pernambuco, Recife (2010)
5. Rieck, J.R.: Statistical analysis for the Birnbaum Saunders fatigue life distribution. Unpublished Ph.D. thesis, Clemson University, Department of Mathematical Science (1989)
6. Rieck, J.R., Nedelman, J.R.A.: Log-linear model for the Birnbaum–Saunders distribution. Technometrics **33**(1), 51–60 (1991)
7. Villegas, C.M.: Modelos log-Birnbaum-Saunders mistos. Tese (Doctorate: Estatística), Universidade de São Paulo, São Paulo (2010)

Chapter 23
Bayesian Weighted Information Measures

Salimeh Yasaei Sekeh

Abstract Following Ebrahimi et al. (J Stat Res Iran 3:113–137, 2006), we study weighted information measure in univariate case. In particular, we address the concept of comparison models based on information measure and, in our case, specially Kullback–Leibler discrimination measure. The main result is presenting the relationship of weighted mutual information measure and weighted entropy. Indeed, the importance of Weibull distribution family in weighted Kullback–Leibler information and Kullback–Leibler information has been carefully examined, which is useful in comparison models. As a notable application of the result, we study normal distributions, which can prove the expected motivation.

23.1 Introduction

The Kullback–Leibler information measure, also known as relative entropy or Kullback–Leibler divergence, between two probability density functions $f(x)$ and $g(x)$ with support S and distribution functions $F(x)$ and $G(x)$ respectively,

$$I(f,g) = \int_S \log \frac{f(x)}{g(x)}\, dF(x), \tag{23.1}$$

is commonly used in statistics as a measure of similarity between two density distributions. The divergence satisfies three properties, hereafter referred to as the divergence properties:

1. Self similarity: $I(f, f) = 0$
2. Self identification: $I(f, g) = 0$ if and only if $f = g$ almost everywhere
3. Positivity: $I(f, g) \geq 0$ for all f, g

An information discrepancy function maps how different are the two distributions, but it is worthwhile to mention that it does not indicate which of the two distributions is more informative. On the other hand, a discrepancy function between distribution F

S. Y. Sekeh (✉)
Department of Statistics, UFSCar, São Carlos, Brazil
e-mail: sa_yasaei@yahoo.com

A. Polpo et al. (eds.), *Interdisciplinary Bayesian Statistics,*
Springer Proceedings in Mathematics & Statistics 118, DOI 10.1007/978-3-319-12454-4_23

and the uniform distribution quantifies the information associated with a probability distribution F, see [7].

The Kullback–Leibler divergence is used in many aspects such as, determining if two acoustic models are similar, optimization by minimizing and maximizing the Kullback–Leibler information between distributions, and hypothesis testing and model evaluation.

The notion of entropy was introduced and developed in the contexts of statistical mechanics and system features, where Shannon, in 1948 [10], proposed the use of a measure of uncertainty for discrete distributions as a measure in information theory. The entropy $H(X)$ of a continuous random variable with support S and having density and distribution functions $f(x)$ and $F(x)$, respectively, is defined as

$$H(X) = -E\left[\log f(X)\right] = -\int_S \log f(x) dF(x). \tag{23.2}$$

Note that entropy is a function of the distribution of X. It does not depend on the actual values taken by the random variable X, but only on the probabilities.

Now, let us mention the conditional entropy of a random variable given as the expected value of the entropies of the conditional distributions, averaged over the conditioning random variable. For more details see [3].

Consider joint random variable (X, Y) with support $(S_1 \times S_2)$, joint and conditional probabilities $f(x, y)$ and $f(y|x)$, respectively, the conditional entropy $H(X|Y)$ and joint entropy $H(X, Y)$ are expressed as,

$$\begin{aligned} H(X|Y) &= -E_{X,Y}\left[\log f(X|Y)\right] = -E_Y\left[H(X|Y = y)\right] \\ &= -\int\int_{S_1\times S_2} f(x, y) \log f(x|y) dx dy, \\ H(X, Y) &= -E_{X,Y}\left[\log f(X, Y)\right] = -\int\int_{S_1\times S_2} f(x, y) \log f(x, y) dx dy. \end{aligned}$$

It can be shown that the naturalness of the definition of joint entropy and conditional entropy is exhibited by the fact that the entropy of a pair of random variables is the entropy of one and the conditional entropy of the other. On the other hand,

$$H(X, Y) = H(Y) + H(X|Y).$$

Note that the above relation is famous as chain rule theorem.

One of the important questions for researchers in statistics is to what extent the use of a variable Y reduces uncertainty about predicting the outcomes of another random variable X. Retzer et al. (2008) provides a comprehensive treatment of this subject. Assume random variable Y has distribution F_Y, information provided by an observation $Y = y$ about predicting outcomes of X can be measured by positive information function $I(f_{X|y}, f_X)$, where f_X is the marginal distribution of X and $f_{X|y}$ is the conditional distribution of X given y.

Note that the information function does not indicate which of the two densities (conditional density $f_{X|y}$ or the marginal density f_X) is more concentrated but it shows how different are the marginal and conditional distributions.

In general, one of the main tools that was used to measure the information provided by an observation $Y = y$ about predicting outcomes of X is the uncertainty difference, $\triangle H(f_X, f_{X|y}) = H(X) - H(X|y)$. However we note that $\triangle H(f_X, f_{X|y})$ can be positive (or negative) when conditional density $f_{X|y}$ is farther (or closer) to uniformity than the marginal density f_X.

Furthermore, to detect the information provided by an observation $Y = y$ about a random variable prospect X means to determine the entropy difference $\triangle H(f_X, f_{X|y})$ or the Kullback–Leibler $I(f_{X|y}, f_X)$ [5]. It can be observed that the expected information, which is known as mutual information, $I(f_{X,Y}, f_X f_Y)$, is given by,

$$I(f_{X,Y}, f_X f_Y) = E_Y [H(X) - H(X|y)] = E_Y \left[I(f_{X|y}, f_X) \right]. \tag{23.3}$$

As it has been seen in (23.3), mutual information is the information discrepancy between the actual joint distributions of two random variables and their joint distribution as if they were independent.

Following the results in [3], we are able to give other representations of mutual information as:

$$I(f_{X,Y}, f_X f_Y) = H(X) - H(X|Y) = H(X) + H(Y) - H(X, Y). \tag{23.4}$$

We continue this chapter by presenting the concept of weighted entropy and information measures.

For n events $E_1, E_2, \ldots, E_n$ and $F_1, F_2, \ldots, F_n$ with probabilities $p_1, p_2, \ldots, p_n$ and $q_1, q_2, \ldots, q_n$ respectively, we have demonstrated that the information supplied by the events E_i and F_i having probabilities p_i, q_i and utility u is given by

$$I_i = I(u_i, p_i, q_i) = u_i \log \frac{p_i}{q_i}.$$

The average of the information supplied by events $E_1, E_2, ..., E_n$ and $F_1, F_2, ..., F_n$ is obtained as

$$I = I(u_1, ..., u_n, p_1, ..., p_n, q_1, ..., q_n) = \sum_{i=1}^{n} p_i I_i = \sum_{i=1}^{n} p_i u_i \log \frac{p_i}{q_i}. \tag{23.5}$$

Let us put $u_1 = u_2 = ... = u_n = 1$; in this case relation (23.5) implies Kullback–Leibler information. In agreement with Belis and Guiasu [1], we recall the information supplied by $E_1, E_2, \ldots, E_n$ and $F_1, F_2, \ldots, F_n$, which is mentioned in (23.5), weighted Kullback–Leibler information.

In continuous case, suppose X and Y be two nonnegative random variables with probability functions $f(x)$ and $g(x)$, respectively, the weighted Kullback–Leibler information denotes by $I^w(f, g)$ is

$$I^w(f, g) = \int_S x \log \frac{f(x)}{g(x)} \, dF(x). \tag{23.6}$$

In the field of transmission systems of communication, it was realized that a general notion capable of abstracting various kinds of transmitted signals was necessary. A possible solution is to consider signals as random abstract events, allowing the possibility of defining the quantitative aspects of information based on the probability of different events. In fact, the occurrence of an event removes a double uncertainty: the quantitative one related to its probability, and the qualitative one related to its utility, where the utility of an event is independent of its probability. In this situation, one can measure the information for the occurrence of an event with probability p and utility u through a quantity depending on both variables. In particular, in the context of theoretical neurobiology, some measures of uncertainty based on the notion of the weighted entropy as the measure of uncertainty for an experiment with finite measurable partition $A_1, A_2, \ldots, A_n, (n > 1)$ have been considered, defined by Belis and Guiasu [1] as

$$H_n^1(p, u) = -\sum_{k=1}^{n} u_k p_k \log p_k,$$

where p_k is the probability of the event A_k with utility u_k.

In agreement with Belis and Guiasu [1] and Guiasu [6], Di Cresenzo and Longobardi [4] have demonstrated an analogous definition for the notion of weighted entropy in continuous case as following:

$$H^w(X) = -E\left[X \log f(X)\right] = -\int_S x \log f(x)\, dF(x). \tag{23.7}$$

Note that the weighted entropy is a function of the distribution of nonnegative random variable X and actual values taken by X.

Bayesian information measures are for an observation y made or to be made from a random variable Y having a probability function $f_{Y|\theta}$ about the parameter Θ when the prior can be described by a probability distribution with density f_Θ. The main target is to measure the information provided by data about the parameter which can be applied to design comparison, data evaluation, and mainly model comparison.

For more details, refer to Lindley [8], Zelner [13, 14], Bernardo [2], and Soofi [9, 11, 12].

Our purpose in this chapter is to introduce Bayesian weighted information measure and compare it with Bayesian information measure in univariate case. The main application of our result (see Sect. 23.3) is directed toward model comparison, although some other interesting examples, such as normal case, have been presented.

The rest of the chapter is organized as follows: Section 23.2 presents the notion of weighted uncertainty and weighted mutual information functions. Furthermore, the weighted general entropy is introduced. However, this concept is used for pointing out the relationship between weighted mutual information and weighted uncertainty. Finally, in this section, by presenting a well-known example as normal distribution, our results are illustrated, which coincide with our expectations.

In Sect. 23.3, the expected weighted information is defined. Afterward, among all distribution functions, the Weibull distribution is considered. By presenting this

example and using the weighted Kullback–Leibler information in Bayesian analysis, different models of Weibull family are compared.

23.2 Weighted Uncertainty and Information Measures

Here the definitions and properties, which we shall exploit later to use, and main purpose of Bayesian analysis are stated.

Definition 1 Assume X and Y as a pair of nonnegative continuous random variables with joint support $S_1 \times S_2$, having joint and conditional density functions $f(x, y)$ and $f(y|x)$ respectively, the weighted joint entropy and conditional entropy are given as:

$$H^w(X, Y) = -E_{X,Y}\left[XY \log f(X, Y)\right] = -\int\int_{S_1 \times S_2} xyf(x, y) \log f(x, y)dxdy, \tag{23.8}$$

$$H^w(Y|X) = -E_{X,Y}\left[XY \log f(Y|X)\right] = -\int\int_{S_1 \times S_2} xyf(x, y) \log f(y|x)dxdy. \tag{23.9}$$

Without different indication, it will be noted that $H^w(X, Y)$ is as defined in [4].

Hereafter, we switch to introduce the extension of the weighted entropy as defined in (23.7), and we shall call it a generalized weighted entropy and denote the same by $H^{w,\phi}(X)$, where ϕ is any nonnegative and differentiable utility function.

$$H^{w,\phi}(X) = -E_X\left[\phi(X) \log f(x)\right] = -\int_S \phi(x) \log f(x)dF(x). \tag{23.10}$$

Taking into account the Definition 1 and relation (23.10), we are entitled to give the following theorem that is known as the chain rule theorem. In fact the following rule shows that the weighted entropy of a pair of random variables is the generalized weighted entropy of one and the weighted conditional entropy of the other.

Theorem 1 *Chain rule: Assume that (X, Y) is a pair of nonnegative random variables, then,*

$$H^w(X, Y) = H^{w,\phi}(X) + H^w(Y|X) \quad \textit{where} \quad \phi(X) = XE_Y\left[Y|X\right].$$

Proof We observe that for nonnegative (X, Y),

$$H^w(X, Y) = -\int_0^\infty\int_0^\infty xyf(x, y) \log f(y|x)dxdy - \int_0^\infty\int_0^\infty xyf(x, y) \log f(x)dxdy$$

$$= H^w(Y|X) - \int_0^\infty x f(x) \log f(x) \left[\int_0^\infty y f(y|x) dy \right] dx$$
$$= H^w(Y|X) - \int_0^\infty x E_Y [Y|x] f(x) \log f(x) dx.$$

Using now the relation (23.10) by replacing $\phi(X) = X E_Y [Y|X]$, we obtain the result. □

Equivalently we can write,

$$H^w(X, Y) = H^{w,\phi}(Y) + H^w(X|Y); \quad \phi(Y) = Y E_X [X|Y].$$

When turning our attention to independent random variables, Di Cresenzo and Longobardi [4] emphasize the role of the mean value in the evaluation of joint weighted entropy, which is the following:

Proposition 1 *If X and Y are nonnegative independent random variables, it can yield,*

$$H^w(X, Y) = E(Y) H^w(X) + E(X) H^w(Y).$$

Moreover, let us continue with a new notion as weighted mutual information, which is a measure of the amount of information that one random variable contains about another random variable; however in weighted case, in spite of probability, the actual values of random variables also are considered.

Definition 2 Consider two nonnegative random variables with joint density functions $f(x, y)$ and marginal density functions $f(x)$ and $f(y)$; the weighted mutual information is weighted Kullback–Leibler information between the joint distribution and product distribution $f(x)\ f(y)$as,

$$I^w(f_{X,Y}, f_X f_Y) = \int_0^\infty \int_0^\infty xy f(x, y) \log \frac{f(x, y)}{f(x) f(y)} dx dy. \tag{23.11}$$

In the following result, we explicitly note that the weighted mutual information is equal to the weighted entropy of X in the reduction of its uncertainty due to the knowledge of Y.

Remark 1 For nonnegative random variables X and Y, we can write,

$$I^w(f_{X,Y}, f_X f_Y) = H^{w,\phi}(X) - H^w(X|Y) \quad ; \quad \phi(X) = X E_Y [Y|X]. \tag{23.12}$$

Proof

$$I^w(f_{X,Y}, f_X f_Y) = \int_0^\infty \int_0^\infty xy f(x, y) \log \frac{f(x, y)}{f(y)} dx dy - \int_0^\infty \int_0^\infty xy f(x, y) \log f(x) dx dy$$
$$= \int_0^\infty \int_0^\infty xy f(x, y) \log f(x|y) dx dy$$

$$-\int_0^\infty \int_0^\infty xyf(y|x)f(x)\log f(x)dxdy$$
$$= -H^w(X|Y) - \int_0^\infty xE_Y[Y|x]f(x)\log f(x)dx.$$

This is actually what we are looking for, hence the final result is achieved. □

The next corollary is an immediate consequence of Theorem 1 and previous remark.

Corollary 1 *With reference to notions and definitions from the beginning of this chapter, we have,*

$$I^w(f_{X,Y}, f_X f_Y) = H^{w,\phi_1}(X) + H^{w,\phi_2}(Y) - H^w(X,Y),$$

where $\phi_1(X) = XE_Y[Y|X]$ *and* $\phi_2(Y) = YE_X[X|Y]$.

At this point of time, the question that arises is to what extent the use of a variable Y reduces the uncertainty about predicting the outcomes of another random variable X but even consists the actual value of variable X. We explicitly answer this question using weighted information provided by an observation $Y = y$ about prediction outcomes of X. Although in the following remark , an alternative relation for this measure in terms of general weighted and conditional entropies can be seen.

Remark 2 Assuming X and Y are nonnegative random variables and recalling the weighted Kullback–Leibler information, we obtain:

$$I^w(f_{X|y}, f_X) = H^{w,\phi'}(X) - H^w(X|y);\ \phi'(X) = \frac{Xf(X|y)}{f(X)}, \tag{23.13}$$

where $f(X|y)$ is the conditional density function.

Moreover, going back to the weighted mutual information, we interpret that this information can also be obtained by using the weighted Kullback–Leibler information with the following expression:

$$I^w(f_{X,Y}, f_X f_Y) = E_Y\left[YI^w(f_{X|y}, f_X)\right] = E_Y\left[Y(H^{w,\phi'}(X) - H^w(X|y))\right]. \tag{23.14}$$

Proof From the definition of $I^w(f,g)$, we observe that,

$$I^w(f_{X|y}, f_X) = \int_0^\infty xf(x|y)\log\frac{f(x|y)}{f(x)}dx$$
$$= \int_0^\infty xf(x|y)\log f(x|y)dx - \int_0^\infty xf(x|y)\log f(x)dx$$
$$= -H^w(X|y) - \int_0^\infty x\frac{f(x|y)}{f(x)}.f(x)\log f(x)dx.$$

We can see that the last expression is exactly our aim result. Moreover, if we multiple the final statement in Y and take the expectation with respect to it, we can obtain the weighted mutual information as,

$$\begin{aligned} E_Y\left[YI^w(f_{X|y}, f_X)\right] &= E_Y\left[Y(H^{w,\phi'}(X) - H^w(X|y))\right] \\ &= -\int_0^\infty \int_0^\infty xyf(x|y)f(y)\log f(x)dxdy \\ &\quad + \int_0^\infty \int_0^\infty xyf(x|y)f(y)\log f(x|y)dxdy \\ &= \int_0^\infty \int_0^\infty xyf(x,y)\log \frac{f(x|y)}{f(x)}dxdy \\ &= \int_0^\infty \int_0^\infty xyf(x,y)\log \frac{f(x,y)}{f(x)f(y)}dxdy \\ &= I^w(f_{X,Y}, f_X f_Y). \end{aligned}$$

□

In this part of work, it is worthwhile to mention that throughout the chapter, weighted entropy with positive utility is considered, on the other words, if random variable is not always nonnegative with support S, then we define weighted entropy as:

$$H^w(X) = -E\left[|X|\log f(X)\right] = -\int_S |x|f(x)\log f(x). \tag{23.15}$$

And moreover,

$$I^w(f,g) = \int_S |x|\log \frac{f(x)}{g(x)}\, dF(x). \tag{23.16}$$

We conclude this section by presenting an example as an application of information, and comparing two Kullback–Leibler and weighted Kullback–Leiber information measures due to normal random variables.

Example 1 Among all the statistical models, consider the famous standard normal distribution for both random variables X and Y. As already we have observed in this case, the joint density function for the pair of random variables (X, Y) with the coefficient correlation ρ is given by,

$$f(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}}\exp\left[\frac{-(x^2 - 2\rho xy + y^2)}{2(1-\rho^2)}\right], \tag{23.17}$$

where $(x, y) \in \mathbb{R}^2$ and $|\rho| \leq 1$.

Taking into account all these assumptions, one can conclude that the conditional random variable $X|Y$ also has the normal distribution $\mathcal{N}(\rho Y, (1-\rho^2))$.

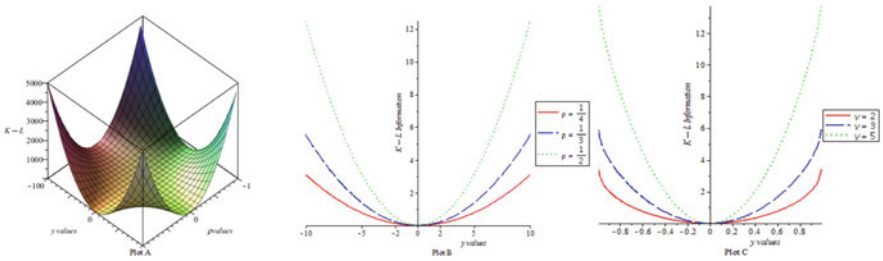

Fig. 23.1 *Plot A*: $I(f_{X|y}, f_X)$ and different value y and ρ in normal distribution, *Plot B*: $I(f_{X|y}, f_X)$ for different value $\rho = \frac{1}{2}, \frac{1}{3}, \frac{1}{4}$, *Plot C*: $I(f_{X|y}, f_X)$ for different values $y = 2, 3, 5$

By applying some simple computation for normal distribution with mean μ and variance σ^2, we can obtain,

$$H(X) = \log \sqrt{2\pi\sigma^2} + \frac{1}{2}.$$

Now let us to recall the conditional random variable $X|Y = y$. It has been seen that entropy does not depend on mean parameters, hence,

$$H(X|y) = \log \sqrt{2\pi(1-\rho^2)} + \frac{1}{2}. \tag{23.18}$$

On the other side, with straightforward computations, we give the following expression:

$$\int_{\mathbb{R}} f(x|y) \log f(x) dx = -\log \sqrt{2\pi} - \frac{1}{2}\left[1 - \rho^2 + y^2\rho^2\right]. \tag{23.19}$$

Thus, it descends directly from (23.18) and (23.19) that the Kullback–Leibler information between $f_{X|y}$ and f_X is a quadratic function, so that we have,

$$I(f_{X|y}, f_X) = \frac{1}{2}\rho^2(y^2 - 1) - \log \sqrt{1-\rho^2}. \tag{23.20}$$

Here, it is worthwhile noting that when the correlation ρ between X and Y is increasing for given $Y = y$, $I(f_{X|y}, f_X)$ is also increasing. In fact, the Fig. 23.1 proves this result. We can see (Plot B) that for different values of ρ, such as $\rho = \frac{1}{2}$, $\rho = \frac{1}{3}$ and $\frac{1}{4}$, the information discrepancy between the distribution of a random variable X by different observations for Y and marginal distribution X raises. Moreover, with fixing the value y, the function (23.20) is an increasing function with respect tothe absolute value of ρ (see Fig. 23.1, Plot C). All this justifies the meaning of information and coincides with our expectations as when X and Y are more dependent, then having information about Y increases the distance between $f_{X|y}$ and f_X.

Now, the question arises that can we have the same result for $I^w(f_{X|y}, f_X)$. For this reason, by recalling relation (23.16), we compute the weighted Kullback–Leibler information as follows:

$$I^w(f_{X|y}, f_X) = H^{w,\phi'}(X) - H^w(X|y). \tag{23.21}$$

Note that the general equation (23.15) for the normal random variable X with mean μ and variance σ^2 yields,

$$H^w(X) = \log\sqrt{2\pi\sigma^2}\left[\frac{2\sigma}{\sqrt{2\pi}}.e^{-\frac{\mu^2}{2\sigma^2}} + \mu\left(\bar{F}_z\left(-\frac{\mu}{\sigma}\right) - F_z\left(-\frac{\mu}{\sigma}\right)\right)\right] + \frac{\sigma}{\sqrt{2\pi}}\left(2+\frac{\mu^2}{\sigma^2}\right).e^{-\frac{\mu^2}{2\sigma^2}} + \frac{\mu}{\sqrt{2\pi}}.\left(\frac{-\mu}{\sigma}\right).e^{-\frac{\mu^2}{2\sigma^2}} - \mu.\sqrt{\pi}.erf\left(\frac{-\mu}{\sqrt{2}.\sigma}\right).$$

Hence, consequently we can write,

$$H^w(X|y) = \log\sqrt{2\pi(1-\rho^2)}\left[\frac{2\sqrt{1-\rho^2}}{\sqrt{2\pi}}.e^{-\frac{(\rho y)^2}{2(1-\rho^2)}} + \rho y\left(\bar{F}_z\left(-\frac{\rho y}{\sqrt{1-\rho^2}}\right) - F_z\left(-\frac{\rho y}{\sqrt{1-\rho^2}}\right)\right)\right] + \frac{\sqrt{1-\rho^2}}{\sqrt{2\pi}}\left(2+\frac{(\rho y)^2}{(1-\rho^2)}\right).e^{-\frac{(\rho y)^2}{2(1-\rho^2)}} + \frac{\rho y}{\sqrt{2\pi}}.\left(\frac{-\rho y}{\sqrt{1-\rho^2}}\right).e^{-\frac{(\rho y)^2}{2(1-\rho^2)}} - \rho y.\sqrt{\pi}.erf\left(\frac{-\rho y}{\sqrt{2}.\sqrt{1-\rho^2}}\right).$$

Besides, using (23.10), straightforward computation shows that,

$$H^{w,\phi'}(X) = \left[\log(\sqrt{2\pi}) + \frac{(\rho y)^2}{2}\right].\left[\frac{2\sqrt{1-\rho^2}}{\sqrt{2\pi}}.e^{-\frac{(\rho y)^2}{2(1-\rho^2)}} + \rho y\left(\bar{F}_z\left(-\frac{\rho y}{\sqrt{1-\rho^2}}\right) - F_z\left(-\frac{\rho y}{\sqrt{1-\rho^2}}\right)\right)\right] + (1-\rho^2).\left[\frac{\sqrt{1-\rho^2}}{\sqrt{2\pi}}(2+\frac{(\rho y)^2}{(1-\rho^2)}).e^{-\frac{(\rho y)^2}{2(1-\rho^2)}} + \frac{\rho y}{\sqrt{2\pi}}.\left(\frac{-\rho y}{\sqrt{1-\rho^2}}\right).e^{-\frac{(\rho y)^2}{2(1-\rho^2)}} - \rho y.\sqrt{\pi}.erf\left(\frac{-\rho y}{\sqrt{2}.\sqrt{1-\rho^2}}\right)\right] + \rho y\sqrt{1-\rho^2}\left[\frac{\sqrt{2}\sqrt{1-\rho^2}}{\sqrt{\pi}}.\left(\frac{-\rho y}{\sqrt{1-\rho^2}}\right).e^{-\frac{(\rho y)^2}{2(1-\rho^2)}} - \sqrt{1-\rho^2}.erf\left(\frac{-\rho y}{\sqrt{2}.\sqrt{1-\rho^2}}\right) + \frac{2\rho y}{\sqrt{2\pi}}.e^{-\frac{(\rho y)^2}{2(1-\rho^2)}}\right].$$

When turning our attention to the influence of dependency between random variables X and Y, we face the situation that with increasing ρ, the weighted information also increases, and what surprises us and is really interesting is the behavior

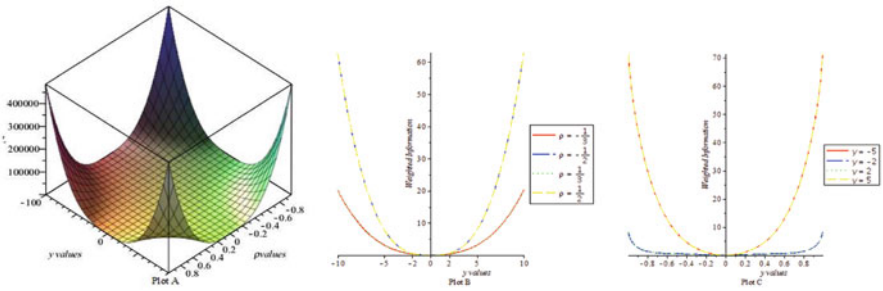

Fig. 23.2 *Plot A*: $I^w(f_{X|y}, f_X)$ and different value y and ρ in normal distribution, *Plot B*: $I^w(f_{X|y}, f_X)$ for different value $\rho = \frac{1}{2}, \frac{1}{3}, \frac{-1}{3}, \frac{-1}{2}$, *Plot C*: $I(f_{X|y}, f_X)$ for different values $y = -5, -2, 2, 5$

of $I^w(f_{X|y}, f_X)$, which is almost the same as $I(f_{X|y}, f_X)$. On the other hand, the weighted information is also an even function with respect to ρ and y. Let us note that in spite of $I(f_{X|y}, f_X)$, $I^w(f_{X|y}, f_X)$ is not always positive, but as you can see in Figs. 23.1 and 23.2, in this special case of normal distribution, both the weighted Kullback–Leilber information and Kullback–Leiber information are positive.

Figure 23.2 shows that the distance between distributions $f_{X|y}$ and f_X increases in terms of different value for ρ such as $\frac{1}{2}$, $\frac{1}{3}$, $\frac{-1}{2}$, and $\frac{-1}{3}$ (see Plot B). Furthermore, in Fig. 23.2 (Plot C), we present the effect of ρ on weighted information for different observations $Y = y$.

This example shows that the weighted Kullback–Leibler information can be a useful method for analyzing the distance between two distribution functions as much as the Kullback–Leibler information, although in the next section, we present that in some cases, weighted information is even more applicable.

23.3 Bayesian Weighted Information Measures

With reference to the notation and the setting outlined in the previous section, we focus on our main interest, which is a pair of random prospects (Θ, Y) that plays the role of (X, Y) of the preceding section. Indeed, the second component, Y, is an observable random variable whose distribution depends on an unknown parameter θ. Considering Y to be a nonnegative random variable representing a lifetime with density function $f_{Y|\theta}$, we observe the random $Y = y$ and evaluate our prior belief about a parameter θ with the prior density function, f_Θ. Therefore, we obtain the posterior distribution having density function $f_{\Theta|y}$ as the following:

$$f_{\Theta|Y}(\theta|y) = \frac{f_{Y|\Theta}(y|\theta) f_\Theta(\theta)}{f_Y(y)}; \quad \theta \in \Theta, y \geq 0.$$

Note that the predictive distribution with the density function $f_Y(y)$ is given by,

$$f_Y(y) = \int f_{Y|\theta}(y|\theta) f_\Theta(\theta) d\theta, \quad y \geq 0.$$

The reason of introducing these distribution functions lies in the fact that we want to find the posterior and predictive information, which is a measure of output information with the knowledge about measure of input information, prior and likelihood information measures. We shall now note that it is possible to compute the weighted uncertainty for prior and posterior density function, i.e., $H(\Theta)$ and $H(\Theta|y)$, respectively, just by replacing the $f_\Theta(\theta)$ and $f_{\Theta|y}(\theta|y)$ in the Eq. (23.7).

Now we present a concept named expected weighted information in the data about the parameter. This constitutes an improvement of the result found by Lindley [8] and Ebrahimi et al. [5]. In fact, we can claim that the expected weighted information precisely is the weighted mutual information of the joint distribution $f(\theta, y)$ and the product distribution $f(\theta)\ f(y)$.

Definition 3 Let $f_{\Theta|y}(\theta|y)$, as we defined before, be a posterior density function, then the expected weighted information is given by,

$$\begin{aligned} \nu^w(\Theta|Y) &= E_Y\left[Y I^w(f_{\theta|y}, f_\theta)\right] = E_Y\left[Y(H^{w,\phi'}(\Theta) - H^w(\Theta|Y))\right] \\ &= H^{w,\phi}(\Theta) - H^w(\Theta|Y), \end{aligned}$$

where E_Y denotes the expectation with respect to the marginal density f_Y, $\Phi'(\Theta) = \frac{\Theta f(\Theta|y)}{f(\Theta)}$ and $\Phi(\Theta) = \Theta E_Y[Y|\Theta]$.

Following the authors in [5], we can recall weighted Kullback–Leibler information in Bayesian analysis, $I^w(f_{\Theta|y}, f_\Theta)$ as,

$$\nu^w(\Theta|y) = H^{w,\phi'}(\Theta) - H^w(\Theta|y).$$

Next example illustrates applications of the Bayesian information measures for Weibull lifetime model.

Example 2 Consider Weibull model for random variables $Y|\theta$ with known positive shape parameter c as the following:

$$f(y|\theta) = c\theta y^{c-1} e^{-\theta y^c}, \quad y > 0, \ \ \theta > 0.$$

Moreover, let the parameter of interest θ has the gamma prior distribution with parameters α and β. Obviously, the posterior distribution is given by gamma distribution with parameters $\alpha + 1$ and $\beta + y^c$.

Now, we shall present the Kullback–Leibler information between prior and posterior distribution by,

$$H(\theta|y) = \log \Gamma(\alpha + 1) - \alpha\Psi(\alpha + 1) - \log(\beta + y^c) + \alpha + 1,$$

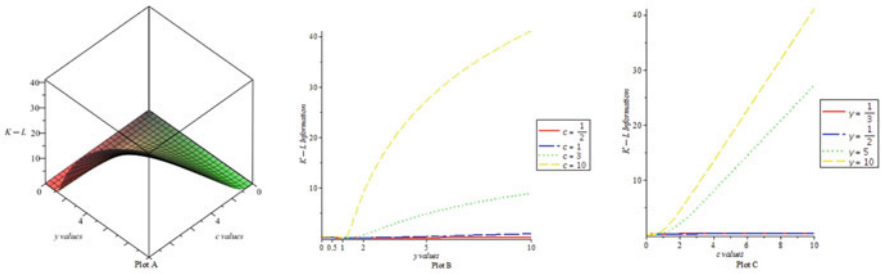

Fig. 23.3 *Plot A*: $I(f_{X|y}, f_X)$ and different value y and c in Weibull distribution, *Plot B*: $I(f_{X|y}, f_X)$ for different value $c = \frac{1}{2}, 1, 3, 10$, *Plot C*: $I(f_{X|y}, f_X)$ for different values $y = \frac{1}{3}, \frac{1}{2}, 5, 10$

therefore,

$$I(f_{\theta|y}, f_\theta) = \alpha \log\left(\beta + y^c\right) - \log\alpha - \alpha \log\beta + \Psi(\alpha+1) + \frac{\beta(\alpha+1)}{(\beta+y^c)} - \alpha - 1. \tag{23.22}$$

Following some computations yield the weighted conditional entropy for posterior distribution as,

$$\begin{aligned} H^w(\Theta|y) = &-\frac{\alpha+1}{\beta+y^c}\log\left[\frac{(\beta+y^c)^{\alpha+1}}{\Gamma(\alpha+1)}\right] - \frac{\alpha(\alpha+1)\Psi(\alpha+2)}{(\beta+y^c)} \\ &+ \frac{\alpha(\alpha+1)}{(\beta+y^c)}\log(\beta+y^c) + \frac{(\alpha+1)(\alpha+2)}{(\beta+y^c)}, \end{aligned} \tag{23.23}$$

where $\Psi(\alpha) = \frac{d}{d\alpha}\log\Gamma(\alpha)$. Note that with the analogous methodology we obtain,

$$\begin{aligned} H^{w,\phi'}(\theta) = &-\frac{\alpha+1}{\beta+y^c}\log\left[\frac{\beta^\alpha}{\Gamma(\alpha)}\right] + \frac{(\alpha-1)(\alpha+1)}{(\beta+y^c)}[\Psi(\alpha+2)] \\ &+ \beta\,\frac{(\alpha+1)(\alpha+2)^2}{(\beta+y^c)^2}. \end{aligned} \tag{23.24}$$

By putting the relations (23.23) and (23.24) in (23.13), we can lay the precise value for the weighted information between posterior and prior distributions as the following,

$$\begin{aligned} I^w(f_{\theta|y}, f_\theta) = &\frac{\alpha+1}{\beta+y^c}\log\left[\frac{(\beta+y^c)^{\alpha+1}}{\alpha\beta^\alpha}\right] + \frac{(\alpha+1)\Psi(\alpha+2)}{(\beta+y^c)} \\ &- \frac{(\alpha+1)}{(\beta+y^c)}\log(\beta+y^c) - \frac{y^c(\alpha+1)(\alpha+2)}{(\beta+y^c)^2}. \end{aligned} \tag{23.25}$$

Referring to Figs. 23.3 and 23.4, we explicitly figure out that $I(f_{\theta|y}, f_\theta)$ is increasing function in y and also c, hence, for any observation, if we choose a model with big value of c, then we would have more information between posterior and prior distributions. As $I^w(f_{\theta|y}, f_\theta)$ is not monotonic with respect to y and c, so for a particular observation y, we can choose an appropriate model in the sense of more information between posterior and prior distributions.

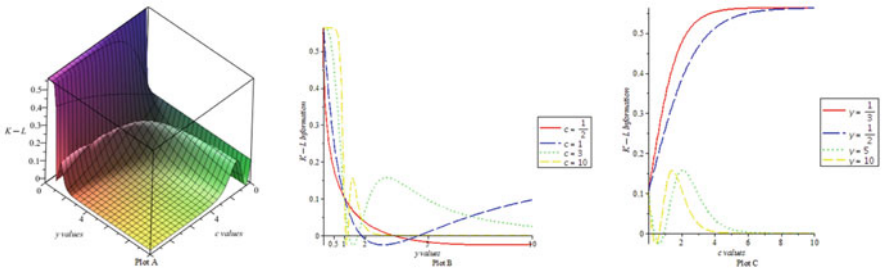

Fig. 23.4 *Plot A*: $I^w(f_{X|y}, f_X)$ and different values y and c in Weibull distribution, *Plot B*: $I^w(f_{X|y}, f_X)$ for different value $c = \frac{1}{2}, 1, 3, 10$, *Plot C*: $I^w(f_{X|y}, f_X)$ for different values $y = \frac{1}{3}, \frac{1}{2}, 5, 10$

23.4 Final Comments

In this chapter, the weighted Kullback–Leibler information measure and conditional weighted entropy have been presented for particular use in the Bayesian analysis. In Sect. 23.2, we have concentrated on the performance of weighted information for normal distributions in order to evaluate and compare its behavior with other common information measure as Kullback–Leibler information. However, the main purpose was to demonstrate that in some cases, such as Bayesian analysis for Weibull distribution, it is more convenient and appropriate to apply weighted information than other common information.

Acknowledgments The author is grateful to the referees for their useful comments. I would like to thank Professor Adriano Polpo for his valuable comments that helped in the improvement of this chapter.

References

1. Belis, M., Guiasu, S.: A quantitative-qualitative measure of lifetime in cybernetic systems. IEEE Trans. Inf. Theory IT **4**, 593–594 (1968)
2. Bernardo, J.M.: Expected information as expected utility. Ann. Stat. **7**, 686–690 (1979)
3. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley, New York (2006)
4. Di Crescenzo, A., Longobardi, M.: On weighted residual and past entropies. Scient. Math. Jpn. **64**, 255–266 (2006)
5. Ebrahimi, N., Kirmani, S.N.U.A., Soofi, E.S.: Dynamic Bayesian information measures. J. Stat. Res. Iran **3**, 113–137 (2006)
6. Guiasu, S.: Grouping data by using the weighted entropy. J. Stat. Plan. Inference **15**, 63–69 (1986)
7. Kullback, S., Leibler, R.A.: On information and sufficiency. Ann. Math. Stat. **22**, 79–86 (1951)
8. Lindley, D.V.: On measure of information provided by an experiment. Ann. Math. Stat. **27**, 986–1005 (1956)
9. Retzer, C.C., Soofi, E.S., Soyer, R.: Information importance of predictors: concepts, measures, Bayesian inference, and application. Comput. Stat. Data Anal. **53**, 2363–2377 (2009)

10. Shannon, C.E.: A mathematical theory of communication. Bell Syst. Tech. J. **27**, 379–423 (1948)
11. Soofi, E.S.: Capturing the intangible concept of information. J. Am. Stat. Assoc. **89**, 1243–1254 (1994)
12. Soofi, E.S.: Principle information theoretic approaches. J. Am. Stat. Assoc. **95**, 1349–1353 (2000)
13. Zellner, A.: An Introduction to Bayesian Inference. Wiley, New York (1971)
14. Zellner, A. Maximal data information prior distributions. In: Aykac, A., Brumat, C. (eds.) New Developments in Application of Bayesian Methods, 211–232. North Holland, Amsterdom (1977)

Chapter 24
Classifying the Origin of Archeological Fragments with Bayesian Networks

Melaine Cristina de Oliveira, Andressa Soreira and Victor Fossaluza

Abstract Classification of archeological fragments is the focus of the present chapter. The fragments were collected from various archeological sites in the state of Mato Grosso do Sul at Lalima village. They are thought to have originated from three Indian tribes: the Guarani (66 %), the Jacadigo (22 %), and the Kadiwéu (12 %).
We use information contained in an archeological researcher's database. It contains qualitative and quantitative observations obtained from the characteristics of the pieces. The researcher's expertise provided a precise classification of about 760 pieces. A supervised model of classification was created to infer the Indian technological traditions of 2300 pieces of fragments collected from the same sites. Bayesian nets were the basis for building the model. Bayesian nets are directed acyclic graphs (DAG) that properly represent the dependency within a set of random covariates. This kind of network represents the joint probability distribution of these variables and a particular factorization of it. Our approach provides a robust classification: it is based on the probabilities of fragment being originated from each one of the three archeological communities. Also, if the probability of technological tradition indicates "low probabilities" for all three groups, there could be an indication of the presence of an additional community. Comparison with alternative methods to build the networks was also presented.

24.1 Introduction

The presence of Indians in the Brazilian territory is much older than was originally established by European explorers and the study of archeology allows the reconstruction of the historical trajectories of populations during their existence. Through the

M. C. de Oliveira (✉) · A. Soreira · V. Fossaluza
IME-USP, São Paulo, Brazil
e-mail: oliveira.mel@gmail.com

A. Soreira
e-mail: dessasoreira@gmail.com

V. Fossaluza
e-mail: victorf@ime.usp.br

A. Polpo et al. (eds.), *Interdisciplinary Bayesian Statistics,*
Springer Proceedings in Mathematics & Statistics 118, DOI 10.1007/978-3-319-12454-4_24

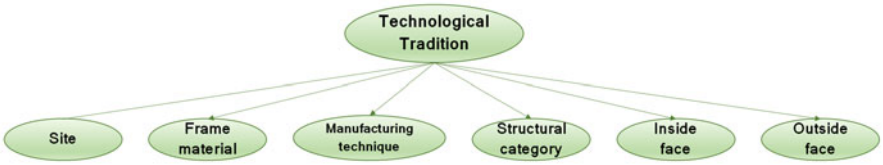

Fig. 24.1 Naive Bayes structure for the ceramic database

characteristics of ceramic objects and historical–cultural knowledge, it is possible to identify which Indian tribes possibly habited a particular region.

Archeological data of ceramic fragments were collected in seven sites from different locations of Lalima village, Mato Grosso do Sul, Brazil, through archeological approaches and with the support of Indians living in that region. Variables relating to composition, shape, and utility of archeological fragments were observed in order to characterize and thus determine whether these working pieces came from one of the following tribes: the Guarani (66 %), the Jacadigo (22 %), or the Kadiwéu (12 %). About 3000 pieces were observed, and using the researcher's expertise, a "precise" classification of 766 pieces into the three technological traditions was accomplished.

The objective of this chapter is to present a model to classify a ceramic fragment in one of the three technological traditions. The classifier method chosen is based on the joint distribution of observed variables. Let T = Technological Tradition; S = Site; IF = Inside Face; OF = Outside Face; MT = Manufacturing Technique; FM = Frame Material; and SC = Structure Category. The joint distribution of a model containing these seven variables can be expressed as

$$P(T,S,FM,MT,SC,IF,OF) = P(T)P(S|T)P(FM|T,S)P(MT|T,S,FM)$$
$$P(SC|T,S,FM,MT)P(IF|T,S,FM,MT,SC)P(OF|T,S,FM,MT,SC,IF).$$

If the seven variables chosen are binaries, then dimension of the sample space is 2^7. In order to reduce dimensionality and have a compact structure of dependence, a simpler and realistic model is investigated. Since the relationship among the variables is unknown, we start modeling with a naive Bayes model that has the unobservable classification variable as the conditioning element that makes the observable variables mutually conditional independent, as illustrated by Fig. 24.1.

Given the structure of the presented network, a possible factorization of the joint distribution used here is as follows:

$$P(T,S,FM,MT,SC,IF,OF) = P(T)P(S|T)P(FM|T)P(MT|T)$$
$$P(SC|T)P(IF|T)P(OF|T)$$

All the variables are independent given T (unobserved variables). This model maybe unrealistic for the data collected. Hence, to contemplate other kinds of dependence among the covariates, more complex Bayesian networks were built by considering the sample fragments already classified by the researcher. The networks obtained were compared and the selected ones were used to allocate the pieces from unknown origins.

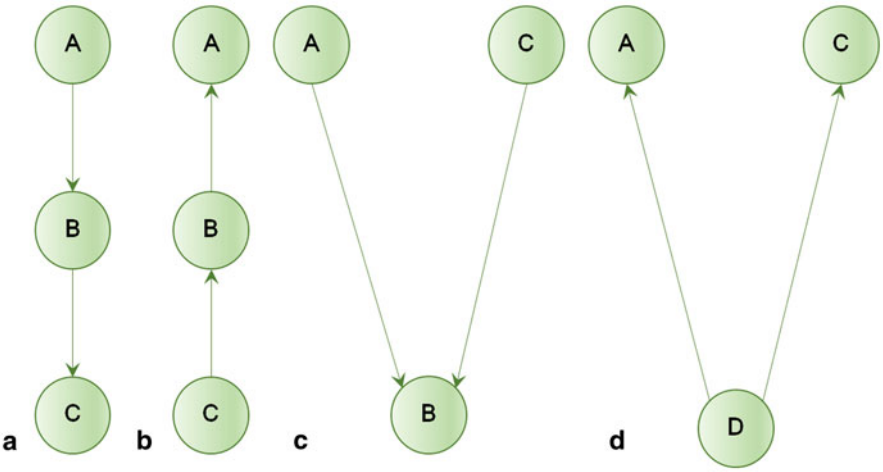

Fig. 24.2 Local structures found on a Bayesian network

Formally, Bayesian networks are directed acyclic graphs (DAG) whose nodes represent random variables, with their conditioning probabilities, in the Bayesian sense: they may be observable quantities, latent variables, unknown parameters, or hypothesis spaces. The edges represent conditional dependencies; nodes that are not connected represent variables that are conditionally independent. Each node is associated with a probability function that takes as input a particular set of values for the node's parent variables and gives the probability of the variable represented by the node given its parents. Figure 24.2 illustrates the structure of Bayesian network with few number of arcs.

We say that A influences C if we have an active path between A and C (nonblocked). If this path is blocked, we say that A and C are d-separated. In the Fig. 24.2a, one can observe that A influences C through B, however if B is to be known, the path is blocked by it and $(C \perp A)|B$. This relationship can also be seen in Fig. 24.2b. Figure 24.2a shows that A is a parent of B and B is parent of C. Figure 24.2b shows that C is a parent of B and B is a parent of A. In Fig. 24.2d, this relation $(A \perp C)|D$ is equally represented, besides B is the parent of A and C. Afterward, in the Fig. 24.2a, b, and d, A and C are d-separated given B. In Fig. 24.2c, we can see that $A \perp C$ if B or any of his ancestors is unknown, but given B, A, and C are dependent. This type of node is known as a collider.

Let $\mathbf{X} = \{X_0, X_1, \ldots, X_m\}$ be a set of variables. We say that X_i is a parent of X_j, $i \neq j$, if there is a link from X_i to X_j. For each i in $0, \ldots, m$, the random variables that are parents of X_i are represented by $pa(X_i)$ and, given $pa(X_i)$, the variable X_i is conditionally independent of its nondescendent variables in $\mathbf{X}$. A Bayesian network Bs over a set of variables $\mathbf{X}$ is a DAG, which represents a joint probability distribution

$$P(\mathbf{X}) = \prod_{i=0}^{m} P(X_i | pa(X_i)) . \tag{24.1}$$

24.2 Methodology: Learning the Bayesian Network

It is difficult to have a full knowledge of conditional independences in the joint distributions of the variables in the study. In these cases, there are some learning algorithms that can find an appropriate—minimum possible number of arcs—Bayesian network Bs given a set of observations from the database $\mathbf{X}$.

The dual nature of a Bayesian network divides the learning process of a Bayesian network in two stages: first, learning the structure of the network, then, learning the probability tables, $P\left(X_i|pa(X_i)\right), i = 0, \ldots, m$.

There are various ways to learn the representation of the joint probability distributions and to choose one, a quality measure considering the network structure was calculated. A quality measure considers the number of variables in the net, the edges between variables, and the model accuracy. Here we consider a *quality measure* $Q\left(B_S|D\right)$ of a network structure B_S given the training data D and looking for the structure that maximizes $Q\left(B_S|D\right)$.

The quality measure used contains the practical properties of attributing a score to the whole network that can be decomposed as the sum (or product) of the score of the individual nodes. This allows for local scoring, which also favors more efficient local search methods.

To describe the methods, we need to introduce some notation. For $i \in \{0, \ldots, m\}$, let r_i be the cardinality of (i.e., the number of values assumed by) X_i. The cardinality of the parent set of X_i in B_S is denoted by q_i, i.e., $q_i = \prod_{\left\{j:\, X_j \in pa(X_i)\right\}} r_j$. Note that $pa(X_i) = \emptyset$ implies $q_i = 1$. For $j \in \{1, \ldots, q_i\}$ and $k \in \{1, \ldots, r_i\}$, we use N_{ijk} to denote the number of records in D for which $pa(X_i)$ takes its jth value and X_i takes its kth value. So, $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$. We use N to denote the number of records in D. Let the *maximum of the log-likelihood function* (sometimes called *entropy metric*) $H\left(B_S, D\right)$ of a network structure B_S and database D be defined as

$$H\left(B_S, D\right) = -N \sum_{i=0}^{m} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} \frac{N_{ijk}}{N} \log \frac{N_{ijk}}{N_{ij}} \tag{24.2}$$

and the *number of parameters* K as

$$K = \sum_{i=1}^{m} (r_i - 1) q_i. \tag{24.3}$$

The quality measures considered in this work are defined below.

Definition 1 The *Akaike information criterion (AIC) metric* [1] $Q_{AIC}\left(B_S, D\right)$ of a Bayesian network structure B_S for a database D is

$$Q_{AIC}\left(B_S, D\right) = H\left(B_S, D\right) + K.$$

Definition 2 The *minimum description length (MDL) metric* [11] $Q_{MDL}(B_S, D)$ of a Bayesian network structure B_S for a database D is defined as

$$Q_{MDL}(B_S, D) = H(B_S, D) + \frac{K}{2} \log N.$$

Definition 3 The *Bayesian metric* [10] $Q_{Bayes}(B_S, D)$ of a Bayesian network structure B_S for a database D is

$$Q_{Bayes}(B_S, D) = P(B_S) \prod_{i=0}^{m} \prod_{j=1}^{q_i} \frac{\Gamma(N'_{ij})}{\Gamma(N'_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(N'_{ijk} + N_{ijk})}{\Gamma(N'_{ijk})}$$

where $P(B_S)$ is the prior on the network structure (taken to be constant here) and $\Gamma(.)$ the gamma function. N'_{ij} and N'_{ijk} represent the priors for the counts, restricted by $N'_{ij} = \sum_{k=1}^{r_i} N'_{ijk}$. With $N'_{ijk} = 1$ (and thus $N'_{ij} = r_i$), we obtain the *K2 metric*.

In the present work, we will consider two methods to obtain the Bayesian network structure. The first one is the *Hill-Climbing* algorithm [4] that adds and deletes arcs with no fixed ordering of variables. At each iteration, it will adjust a single arc in B_S and determine whether the changes improve the value of $Q(B_S, D)$. Any change that improves $Q(B_S, D)$ is accepted, and the process continues until no other change can be found to improve the value of $Q(B_S, D)$. The second method is called *TAN* (Tree Augmented Naive Bayes) [6, 8], where the tree is formed by calculating the maximum weight spanning tree using Chow and Liu algorithm [7]. Both methods were implemented using the software WEKA [9].

24.3 Results and Discussion

After an intensive descriptive analysis, the variables that best contributed to the distinction between three classifications of archeological fragments were selected and their categorizations were defined for the database of 760 pieces. This database was used to find the best representation for a structure of data (relation of dependency between the variables). To avoid overfitting, we used 10 % of the data for cross-validation and the adjustment measures in Tables 24.1 and 24.2 come from this subsample.

First, a model (saturated) with 13 variables that were shown relevant to infer the classification of the pieces were constructed. After the modeling process, only the most relevant variables (6) stay in the final model (reduced). We can perceive looking at Tables 24.1 and 24.2 that the final model has a similar accuracy to the saturated model, showing that the more parsimonious model has similar adjustments.

Table 24.1 presents the fit of two implemented Bayesian networks through the learning method *TAN*: saturated and reduced model, with the use of four different quality measures. Similarly, Table 24.2 presents these models implemented through the *Hill-Climbing* algorithm.

Table 24.1 Statistics fit measures for the *Hill-Climbing* saturated and reduced models ($n = 766$)

	Saturated models				Reduced models			
Quality measures	MDL	Bayes	AIC	Entropy	MDL	Bayes	AIC	Entropy
Bayes	−5803	−5610	−5619	−5648	−3280	−3176	−3182	−3204
Bdeu	−6312	−6934	−6987	−7237	−3445	−3898	−3951	−4251
MDL	−6302	−6840	−6881	−7101	−3466	−3845	−3894	−4154
Entropy	−5767	−5698	−5699	−5759	−3250	−3230	−3240	−3300
AIC	−5928	−6042	−6055	−6163	−3315	−3415	−3437	−3557
Correctly classified instances	750	749	748	750	757	753	751	750
Kappa statistic	0.96	0.96	0.95	0.96	0.98	0.97	0.96	0.96
Relative absolute error	6.20	6.17	7.34	7.09	5.88	6.19	7.17	7.00
Guarani (FP)	0.004	0.019	0.015	0.015	0.004	0.019	0.015	0.012
Jacadigo (FP)	0.017	0.013	0.018	0.017	0.013	0.010	0.013	0.018
Kadiwéu (FP)	0.007	0.006	0.004	0.003	0.000	0.003	0.004	0.003
Average FP rate	0.007	0.016	0.015	0.014	0.005	0.015	0.014	0.012

FP false positive rate

Observing the two tables of fitted models, adding the knowledge about the problem, and having an expected relation between the variables, we decided to use the reduced model with the MDL metric as the quality measure. Thus, Figs. 24.3 and 24.4 show two Bayesian networks considering the ceramic pieces already classified by the researcher using the *TAN* and *Hill-Climbing* methods, respectively.

Another objective of developing the above models, attributing probabilities belonging to one of the three technological traditions (Guarani, Jacadigo, or Kadiwéu), is to help the researcher understand better the origin of the pieces that (s)he was unable to classify. As it can be seen, there is no quantitative reason to prefer one of the two used methods, since both present similar and satisfactory fits. However, the Bayesian network provides information about dependency and the researcher can choose the model that best represents reality from the point of view of an expert. We can observe that some dependency relationships are maintained in both models, but others differ. Two Bayesian networks are said to be equivalent if their joint probabilities are equal [5]. So, we can say that the Bayesian networks produced through the methods *TAN* and *Hill-Climbing* are not equivalent.

In the Fig. 24.3, we have the following joint distribution:

$$\begin{aligned} P_{TAN}(T,S,IF,OF,SC,MT,FM) = {} & P(T)P(S|T)P(IF|MT,T) \\ & P(FM|S)P(OF|MT,T)P(MT|FM,T) \\ & P(SC|MT,T). \end{aligned}$$

Table 24.2 Statistics fit measures for the *TAN* saturated and reduced models ($n = 766$)

	Saturated models				Reduced models			
Quality measures	MDL	Bayes	AIC	Entropy	MDL	Bayes	AIC	Entropy
Bayes	−5774	−5633	−5650	−5689	−3252	−3186	−3191	−3217
Bdeu	−6485	−6994	−6945	−7471	−3581	−3931	−3942	−4259
MDL	−6452	−6892	−6834	−7284	−3587	−3876	−3884	−4160
Entropy	−5739	−5720	−5711	−5823	−3232	−3222	−3250	−3307
AIC	−5954	−6073	−6049	−6263	−3339	−3419	−3441	−3564
Correctly classified instances	751	750	748	746	758	752	753	748
Kappa statistic	0.96	0.96	0.95	0.95	0.98	0.96	0.97	0.95
Relative absolute error	6.71	6.66	6.84	7.48	6.06	6.72	6.53	7.44
Guarani (FP)	0.008	0.015	0.015	0.027	0.008	0.019	0.015	0.015
Jacadigo (FP)	0.017	0.015	0.017	0.018	0.010	0.012	0.012	0.020
Kadiwéu (FP)	0.004	0.004	0.006	0.003	0.000	0.003	0.003	0.003
Average FP rate	0.009	0.014	0.015	0.022	0.007	0.016	0.013	0.015

FP false positive rate

For Fig. 24.4, we have the following joint distribution:

$$P_{HC}(T, S, IF, OF, SC, MT, FM) = P(T)P(S|FM, T)P(IF|T)P(FM|T) \\ P(OF|T)P(MT|T)P(SC|T).$$

To check the dependency relationships presented in Bayesian network, it is necessary to pay attention to its structure, shown in Figs. 24.3 and 24.4.

- For the Bayesian network built using the Hill-Climbing algorithm (Fig. 24.3), we can see that:
  - *Inside face*, *Outside face*, *Structural category*, *Manufacturing technique*, and *Frame material* are mutually independent, given *Technological Tradition*;
  - *Site* is independent of all other variables, given *Frame material* and *Technological Tradition*;
- For the Bayesian network built using the TAN method (Fig. 24.4), we can see that:
  - *Site* depends only on *Technological Tradition*;
  - *Inside face*, *Outside face*, and *Structural category* are mutually independent, given *Manufacturing technique* and *Technological Tradition*;
  - *Frame material* depends only on *Site*;
  - *Frame material* and *Technological Tradition* are parents of *Manufacturing technique*.

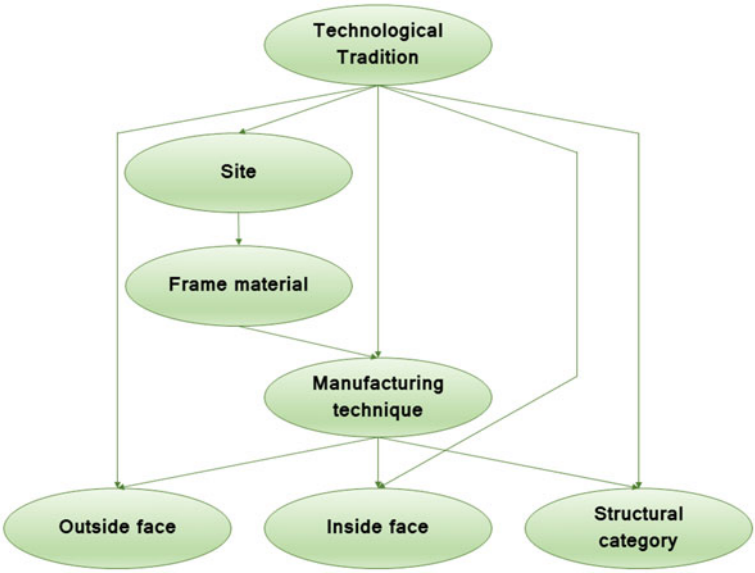

Fig. 24.3 Reduced Bayesian network model built by *TAN* method considering the MDL metric. This Bayesian network was built using the database of ceramic pieces already classified by the researcher to learning distribution structure

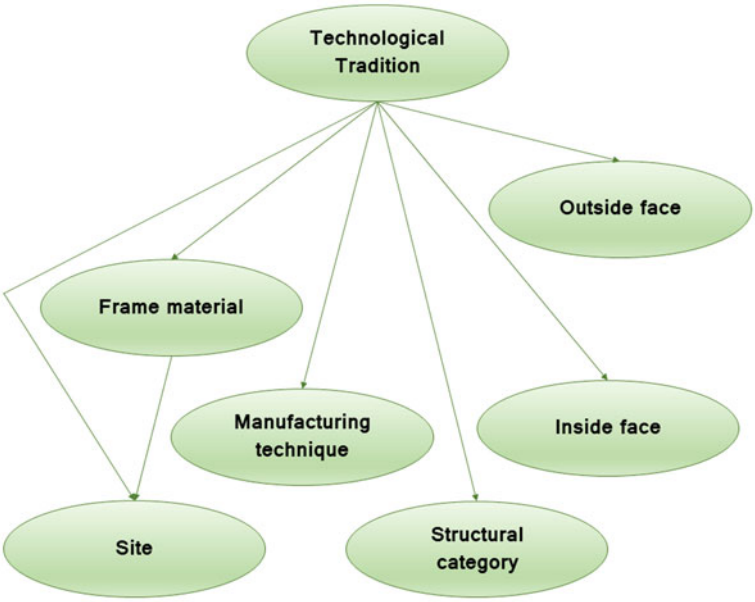

Fig. 24.4 Reduced Bayesian Network model built by *Hill-Climbing* algorithm considering the MDL metric. This Bayesian network was built using the database of ceramic pieces already classified by the researcher to learning the distribution structure

Table 24.3 Classification by *TAN* and *Hill-Climbing* reduced models ($n = 2320$)

Hill-Climbing classification	TAN classification				
	Guarani (%)	Jacadigo (%)	Kadiwéu (%)	NC (%)	Total
Guarani	6.6	0.4	2.1	0.0	9.1 %
Jacadigo	0.3	58.3	0.0	0.0	58.5 %
Kadiwéu	3.0	0.0	27.9	0.1	31.0 %
NC	0.0	0.0	1.4	0.0	1.4 %
Total	9.8	58.7	31.4	0.1	2,320

NC nonclassified

Table 24.3 shows that the classification of 2320 fragments from unknown origins is very similar for both models. There are 93 % of the pieces classified in the same Technological Tradition. We consider *NC* (nonclassified) when the model attribute to the fragment less than 50 % of probability belongs to one of the Technological Traditions.

Additional Remarks: For the basic theory of Bayesian network, we refer to Barlow and Pereira [3] and for hypotheses tests that reduce the number of arcs, the reference could be Andrade et al. [2].

Acknowledgements The authors gratefully acknowledge ISBrA (the Brazilian chapter of the International Society for Bayesian Analysis) and FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo) for financial support, CEA-IME-USP (Center of Applied Statistics of the Institute of Mathematics and Statistics from Universidade de São Paulo) for providing the database. Thanks to Prof. Carlos A. B. Pereira, Pablo M. Andrade, and Rafael Izbicki for helpful comments.

References

1. Akaike, H.: A new look at the statistical model identification. IEEE Trans. Autom. Control **19**(6), 716–723 (1974)
2. Andrade, P.M., Stern, J.M., Pereira, C.A.B.: Bayesian test of significance for conditional independence: the multinomial model. Entropy **16**(3), 1376–1395 (2014)
3. Barlow, R.E., Pereira, C.A.B.: Conditional independence and probabilistic influence diagrams. In: Block, H.W., Sampson, A.R., Savits, T.H. (eds.) Topics in Statistical Dependence, pp. 19–44. IMS, Pittisburgh (1990)
4. Buntine, W.L.: A guide to the literature on learning probabilistic networks from data. IEEE Trans. Knowl. Data Eng. **8**, 195–210 (1996)
5. Cheng, J., Bell, D.A., Liu, W.: Learning belief networks from data: an information theory based approach. In Proceedings of the Sixth International Conference on Information and Knowledge Management, pp. 325–331. ACM. (1997, January)
6. Cheng, J., Greiner, R.: Comparing Bayesian network classifiers. Proceedings UAI, pp. 101–107. Sweden (1999)
7. Chow, C.I., Liu, C.N.: Approximating discrete probability distributions with dependence trees. IEEE Trans. Inf. Theory **14**(3), 462–467 (1968)

8. Friedman, N., Geiger, D., Goldszmidt, M.: Bayesian network classifiers. Mach. Learn. **29**, 131–163 (1997)
9. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. ACM SIGKDD Explor. Newsl. **11**(1), 10–18 (2009)
10. Koller, D., Friedman, N.: Probabilistic Graphical Models: Principles and Techniques. MIT Press, Cambridge (2009)
11. Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **6**(2), 461–464 (1978)

Chapter 25
A Note on Bayesian Inference for Long-Range Dependence of a Stationary Two-State Process

Plinio L. D. Andrade and Laura L. R. Rifo

Abstract In this work we propose a Bayesian approach for selecting the range of a stationary process with two states. The analysis is based on approximate posterior distributions of the Hurst index obtained from a likelihood-free method. Our empirical study shows that a main advantage of our approach, along with its of simplicity, is the possibility of obtaining an approximate sample of the posterior distribution on the Hurst index, thus providing better estimates. Furthermore, there is no need for Gaussian nor asymptotic assumptions.

25.1 Introduction

The problem of estimating the range of a binary process has been extensively studied and its application for long-range phenomena can be found in several fields [6, 10, 15]. The central point is the fact that we are not dealing with a Gaussian process, and estimates based on this assumption are not accurate. We propose a simple and intuitive Bayesian approach to estimate such range based on a likelihood-free methodology. The advantages of our proposal are its simplicity, the fact that it provides an approximate sample from the posterior distribution on the memory parameter and, mainly, it can be used straightforwardly in Gaussian and non-Gaussian long-memory estimation problems.

25.2 Framework

This section defines long-memory and reviews some statistics that will be used in our analysis.

P. L. D. Andrade (✉)
Institute of Mathematics and Statistics, University of São Paulo, São Paulo, Brazil
e-mail: plinio@ime.usp.br

L. L. R. Rifo
Institute of Mathematics and Statistics, University of Campinas, Campinas, Brazil
e-mail: lramos@ime.unicamp.br

A. Polpo et al. (eds.), *Interdisciplinary Bayesian Statistics,*
Springer Proceedings in Mathematics & Statistics 118, DOI 10.1007/978-3-319-12454-4_25

Definition 1 Let $X = \{X_t, t \in \mathbb{Z}\}$ be a second-order stationary process with autocorrelation function $\rho(n)$. X is said to exhibit long memory or long-range dependence if, for some H

$$\liminf_{n\to\infty} \frac{\rho(n)}{n^{2H-2}} > 0.$$

H is called a memory parameter, memory index, or Hurst index (see [7, 21]). The long-range memory is expressed by values of H such that $0.5 < H < 1$, where $H = 0.5$ indicates a short-memory process, and as H increases, so does the memory effect.

In order to estimate H we consider three well-known approaches, the ideas of which will be given in the next subsections: *R/S analysis*, the *fully extended local Whittle* (FELW) analysis, and the *wavelet* analysis. We compare these three methods with our proposed Bayesian method.

25.2.1 R/S Analysis

In this subsection we review a heuristic method to estimate the Hurst index. This method is recommended to detect long-range dependence rather than concrete estimation (see the introduction in [7]).

Let us take an observed sample path $\mathbf{x} = (x_1, \ldots, x_n)$ from a second-order stationary time series $X = \{X_t, t \in \mathbb{Z}\}$.

The rescaled range statistic R/S, introduced by [16], is given by

$$R/S(n) = \frac{1}{\hat{s}_n}\left[\max_{1\le k\le n}\sum_{i=1}^{k}(x_i - \bar{x}_n) - \min_{1\le k\le n}\sum_{i=1}^{k}(x_i - \bar{x}_n)\right],$$

where $\bar{x}_n$ and $\hat{s}_n^2$ are the sample mean and sample variance, respectively. It is known that

$$\frac{1}{n^H} R/S(n) \xrightarrow{D} C_H^{R/S} \quad (n \to \infty),$$

where $C_H^{R/S}$ is a nondegenerate random variable and $(\xrightarrow{D})$ means convergence in distribution. Thus, we have

$$\log R/S(n) \approx H \log n + \log C_H^{R/S}.$$

Therefore, the estimator $\hat{H}_{R/S}$ of H is the linear coefficient of the regression of $R/S(n)$ on n in loglog scale.

25.2.2 Whittle's Method

Consider a second-order stationary process $X = \{X_t, t \in \mathbb{Z}\}$ observed at times $t = 1, \ldots, n$ and assume that the spectral density has the following form

$$f(\lambda) = |\lambda|^{1-2H} w(\lambda), \quad \lambda \in [-\pi, \pi],$$

where $w(\lambda) \to b \in (0, \infty)$, as $|\lambda| \to 0$.

When the spectral density above is correctly specified by a finite dimensional parameter θ, that is, $w(\lambda) = w(\lambda, \theta)$, then under regularity conditions the parameters H and θ can be consistently estimated by the parametric Whittle estimator (see [5, 7] for a detailed exposition). The Whittle estimator is the value $\hat{\eta}$ that minimizes

$$Q(\eta) = \sum_{j=1}^{[(n-1)/2]} \frac{I(\lambda_j)}{g(\lambda_j; \eta)}$$

where, $I(\lambda_j)$ is the periodogram at Fourier frequencies $\lambda_j = 2\pi j/n$, $[\alpha]$ denotes the integer part of α, and $g(\lambda; \eta)$ is a renormalization of the spectral density function $f(\lambda; \eta)$ such that $\int_{-\pi}^{\pi} \log f(\lambda; \eta) d\lambda = 0$. When dealing with the increments of the *fractional Brownian motion* or *fractional*-ARIMA$(0, d, 0)$, η is simply the parameter H or $d = H - 1/2$ respectively, and for *fractional*-ARIMA(p, d, q), η has autoregressive and moving average parameters.

An empirical study [23] shows that for such processes the Whittle estimator is the best among several estimators. Such method assumes that the parametric form of the spectral density is known, which is not common in practice and that is why some authors consider semiparametric approaches based on the Whittle method.

The local Whittle approach to estimate H requires only the knowledge of $f(\lambda)$ near the origin, it is robust with respect to model specification and may be applied to a wider class of processes. The first relevant work on this topic is due to Robinson [20]. In the simulation study, we consider the FELW estimator detailed in [1]. The FELW estimator is applicable not only for traditional cases but also for nonlinear and non-Gaussian processes with long memory, and the variation interval for H is extended to $-1 < H < \infty$ in order to make possible its application in some important processes, which appear in financial econometrics such as stochastic volatility studies.

25.2.3 Wavelet Analysis

The wavelet approach for the estimation of H was first proposed by [2] and several results concerning this estimator were developed since then (see [3] and references therein).

Let $X = \{X_t, t \in \mathbb{R}\}$ be a stationary long-memory process and $X_1, \ldots, X_n$, a sample path from this process. We call a *mother wavelet* any continuous function

$\psi : \mathbb{R} \to \mathbb{R}$ with support on the interval $[0, 1]$ satisfying

$$\int_{\mathbb{R}} t^p \psi(t) dt = 0 \quad (p = 0, 1, \ldots, \mathrm{Q} - 1)$$

and

$$\int_{\mathbb{R}} t^{\mathrm{Q}} \psi(t) dt \neq 0,$$

where $\mathrm{Q} \geq 1$ is an integer called the *number of vanishing moments*. For a *scale* $a \in \mathbb{N}^*$, the wavelet coefficient of X is given by

$$d(a, i) = \frac{1}{\sqrt{a}} \int_{\mathbb{R}} \psi \left(\frac{t}{a} - i \right) X_t dt,$$

for $i = 1, 2, \ldots, N_a$ with $N_a = [n/a] - 1$. Now, for (a, b), we can define the approximate wavelet coefficient of $d(a, b)$ as

$$e(a, b) = \frac{1}{\sqrt{a}} \sum_{k=1}^{n} \psi \left(\frac{k}{a} - b \right) X_k.$$

Under appropriate regularity conditions and large a, it can be shown that $\mathrm{var}[d(a, i)]$ is a power-law function of a with exponent $2H - 1$. Thus, if we consider the following statistic

$$V_n(a) = \frac{1}{N_a} \sum_{i=1}^{N_a} e^2(a, i),$$

then an estimator $\hat{H}_{Wa}$ of H can be obtained by a loglog regression of $V_n(a)$ over a.

25.3 ABC Method

When dealing with a nonstandard posterior distribution, we usually use Monte Carlo simulation (or, more specifically, Markov chain Monte Carlo (MCMC) methods) to produce a sample from it.

However, there are cases where the likelihood function is untractable and MCMC methods cannot be implemented. The class of likelihood-free methods termed *Approximate Bayesian Computation* (ABC) addresses this issue as long as we are able to simulate from the model and a suitable summary statistic is available. The ABC idea was proposed by [19], among many others.

In our context, given a sample $\mathbf{x} = (x_1, \ldots, x_n)$ associated with some distribution $f(\cdot \mid \theta)$, a summary statistic T, and a prior π for θ, the simplest ABC algorithm is described as follows:

ABC rejection sampler

Step 1 generate θ_j from the prior π;
Step 2 generate $\mathbf{y} = (y_1, \ldots, y_n)$ from $f(\cdot \mid \theta_j)$;
Step 3 compute the distance $|T(\mathbf{y}) - T(\mathbf{x})|$;
Step 4 accept θ_j if $|T(\mathbf{y}) - T(\mathbf{x})| \leq \epsilon$, otherwise go back to Step 1.

The main idea is that the summary statistic T coupled with a small value of ϵ should provide a good approximation of the posterior distribution for θ.

The accuracy of this approximation will depend on the choice of the statistic T and a suitable value of ϵ (see [14] and the references therein). When T is a sufficient statistic and ϵ is small enough, the accepted values will form a sample from the posterior distribution for θ. Unfortunately, this optimal situation rarely occurs in practice. Regarding the choice of ϵ, [14] proposes to choose ϵ as the 1 % quantile of the simulated distances by repeating the Step 3 of the ABC algorithm, i.e., we can previously fix the number of accepted values to $N_{sim} \times (1\,\%)$, where N_{sim} is the number of simulations. We use that value for ϵ in our simulation study.

When looking for a good statistic to use in the ABC algorithm, we are motivated by a definition of sufficiency derived from information theory (see [12]).

Definition 2 A function $T(X)$ is said to be a sufficient statistic relative to the family $\{f(x|\theta); \theta \in \Theta\}$ if and only if, $I(\theta; T(X)) = I(\theta; X)$, where $I(X, Y)$ is the *mutual information* between X and Y, defined by

$$I(X; Y) = \mathbb{E}_{P_{XY}}\left(\log \frac{P_{XY}}{P_X P_Y}\right),$$

where $\mathbb{E}_P$ is the expected value over the distribution P, P_{XY} is a joint probability distribution of X and Y, while P_X and P_Y are their marginals.

In general if $T(X)$ is any function of the sample, we have $I(\theta; T(X)) \leq I(\theta; X)$. Definition 2 suggests that a good statistic for the ABC procedure is the one that is close to sufficiency in an informative sense. In other words, we seek a statistic $T(X)$ that maximizes the mutual information $I(\theta; T(X))$. It can be easily shown that

$$I(\theta; T(X)) = \mathbf{h}(\theta) - \mathbf{h}(\theta|T(X)), \tag{25.1}$$

where $\mathbf{h}(X)$ is the entropy of X, $\mathbf{h}(X) = -\mathbb{E}_{P_X}(\log P_X)$.

From the inequality above and (25.1), we see that maximizing mutual information $I(\theta; T(X))$ is equivalent to minimizing $\mathbf{h}(\theta|T(X))$. We can use this minimization criterion to select the nearly optimal statistic from a set of available statistics. The estimation of entropy is a well-developed field and there are many sample-based estimators that can be readily applied to the sample obtained by the ABC algorithm (see [4, 17, 22]).

The ABC procedure is now summarized as follow: Given a set $S = \{S_1, S_2, \ldots, S_k\}$ of summary statistics, we replace Steps 3 and 4 in the ABC algorithm by

Step 3' compute the distances $|S_i(\mathbf{y}) - S_i(\mathbf{x})|$, $i = 1, 2, \ldots, k$;

Step 4' accept H_j in the sample i if $|S_i(\mathbf{y}) - S_i(\mathbf{x})| < \epsilon_i$, otherwise go back to step 1, and a new step is added.

Step 5 For each sample i obtained, compute $\hat{\mathbf{h}}_i$ $(i = 1, 2, \ldots, k)$ and choose S_l in S such that

$$\hat{\mathbf{h}}_l = \min_{1 \le i \le k} \{\hat{\mathbf{h}}_i\}.$$

Such a sample provides an approximate sample from the posterior distribution for θ.

The estimator $\hat{H}_{ABC}$ of H obtained from the ABC algorithm will be considered here as the mean of the approximate posterior distribution.

25.3.1 Summary Statistics

In order to use ABC algorithm, we need to choose a set of summary statistics that provides some information about the parameters of interest in the posterior analysis.

In the context of memory estimation problems, one can consider the R/S statistic presented in Sect. 25.2.1. We can see that the statistic $\log R/S(N)/\log(N)$ is related to the Hurst index H, so that this statistic is a natural candidate to be included in the set of summary statistics for an ABC algorithm. There are other statistics used to estimate the memory parameter heuristically [13, 18], but the results obtained are similar to those with the R/S statistic.

We also consider another set of statistics of the quadratic variations family. Such statistics are used to estimate the memory index in self-similar processes as proposed in [9], defined for filtered processes.

Definition 3 A filter f of length $l \in \mathbb{N}$ and order $p \in \mathbb{N}^*$ is an $(l+1)$-dimensional vector $f = \{f_0, f_1, \ldots, f_l\}$ satisfying

$$\sum_{q=0}^{l} f_q q^r = 0, \quad \text{for } 0 \le r \le p-1, \quad r \in \mathbb{Z},$$

$$\sum_{q=0}^{l} f_q q^p \neq 0,$$

with the usual convention $0^0 = 1$.

Definition 4 Assuming that we observe the process in discrete times $\{1, \ldots, n\}$, a filtered process $\mathbf{X}(f)$ is defined as

$$\mathbf{X}(f) := \sum_{q=0}^{l} f_q X_{i-q}, \quad \text{for } i = l+1, \ldots, n,$$

and the quadratic variation of $\mathbf{X}(f)$ is given by

$$S_n(f) = \frac{1}{\text{n-l}} \sum_{i=l+1}^{n} [\mathbf{X}(f)]^2 = \frac{1}{\text{n-l}} \sum_{i=l+1}^{n} \left(\sum_{q=0}^{l} f_q X_{i-q} \right)^2 .$$

In [8], it is observed that a logarithmic transformation of the summary statistics can improve the performance of the ABC algorithm. So, in addition to the statistics considered above, we will consider transformations such as logarithm, square root, and reciprocal of these statistics, and include them in the set of summary statistics.

25.4 Results

For the simulation study, we obtained a sample from each estimator defined in the previous section, as follows:

- We considered three nominal values for the Hurst index H, namely 0.5, 0.7, and 0.9.
- For each nominal value of H, we simulated 1000 sample paths with length 1000 of the long-memory stationary two-state process taking values in $\{0, 1\}$. The paths were simulated through a dichotomization of a sample path from the fractional Brownian motion, obtained using the circulant matrix procedure [11].
- Each method presented previously was applied to each of those data.

For the ABC algorithm, we created a database of 100,000 simulated paths under a uniform prior distribution over $(0, 1)$. The set of summary statistics, S, includes $\log R/S(N)/\log(N)$, the quadratic variation statistic based on the finite difference filters of orders 1–5, as well as the transformations of those statistics (square root, logarithm, and reciprocal).

For each estimation method, we obtained 1000 estimated values of H, say $\hat{\mathbf{H}}_M = \{H_i, i = 1, \ldots, 1000\}$, where M stands for the method (R/S, FELW, Wavelet, or ABC). We have computed their mean and standard deviation. The results are given in Table 25.1.

Figure 25.1 shows the boxplots of the deviations from the nominal value of H for each estimation method. The letters A, B, C, and D in each graph stand for the estimators R/S, FELW, Wavelet, and ABC, respectively.

As expected, the performance of the R/S analysis has high variability, sustaining that heuristic approaches are only recommended as a first visual inspection and detection of long-range dependence in the data. The performance of the ABC method is remarkable for all nominal values, especially when $H = 0.9$.

With respect to the minimum entropy criterion, when the nominal value is 0.5, the quadratic variation statistic with a finite difference filter of order 1 without any transformation is selected in 98.3 % of all simulations. For $H = 0.7$, the same statistic is selected in 98.6 % of the cases, while for $H = 0.9$, it is selected in 69.1 %

Table 25.1 Estimates for H using 1000 independent realizations from the stationary 0–1 process with $n = 1000$ steps

Estimation method	Nominal H	0.5	0.7	0.9
R/S	Mean $\hat{\mathbf{H}}_{R/S}$	0.5438	0.6781	0.8269
	Std $\hat{\mathbf{H}}_{R/S}$	0.0852	0.1043	0.1303
Fully extended local Whittle	Mean $\hat{\mathbf{H}}_{FELW}$	0.4974	0.6563	0.8443
	Std $\hat{\mathbf{H}}_{FELW}$	0.0582	0.0596	0.0584
Wavelet	Mean $\hat{\mathbf{H}}_{Wa}$	0.4816	0.6363	0.8033
	Std $\hat{\mathbf{H}}_{Wa}$	0.0613	0.0592	0.0649
Posterior mean	Mean $\hat{\mathbf{H}}_{ABC}$	0.4965	0.6956	0.8989
	Std $\hat{\mathbf{H}}_{ABC}$	0.0362	0.0316	0.0338

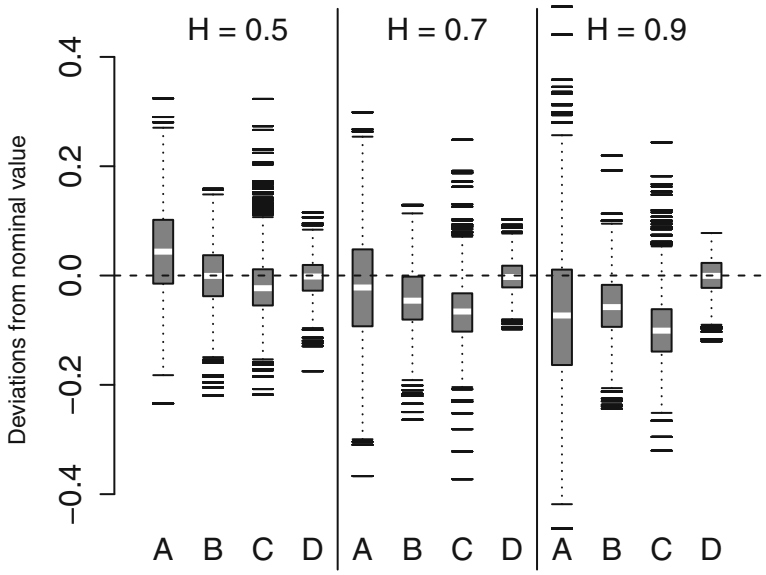

Fig. 25.1 Boxplots of the deviations from the true value based on 1000 independent realizations of the long-range dependent two-state process for each H. The letters A, B, C and D stand for the estimators R/S, FELW, Wavelet and ABC respectively

of all simulations. In 17.9 %, the selected statistic is the quadratic variation with a finite difference filter of order 2. We see that using a higher than two-order filter or the R/S statistic does not improve the performance of the ABC algorithm in this particular case.

25.5 Discussion

We believe that it is possible to make simple and efficient inferences about the memory index of a process using Bayesian techniques. We see through simulations that our proposed approach provides estimates that are pretty close to the nominal values and have had low variability compared to its frequentist competitors. The approach presented here can be applied to more general contexts and can be extended to a wider class of memory estimation problems. The only restriction is the availability of good statistics and goods models that may be simulated. Moreover, high-order difference filters as well as wavelet filters could be considered depending on the nature of the problem.

Acknowledgements The first author is a PhD student with CNPq grant at the University of São Paulo. For the second author, this work was produced as part of the activities of FAPESP Center for Neuromathematics (grant 2013/ 07699-0 , S. Paulo Research Foundation).

References

1. Abadir, K.M., Distaso, W., Giraitis, L.: Nonstationarity-extended local Whittle estimation. J. Econom. **141**, 1353–1384 (2007)
2. Abry, P., Veitch, D.: Wavelet analysis of long-range-dependent traffic. IEEE Trans. Inf. Theory **44**(1), 2–15 (1998)
3. Bardet, J.-M., Bibi, H.: Adaptive semiparametric wavelet estimator and goodness-of-fit test for long memory linear processes. Electron. J. Stat. **6**, 2383–2419 (2012)
4. Beirlant, J., Dudewicz, E.J., Györfi, L., van der Meulen, E.C.: Nonparametric estimation of entropy: an overview. Int. J. Math. Stat. Sci. **6**, 17–39 (1997)
5. Beran, J.: Statistics for Long-Memory Process. Chapman & Hall, New York (1994)
6. Beran, J., Sherman, R., Taqqu, M.S., Willinger, W.: Long-range dependence in variable-bit-rate video traffic. IEEE Trans. Commun. **43**(234), 1566–1579 (1995)
7. Beran, J., Feng, Y., Ghosh, S., Kulik, R.: Long-Memory Process—Probabilistic Properties and Statistical Methods. Springer, New York (2013)
8. Blum, M.G.B.: Choosing the summary statistics and the acceptance rate in the approximate Bayesian computation. COMPSTAT 2010 Proceedings in Computational Statistics, pp. 47–56 (2010)
9. Chronopoulou, A., Viens, F.G.: Hurst index estimation for self-similar processes with long-memory. Recent Adv. Stoch. Dyn. Stoch. Anal. **1**, 85–112 (2009)
10. Churchill, G.A.: Hidden Markov chains and the analysis of genome structure. Comput. Chem. **16**(2), 107–115 (1992)
11. Coeurjolly, J.-F.: Simulation and identification of the fractional Brownian motion: a bibliographic and comparative study. J. Stat. Softw. **5**, 1–53 (2001)
12. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley-Interscience, New York (2006)
13. Giraitis, L., Kokoszka, P., Leipus, R., Teyssiere, G.: Rescaled variance and related testes for long memory in volatility and levels. J. Econom. **112**(2), 265–294 (2003)
14. Grelaud, A., Robert, C.P., Marin, J.M., Rodolphe, F., Taly, J.F.: ABC likelihood-free methods for model choice in Gibbs random fields. Bayesian Anal. **4**(2), 317–335 (2009)
15. Heath, D., Resnick, S., Samorodnitsky, G.: Heavy tails and long range dependence in on/off processes and associated fluid models. Math. Oper. Res. **23**(1), 145–165 (1998)

16. Hurst, H.: Long term storage capacity of reservoirs. Trans. Am. Soc. Civ. Eng. **116** 770–799 (1951)
17. Kozachenko, L.F., Leonenko, N.N.: Sample estimates of entropy of a random vector. Probl. Inf. Transm. **23**, 95–101 (1987)
18. Kwiatkowski, D., Phillips, P.C.B., Schmidt, P., Shin, Y.: Testing the null hypothesis of stationarity against the alternative of a unit root: how sure are we that economic time series have a unit root? J. Econom. **54**, 159–178 (1992)
19. Pritchard, J.K., Seielstad, M.T., Perez-Lezaun, A., Feldman, M.W.: Population growth of human Y chromosomes: a study of Y chromossome microsatellites. Mol. Biol. Evol. **16**, 1791–1798 (1999)
20. Robinson, P.M.: Gaussian semiparametric estimation of long range dependence. Ann. Stat. **23**(5), 1630–1661 (1995)
21. Samorodnitsky, G.: Long range dependence. Found. Trends Stoch. Syst. **1**(3), 163–257 (2006)
22. Singh, H.V., Misra, N., Hnizdo, V., Fedorowicz, A., Demchuk, E.: Nearest neighbor estimates of entropy. Am. J. Math. Man. Sci. **23**, 301–321 (2003)
23. Taqqu, M.S., Teverovsky, V., Willinger, W.: Estimators for long-range dependence: an empirical study. Fractals **3**, 785–798 (1995)

Chapter 26
Bayesian Partition for Variable Selection in the Power Series Cure Rate Model

Jhon F. B. Gonzales, Vera. L. D. Tomazella and Mário de Castro

Abstract In this chapter we present a model of survival with a cure fraction where a feature of the model is that variable selection is performed by Bayesian partition model. To this end we consider a orthogonal hyperplane tessellation to obtain a local structure on space covariates. The proposed model is based on the promotion time where the number of competitive causes follows a power series distribution.

26.1 Introduction

With a rapid progress in the medical and health sciences, many datasets dealing with time to relapse now reveal a substantial proportion of patients who are expected to be non-susceptible to the occurrence of the event of interest (i.e. who are cured). Lifetime data in which there are sampling units non-susceptible to the occurrence of the event of interest, which can usually be caused by different latent competing causes, are common in applications from various areas, such as medical, financial, and industrial ones. The competing causes are latent in the sense that there is no information about which factor was responsible for the component failure (or individual death).

The statistical literature for modeling lifetime data in presence of a cure fraction and latent competing causes is by now vast and growing rapidly. Interested readers can refer to [3, 4, 12, 14] and [1] among others.

Typically, in medical and epidemiological studies, often interest focuses on studying nominal qualitative variables with more than two categories or ordinal qualitative variables. For example, researchers may be interested on study of cancer of melanoma, where we have important factors as disease stage, tumor size, category node, among other. In this sense, cure rate models relate to the cure fraction with the covariates through a (generalized) linear model. However, instead of considering

J. F. B. Gonzales (✉) · V. L. D. Tomazella
Departamento de Estatística, Universidade Federal de São Carlos,
Rod. Washington Luíz, km 235, 13565-905 São Carlos, SP, Brazil,
e-mail: jhonbg@gmail.com

M. de Castro
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo,
Av. Trabalhador São-carlense, 400 - Centro, São Carlos 13566-590, SP, Brazil

A. Polpo et al. (eds.), *Interdisciplinary Bayesian Statistics,*
Springer Proceedings in Mathematics & Statistics 118, DOI 10.1007/978-3-319-12454-4_26

a link function to connect the cure fraction to covariates, in this work uses a local structure generated by a tessellation on covariate space. So, we model the covariate effect locally in the cure fraction based on methodology of Bayesian partition model (BPM) proposed by [8].

The aim of this chapter is to present flexible methodologies suited to incorporate information on nominal qualitative variables with more than two categories or ordinal qualitative variables and perform variable selection. The cure rate model proposed assumes that the number of competing causes follow a power series distribution.

The paper is organized as follows. In Sect. 26.2 we present the power series cure rate model (PSCRM). In Sect. 26.3 we formulate the PSCRM model with BPM model. *In Sect. 26.4 we will apply the proposed model to a real dataset on melanoma data. This chapter concludes with Sect. 26.5 where we present some final comments.*

26.2 The Power Series Cure Rate Model

Let N be a discrete random variable representing the latent number of competing causes needed to the occurrence of a particular event of interest. We assume that N has the power series distribution [10] with probability mass function

$$P[N = k; \theta] = \frac{a_k \theta^k}{\eta(\theta)}, \quad k = 0, 1, 2, \ldots, \quad \theta > 0, \tag{26.1}$$

where $a_k > 0$ and $\eta(\theta) = \sum_{k=0}^{\infty} a_k \theta^k < +\infty$. In (26.1), θ and $\eta(.)$ are called the power parameter and the series function, respectively. The probability generating function of N is given by $G(s) = \eta(\theta s)/\eta(\theta)$, where $s \in (0, 1)$.

For different series functions $\eta(\cdot)$, we obtain different results, for example, if $\eta(\theta) = (1 + \theta)^K$ where K is positive integer and $a_k = \binom{K}{k}$, then (26.1) defines the binomial distribution, $N \sim \text{Bi}(K, \theta^*)$ where $\theta^* = \theta/(1 + \theta)$. If $\eta(\theta) = e^\theta$ and $a_k = 1/k!$, then (26.1) defines the Poisson distribution, $N \sim \text{Poi}(\theta)$. If $\eta(\theta) = (1 - \theta)^{-\tau}$, $\theta \in (0, 1)$ where τ is positive integer and $\binom{\tau+k-1}{\tau-1}$, then (26.1) defines the negative binomial distribution, $N \sim \text{Nb}(\tau, \theta)$. If $\eta(\theta) = -\log(1 - \theta)/\theta$, $\theta \in (0, 1)$ and $a_k = 1/(k + 1)$ then (26.1) defines the logarithmic distribution, $N \sim \text{Lg}(\theta)$.

Conditioned on N, let Z_v, $v = 1, \ldots, N$ be i. i. d. random variables with cumulative distribution function $F(t)$ and survival function $S(t) = 1 - F(t)$, where Z_v is the time of occurrence of a particular event of interest due to the v-th cause. For instance, in a biological scenario N may denote the number of carcinogenic cells which can produce a detectable tumor [16]. The observable time of occurrence of event of interest is $T = \min\{Z_1 \ldots, Z_N\}$. Under this setup, according to [15] and [13], the survival function for the population is given by

$$S_{pop}(t) = G(S(t)) = \frac{\eta(\theta S(t))}{\eta(\theta)}, \tag{26.2}$$

where $t \geq 0$.

The survival function $S_{pop}(t)$ given in (26.2) is not a proper survival function by the fact that $p_0 = \lim_{t\to\infty} S_{pop}(t) = a_0/\eta(\theta) < 1$, where p_0 denotes the cure fraction. So, the improper density and risk functions associated with long-term survival function in (26.2) are given respectively by

$$f_{pop}(t) = \frac{\eta'(\theta S(t))}{\eta(\theta)}\theta f(t) \text{ and } h_{pop}(t) = \frac{\eta'(\theta S(t))}{\eta(\theta S(t))}\theta f(t), \tag{26.3}$$

where $f(t)$ denotes the (proper) density function of the lifetime Z.

The PSCRM model considers some particular cases, for instance, if $\eta(\theta) = (1+\theta)^K$ we obtain the binomial cure rate model $S_{pop}(t) = (1-\theta^* + \theta^* S(t))^K$. If $\eta(\theta) = e^\theta$ we obtain the Poisson cure rate model, $S_{pop}(t) = \exp(-\theta F(t))$. If $N \sim \text{Nb}(\tau, \theta)$ we obtain the negative binomial cure rate model $S_{pop}(t) = \left((1-\theta)(1-\theta S(t))^{-1}\right)^\tau$. If $N \sim \text{Lg}(\theta)$ we obtain the logarithmic cure rate model,

$$S_{pop}(t) = \log(1-\theta S(t))/S(t)\log(1-\theta).$$

26.3 Bayesian Partition Modeling for Power Series Cure Rate Model

The partition models are methods that split some domain of interest $\boldsymbol{\mathcal{X}} \subset \mathbb{R}^p$ $(p \geq 1)$ in disjoint regions, and assign the same probability distribution for the response variable Y in each region of $\boldsymbol{\mathcal{X}}$. So, the BPM model partitioning $\boldsymbol{\mathcal{X}}$ by a tessellation of a structure $\mathbf{T}$ defining regions $R_m \subseteq \boldsymbol{\mathcal{X}}$, $m = 1, \ldots, M$.

One characteristic of the BPM is that assigning conjugate priors within the disjoint regions, the marginal likelihood is available for any tessellation structure. The availability of the marginal likelihood function for the tessellation structure greatly reduces the space of the models as well as the dimension of the parameter space.

In this paper we consider orthogonal hyperplanes tessellation, the hyperplanes are defined by split points $\boldsymbol{h}_{j^*}$, $j^* = 1, \ldots, p$ in each covariate. So, the tessellation structure is given by hyperplanes $\boldsymbol{h} = (\boldsymbol{h}_1, \ldots, \boldsymbol{h}_p)$ and the number of regions M, $\mathbf{T} = \{\boldsymbol{h}, M\}$.

Considering that we have M regions in $\mathcal{X}$, let N_{mj} be the number of latent causes of the event of interest for the j-observation with power series distribution with parameter θ_m, $j = 1, \ldots, n_m$ in the region R_m. So, given N_{mj}, let $Z_{mj}^1, \ldots, Z_{mj}^{N_{mj}}$ be times of occurrence of the event of interest with cumulative distribution function $F(\cdot|\boldsymbol{\gamma}) = 1 - S(\cdot|\boldsymbol{\gamma})$, where $\boldsymbol{\gamma}$ is the vector of parameters.

Let $T_{mj} = \min\left\{Z_{mj}^1, \ldots, Z_{mj}^{N_{mj}}\right\}$ and C_{mj} the censoring time. We observe $Y_{mj} = \min\{T_{mj}, C_{mj}\}$ and δ_{mj} be the censoring indicator with $\delta_{mj} = 1$ if $Y_{mj} = T_{mj}$ and $\delta_{mj} = 0$ otherwise. We assume a Weibull distribution with vector of parameters $\boldsymbol{\gamma} = (\alpha, \lambda)^\top$ for the event time Z_{mj}. The cumulative distribution is given by

$$F(y|\boldsymbol{\gamma}) = 1 - \exp(-y^\alpha e^\lambda),$$

where $\alpha > 0$ and $\lambda \in \mathbb{R}$. Then, the likelihood function for the complete data under uninformative censoring is given by

$$L(\boldsymbol{\gamma}, \boldsymbol{\theta}, \mathbf{T}|\boldsymbol{N}, \boldsymbol{y}, \boldsymbol{\delta}) = \prod_{m=1}^{M} \exp\left(-e^{\lambda} \sum_{j=1}^{n_m} y_{mj}^{\alpha} N_{mj}\right) \prod_{j=1}^{n_m} \left(N_{mj} \alpha e^{\lambda} y_{mj}^{\alpha-1}\right)^{\delta_{mj}} p(N_{mj}|\theta_m).$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_M)^{\top}$ and $\boldsymbol{N} = (N_1, \ldots, N_n)^{\top}$ is a vector of latent variables. Note that in each region R_m the number of causes for the event of interest N_{mj} has the same probability distribution (e.g. Poisson).

26.3.1 Prior and Posterior Distribution

The joint prior distribution for $(\boldsymbol{\gamma}, \boldsymbol{\theta}, \mathbf{T})$ is given by

$$p(\boldsymbol{\gamma}, \boldsymbol{\theta}, \mathbf{T}) = p(\boldsymbol{\gamma}) p(\boldsymbol{\theta}, \mathbf{T}) = p(\boldsymbol{\gamma}) p(\boldsymbol{\theta}|\mathbf{T}) p(\mathbf{T}),$$

and we assume that the parameters of the Weibull distribution being independent, so $p(\boldsymbol{\gamma}) = p(\alpha)p(\lambda)$ where $\alpha \sim \text{Gamma}(\mu_\alpha, \sigma_\alpha)$ and $\lambda \sim \text{N}(\mu_\lambda, \sigma_\lambda^2)$, where μ_α, σ_α, μ_λ, and σ_λ are hyperparameters. Considering that the parameters between regions of $\boldsymbol{\mathcal{X}}$ are independent, we have $p(\boldsymbol{\theta}|\mathbf{T}) = \prod_{m=1}^{M} p(\theta_m|\mathbf{T})$. Moreover, we assume a geometric distribution for M, $M \sim \text{Geo}(\psi)$.

To sample from posterior distribution $p(\boldsymbol{\gamma}, \boldsymbol{\theta}, \mathbf{T}|\boldsymbol{y}, \boldsymbol{\delta})$, we introduce the latent variables N_{mj}

$$p(\boldsymbol{\gamma}, \boldsymbol{\theta}, \mathbf{T}, \boldsymbol{N}|\boldsymbol{y}, \boldsymbol{\delta}) \propto \prod_{m=1}^{M} \exp\left\{e^{\lambda} \sum_{j=1}^{n_m} y_{mj}^{\alpha} N_{mj}\right\} \prod_{j=1}^{n_m} \left(N_{mj} \alpha e^{\lambda} y_{mj}^{\alpha-1}\right)^{\delta_{mj}} p(N_{mj}|\theta_m) \times p(\boldsymbol{\gamma}) p(\boldsymbol{\theta}, \mathbf{T}).$$

Therefore, we need to sample from the full conditional distributions $(\boldsymbol{\theta}, \mathbf{T}|\boldsymbol{N}, \boldsymbol{\gamma}, \boldsymbol{y}, \boldsymbol{\delta})$, $(\boldsymbol{N}|\boldsymbol{\theta}, \mathbf{T}, \boldsymbol{\gamma}, \boldsymbol{y}, \boldsymbol{\delta})$, and $(\boldsymbol{\gamma}|\boldsymbol{\theta}, \mathbf{T}, \boldsymbol{N}, \boldsymbol{y}, \boldsymbol{\delta})$. So, note that to sample from $(\boldsymbol{\theta}, \mathbf{T}|\boldsymbol{\gamma}, \boldsymbol{y}, \boldsymbol{\delta})$ we consider the full conditional distributions given by

$$p(\boldsymbol{\theta}, \mathbf{T}|\boldsymbol{N}, \boldsymbol{\gamma}, \boldsymbol{y}, \boldsymbol{\delta}) = p(\mathbf{T}|\boldsymbol{N}, \boldsymbol{\gamma}, \boldsymbol{y}, \boldsymbol{\delta}) p(\boldsymbol{\theta}|\mathbf{T}, \boldsymbol{N}, \boldsymbol{\gamma}, \boldsymbol{y}, \boldsymbol{\delta}).$$

In the BPM model supposes the parameters between each region of $\boldsymbol{\mathcal{X}}$ are independent and so the full conditional distribution for $(\mathbf{T}|\boldsymbol{N}, \boldsymbol{\gamma}, \boldsymbol{y}, \boldsymbol{\delta})$ is given by

$$p(\mathbf{T}|\boldsymbol{N}, \boldsymbol{\gamma}, \boldsymbol{y}, \boldsymbol{\delta}) \propto \int p(\boldsymbol{N}|\boldsymbol{\theta}, \mathbf{T}) p(\boldsymbol{\theta}|\mathbf{T}) p(\mathbf{T}) \mathrm{d}\boldsymbol{\theta} = p(\boldsymbol{N}|\mathbf{T}) p(\mathbf{T}).$$

In order to obtain a closed form for $p(\boldsymbol{N}|\mathbf{T})$ it is important to assign a conjugate prior for $\boldsymbol{\theta}$. However depending on the series function, $\eta(\theta)$, we have different distributions for N and so the prior distributions for θ are also different.

Thus, if $N \sim \text{Bi}(K, \theta^*)$ we choose a beta prior, $\theta_m^* \sim \text{Beta}(a_0, a_1)$. Therefore

$$p(\boldsymbol{N}|\mathbf{T}) = \prod_{i=1}^{n} \binom{K}{N_i} \prod_{m=1}^{M} \frac{\mathcal{B}(\sum_{j=1}^{n_m} N_{mj} + a_0, Kn_m - \sum_{j=1}^{n_m} N_{mj} + a_1)}{\mathcal{B}(a_0, a_1)}, \quad (26.4)$$

the full conditional distribution for θ_m^* and N_{mj} are given respectively by

$$\theta_m^*|\boldsymbol{N}, \mathbf{T} \sim \text{Beta}\left(\sum_{j=1}^{n_m} N_{mj} + a_0, n_m - \sum_{j=1}^{n_m} N_{mj} + a_1\right),$$

and

$$N_{mj}|\mathbf{T}, \boldsymbol{y}, \boldsymbol{\delta} \sim \text{Bi}\left\{K - \delta_{mj}, \frac{\theta_m^* S(y_{mj}|\gamma)}{\theta_m^* S(y_{mj}|\gamma) + 1 - \theta_m^*}\right\} + \delta_{mj}.$$

In case that $N \sim \text{Poi}(\theta)$, we assign a gamma distribution as prior to θ_m, $\theta_m \sim \text{Gamma}(b_0, b_1)$. So

$$p(\boldsymbol{N}|\mathbf{T}) = \prod_{m=1}^{M} \frac{1}{\prod_{j=1}^{n_m} N_{mj}!} \frac{b_1^{b_0}}{\Gamma(b_1)} \frac{\Gamma(\sum_{j=1}^{n_m} N_{mj} + b_0)}{(n_m + b_1)^{\sum_{j=1}^{n_m} N_{mj} + b_0}},$$

and the full conditional distribution for θ_m and N_{mj} are given respectively by

$$\theta_m|\boldsymbol{N}, \mathbf{T} \sim \text{Gamma}\left(\sum_{j=1}^{n_m} N_{mj} + b_0, n_m + b_1\right),$$

and

$$N_{mj}|\boldsymbol{\gamma}, \boldsymbol{\theta}, \mathbf{T}, \boldsymbol{y}, \boldsymbol{\delta} \sim \text{Poi}\left\{\theta_m S(y_{mj}|\gamma)\right\} + \delta_{mj}.$$

If $N \sim \text{Bn}(\tau, \theta)$, we choose a beta prior, $\theta_m \sim \text{Beta}(c_0, c_1)$, so

$$p(\boldsymbol{N}|\mathbf{T}) = \prod_{i=1}^{n} \binom{\tau + N_i - 1}{\tau - 1} \prod_{m=1}^{M} \frac{\mathcal{B}(\tau n_m + c_0, \sum_{j=1}^{n_m} N_{mj} + c_1)}{\mathcal{B}(c_0, c_1)},$$

the full conditional distribution for θ_m and N_{mj} are given respectively by

$$\theta_m|\mathbf{T}, \boldsymbol{N} \sim \text{Beta}\left(\sum_{j=1}^{n_m} N_{mj} + c_0, \tau n_m + c_1\right),$$

and

$$N_{mj}|\boldsymbol{\gamma}, \boldsymbol{\theta}, \mathbf{T}, \boldsymbol{y}, \boldsymbol{\delta} \sim \text{Nb}\left\{\tau + \delta_{mj}, \theta_m \exp\left(-e^{\lambda} y_{mj}^{\alpha}\right)\right\} + \delta_{mj}.$$

If $N \sim \text{Lg}(\theta)$, we assign the beta distribution as prior to θ_m, $\theta_m \sim \text{Beta}(d_0, d_1)$. Next, we do not have a closed form for $p(\boldsymbol{N}|\mathbf{T})$ and so we use numerical integration.

The full conditional for θ_m is given by

$$p(\theta_m|\boldsymbol{N}, \mathbf{T}) \propto \frac{\theta_m^{n_m+\sum_{j=1}^{n_m} N_{mj}+d_0-1}(1-\theta_m)^{d_1-1}}{\{-\log(1-\theta_m)\}^{n_m}}.$$

If $\delta_{mj} = 0$ the conditional distribution for N_{mj} is

$$N_{mj}|\boldsymbol{\gamma}, \boldsymbol{\theta}, \mathbf{T}, \boldsymbol{y}, \boldsymbol{\delta} \sim \text{Lg}(\theta_m S(y_{mj}|\boldsymbol{\gamma})).$$

If $\delta_{mj} = 1$ then

$$p(N_{mj}|\boldsymbol{\gamma}, \boldsymbol{\theta}, \mathbf{T}, \boldsymbol{y}, \boldsymbol{\delta}) \propto \frac{N_{mj}}{N_{mj}+1}\{\theta_m S(y_{mj}|\boldsymbol{\gamma})\}^{N_{mj}}.$$

Summarizing the hyperpameters for different priors on θ_m are $a_0, a_1, b_0, b_1, c_0, c_1, d_0$ and d_1.

The full conditional distributions for the parameters of the Weibull distribution are

$$p(\lambda|\alpha, \boldsymbol{N}, \mathbf{T}, \boldsymbol{y}, \boldsymbol{\delta}) \propto e^{d\lambda} \exp\left(-e^{\lambda}\sum_{i=1}^{n} N_i y_i^{\alpha}\right) \exp\left(-\frac{(\lambda-\mu_\lambda)^2}{2\sigma_\lambda^2}\right),$$

$$p(\alpha|\lambda, \boldsymbol{N}, \mathbf{T}, \boldsymbol{y}, \boldsymbol{\delta}) \propto \alpha^d \left(\prod_{i=1}^{n} y_i^{\delta_i}\right)^{\alpha} \exp\left(-e^{\lambda}\sum_{i=1}^{n} N_i y_i^{\alpha}\right) p(\alpha|\mu_\alpha, \sigma_\alpha),$$

where $d = \sum_{i=1}^{n} \delta_i$.

26.3.2 Computational Strategy

The strategy computational for numerical and dichotomous predictors is given in [7]. Generally, categorical predictors are not necessarily dichotomous therefore the algorithm has to be modified to handle with qualitative predictors but in general form. In this context, let $X_{\mathbf{T}}$ a categorical covariate with $C_{\mathbf{T}}$ categories, $X_{\mathbf{T}} \in \{1, 2, \ldots, C_{\mathbf{T}}\}$, and suppose that ρ is a partition of $X_{\mathbf{T}}$ and M_ρ the number of cluster(subsets) for partition ρ of $X_{\mathbf{T}}$. The partition ρ is unknown, we need to assign it as a prior probability $p(\rho)$. We assume that $p(\rho)$ is a discrete uniform distribution on $\{1, \ldots, n_\rho\}$ where n_ρ is the number different partitions of $X_{\mathbf{T}}$. So, a natural relationship among the number, the total numbers n_ρ of partitions, and scale of $X_{\mathbf{T}}$. In the case that $X_{\mathbf{T}}$ has a nominal scale the number n_ρ is much greater than when $X_{\mathbf{T}}$ is ordinal scale.

So, we denote by $\mathcal{I}$ the index set of predictor variables $\mathcal{I} = \{1, \ldots, p\}$, and $\mathcal{I}_{\mathbf{T}}$ is the index set of predictor variables present in the tessellation $\mathbf{T}$. Considering that $M = 1$ (i.e., $\mathcal{I}_{\mathbf{T}} = \emptyset$) then starting the algorithm with initializing the tessellation

structure, $\mathbf{T}$, with just one randomly drawn predictor variable and choose a split point. In each iteration of the algorithm and when $1 < M < n$, we try with probability $1/3$, the first three moves. The first two moves concern the selection of covariate. The last three moves involve the categorical predictor with more than two categories.

- **Add.** A new partition can be added to the tessellation structure $\mathbf{T}$ by choosing a new splitting point of a predictor variable in $\mathcal{I}$. The splitting point is selected from the empirical distribution of the predictor variable chosen.
- **Delete.** A hyperplane can be eliminated by choosing a random predictor variable r^* present in the tessellation, $r^* \in \mathcal{I}_{\mathbf{T}}$.
- **Move.** A hyperplane can be changed by selecting a new splitting point of the empirical distribution of the selected variable in $\mathcal{I}_{\mathbf{T}}$.
- **Merge.** The number of clusters in the covariate $X_{\mathbf{T}}$ is decreased, by merging two subsets.
- **Split.** The number of clusters in the covariate $X_{\mathbf{T}}$ is increased, by splitting up one subset into two new subsets.
- **Alter.** The partition ρ for $X_{\mathbf{T}}$ is altered, but the number of subsets being equal.

The new tessellation $\mathbf{T}'$ proposal is accepted with probability:

$$\alpha(\mathbf{T}', \mathbf{T}) = \min\left\{1, \frac{p(\boldsymbol{N}|\mathbf{T}')p(\mathbf{T}')}{p(\boldsymbol{N}|\mathbf{T})p(\mathbf{T})}\right\}. \tag{26.5}$$

[7] proposed the first three moves. In this chapter, added the last 3 moves, which only concern qualitative covariate and are the main novelty of our approach. Details of the movements of Markov chain Monte Carlo (MCMC) sampler can be found in [11].

26.4 Application

The data set for illustrating our methodology was extracted from a melanoma study (the melanoma is a type of malignant cancer). The objective of this study is to evaluate the effectiveness of applying a high dosage of interferon alfa-2b as a way to prevent the recurrence of cancer. Patients were included in the study between 1991 to 1995, and follow-up was conducted until 1998. The response variable Y represents the time from patient to death or time of censoring. The original sample comprises 427 patients, 10 of whom were removed from analysis, since their tumor thickness data are missing. Therefore we have $n = 417$ patients, with 56% of censored observations. The variables include y: time (in years); x_1: treatment (0: observation, $n = 204$; 1: interferon, $n = 213$); x_2: age(in years); x_3: nodule category(1,$n = 82$; 2,$n = 87$; 3,$n = 137$; 4,$n = 111$); x_4: sex (0: male, $n = 263$; 1: female, $n = 154$); x_5: performance status(PS) means patient's functional capacity as regards his/her daily activities (0: fully active, $n = 363$ 1: other, $n = 54$) and x_6: tumor thickness (in mm). For more details related to the melanoma data [9] may be consulted.

Table 26.1 Probability of splitting for covariates for different models

Model	x_1	x_2	x_3	x_4	x_5	x_6
Binomial ($K = 10$)	0.001	0.123	0.999	0.001	0.002	0.020
Poisson	0.018	0.307	1.000	0.025	0.024	0.102
Negative binomial ($\tau = 1$)	0.020	0.256	1.000	0.019	0.017	0.092
Logarithmic	0.010	0.133	1.000	0.012	0.009	0.131

We consider hyperparameters $\mu_\alpha = \sigma_\alpha = 0.1$ for the gamma distribution of the parameter α and the normal distribution with mean $\mu_\lambda = 0$ and variance $\sigma^2_\lambda = 100$ for the parameter λ. The hyperparameters for beta distribution are $a_0 = a_1 = c_0 = c_1 = d_0 = d_1 = 1$ and for gamma distribution we assume $b_0 = b_1 = 0.1$. For number of regions M we assume that, $M \sim \text{Geo}(0.1)$.

Considering these prior densities we generated two parallel independent runs of the MCMC sampler with size 700, 000 for each parameter, disregarding the first 300, 000 iterations to eliminate the effect of the initial values and, to avoid correlation problems, we consider a spacing of length 100, obtaining a sample of size 4000 in each chain. To monitor the convergence of the MCMC sampler we resorted to the methods recommended by [5]. Consider in the first chain initial values for λ and α equal to -5 and 5 respectively in the second chain were 5 and 9, in both chains initiate the algorithm with $\boldsymbol{N} = (1, \ldots, 1)$.

In the data set x_1, x_4, and x_5 are binary variables then the division of any of these variables follows as, if it occurs will result in two groups, for example, the variable x_4 which represents sex will be divided in male and female. In the case of the covariate x_3 with four categories $x_3 \in \{1, 2, 3, 4\}$, the idea of partition of x_3 was made considering the Sect. (26.3.2) except that the choice of partitions has an order.

We tried different binomial and negative binomial models by taking $K = 1, 2, 7, 10$ and $\tau = 1, 3, 7, 13$ respectively. For the binomial model, the best fit is when $K = 10$. In the negative binomial model the best fit was when $\tau = 1$ i.e. the geometric model.

Table 26.1 presents the probability of splitting for each of the covariates for each model. We note that the ordinal covariate x_3 is almost always in the tessellation, so the tessellation by orthogonal hyperplanes identifies that x_3 has a significant effect in the models. One consequence is that the splitting probability of x_3 is very close to 1. Also, the variables x_2 and x_6 have a minor effect in the models. For the other covariates, the probability of splitting is close to zero which means that they are noninformative.

Moreover, the partition with largest posterior probability for x_3 is the partition formed by $\{\{1, 2, 3\}, \{4\}\}$ for binomial (0.639), Poison (0.766), and negative binomial (0.340) models. Nevertheless, in the logarithmic model the partition $\{\{1, 2\}\{3, 4\}\}$ have larger posterior probability.

To asses the goodness of fit of the models, we use the logarithm of pseudomarginal likelihood (LPML) given in [9, Chap. 6]. LPML is a well-established Bayesian model comparison criterion based on the conditional predictive ordinate (CPO) statistics, which is particularly suitable for the cure rate models.

Table 26.2 Posterior summaries for the parameters of the Weibull distribution

Model	LPML	Parameter	Mean	SD	95% HPD
Binomial†	−521.775	α	1.599	0.109	(1.394; 1.820)
		λ	−1.295	0.125	(−1.532; −1.050)
Poisson	−521.482	α	1.721	0.116	(1.495; 1.947)
		λ	−1.645	0.135	(−1.920; −1.388)
Geometric	−519.892	α	1.869	0.125	(1.624; 2.105)
		λ	−2.069	0.125	(−2.390; −1.757)
Logarithmic	−519.004	α	2.044	0.136	(1.766; 2.293)
		λ	−2.454	0.213	(−2.890; −2.071)

† $K = 10$

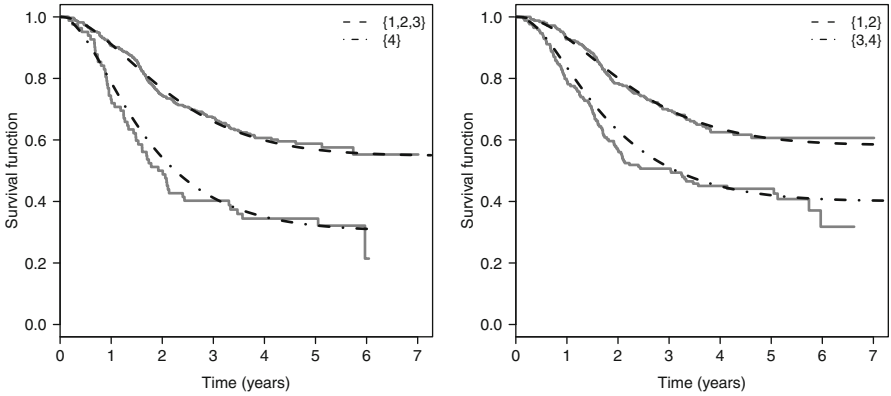

Fig. 26.1 **a** K-M curves stratified by nodule category for the clusters $\{\{1, 2, 3\}, \{4\}\}$ (*lower curve*) for geometric model. **b** Estimates of the survival function according with cluster $\{\{1, 2\}, \{3, 4\}\}$ for logarithmic model

Table 26.2 gives LPML, posterior means, standard deviations (SD), and 95 % highest posterior density (HPD) interval for the parameters of Weibull for all models. Also, we calculated the estimated potential scale reduction $\widehat{R}$ [6] for the parameters of Weibull distribution, which for all parameter is close to 1, indicating good convergence. We also note from Table 26.2 that, based on the LPML statistics the logarithmic model is deemed as the best fitting model. Note that, the SD of posterior estimates of parameter λ in the binomial, Poisson and negative binomial are close, but in the logarithmic model is larger than the others models.

Figure 26.1a shows the Kaplan-Meier(K-M) estimates of survival function and estimates obtained from negative binomial model considering the partition $\{\{1, 2, 3\}, \{4\}\}$ and in the Fig. 26.1b display the K-M estimates of survival function and estimates of logarithmic model considering the partition $\{\{1, 2\}, \{3, 4\}\}$ for covariate x_3.

26.5 Discussion

In this paper, we proposed the power series cure rate based in the Bayesian partition modelling. The model proposed is a extension nonparametric for the power series cure rate [2].

We propose a strategy computational that considers quantitative, dichotomous, and also qualitative covariates with two more categories and order. Moreover, the partition of ordinal covariates is performed respecting an order and this is novelty.

Thus, the methodology proposed extend the model proposed by [7].

Considering the hyperplane orthogonal tessellation in this modeling the variable selection is performed. In this sense, each hyperplane divides the data set in only one covariate and thus the hyperplanes are included when the covariate affects the fit of the model.

Acknowledgement The research was partially supported by the Brazilian Organizations FAPESP, CNPq and CAPES.

References

1. Cancho, V.G., Rodrigues, J., de Castro, M.: A flexible model for survival data with a cure rate: a Bayesian approach . J. Appl. Stat. **38**, 57 – 70 (2011)
2. Cancho, V.G., Louzada, F., Ortega, E.M.: The power series cure rate model: an application to a cutaneous melanoma data. Commun. Stat. Simul. Comput. **42**, 586–602 (2013)
3. Chen, M.H., Ibrahim, J.G., Sinha, D.: A new Bayesian model for survival data with a surviving fraction. J. Am. Stat. Assoc. **94**, 909–919 (1999)
4. Cooner, F., Banerjee, S., Carlin, B.P., Sinha, D.: Flexible cure rate modeling under latent activation schemes. J. Am. Stat. Assoc. **102**, 560–572 (2007)
5. Cowles, M.K., Carlin, B.P.: Markov chain Monte Carlo convergence diagnostics: a comparative review. J. Am. Stat. Assoc. **91**, 883–904 (1996)
6. Gelman, A., Rubin, D.B.: Inference from iterative simulation using multiple sequences. Stat. Sci. **7**, 457–472 (1992)
7. Hoggart, C., Griffin, J.E.: A Bayesian partition model for customer attrition. In: E.I. George (ed.) Bayesian Methods with Applications to Science, Policy, and Official Statistics(Selected Papers from ISBA 2000), pp. 61–70. International Society for Bayesian Analysis, Proceedings of the the Sixth World Meeting of the International Society for Bayesian Analysis, Creta, Greece (2001)
8. Holmes, C.C., Denison, D.G.T., Ray, S., Mallick, B.K.: Bayesian prediction via partitioning. J. Comput. Gr. Stat. **14**, 811–830 (2005)
9. Ibrahim, J.G., Chen, M.H., Sinha, D.: Bayesian Survival Analysis. Springer, New York (2001)
10. Johnson, N.L., Kemp, A.W., Kotz, S.: Univariate Discrete Distributions, 3rd edn. Wiley, Hoboken (2005)
11. Louzada, F., de Castro, M., Tomazella, V., Gonzales, J.F.B.: Modeling categorical covariates for lifetime data in the presence of cure fraction by Bayesian partition structures. J. Appl. Stat. **41**, 622–634 (2014)
12. Maller, R.A., Zhou, S.: Survival Analysis with Long-Term Survivors. Wiley, New York (1996)
13. Rodrigues, J., Cancho, V.G., de Castro, M., Louzada-Neto, F.: On the unification of long-term survival models.Stat Probab. Lett. **79**, 753–759 (2009)

14. Rodrigues, J., de Castro, M., Cancho, V.G., Balakrishnan, N.: COM-Poisson cure rate survival models and an application to a cutaneous melanoma data. J. Stat. Plan. Inference **139**, 3605–3611 (2009)
15. Tsodikov, A.D., Ibrahim, J.G., Yakovlev, A.Y.: Estimating cure rates from survival data: an alternative to two-component mixture models. J. Am. Stat. Assoc. **98**, 1063–1078 (2003)
16. Yakovlev, A.Y., Tsodikov, A.D.: Stochastic Models of Tumor Latency and their Biostatistical Applications. World Scientific, Singapore (1996)

Chapter 27
Bayesian Semiparametric Symmetric Models for Binary Data

Marcio Augusto Diniz, Carlos Alberto de Bragança Pereira and Adriano Polpo

Abstract This work proposes a general Bayesian semiparametric model for binary data. Symmetric prior probability curves as an extension for discussed ideas from Basu and Mukhopadhyay (Generalized Linear Models: A Bayesian Perspective, pp. 231–241, 1998) are considered using the blocked Gibbs sampler, which is more general than the Polya urn Gibbs sampler. The Bayesian semiparametric approach allows us to incorporate uncertainty around the F distribution of the latent data and to model heavy-tailed or light-tailed distributions. In particular, the Bayesian semiparametric *logistic* model is introduced, which enables one to elicit prior distributions for regression coefficients from information about odds ratios; this is quite interesting in applied research. Then, this framework opens several possibilities to deal with binary data in the Bayesian perspective.

27.1 Introduction

The binary data modeling is a recurring challenge in several applied research areas considering that the *logistic* regression popularized by Cox [10] and the *Probit* model introduced by Bliss [6] are the strategies adopted often.

These models are obtained when the probability curve of success is defined as a distribution function F evaluated on some covariables in the case of *logistic* and *normal* distributions.

The distribution function F usually is symmetric with mean $\mu = 0$ and precision $\tau = 1$; thereby, the probability of success for a binary response approximates to the value zero with the same rate that it approximates to value one. Furthermore, the distribution F could be a scale mixture of a G distribution by a H distribution.

M. A. Diniz (✉) · C. A. de Bragança Pereira
Institute of Mathematics and Statistics, University of São Paulo, São Paulo, Brazil
e-mail: dnz.marcio@gmail.com

C. A. de B. Pereira
e-mail: cpereira@ime.usp.br

A. Polpo
Federal University of Sao Carlos, Rod. Washington Luiz,
km 235, Sao Carlos 13565-905, SP, Brazil
e-mail: polpo@ufscar.br

A. Polpo et al. (eds.), *Interdisciplinary Bayesian Statistics,*
Springer Proceedings in Mathematics & Statistics 118, DOI 10.1007/978-3-319-12454-4_27

In the Bayesian parametric approach, a binary random variable is treated by Albert and Chib [1] as a discretization of a latent variable with the F distribution. In this context, a scale mixture of distributions is considered such that G is defined as the *normal* distribution and H as the *gamma* distribution, resulting in F as the t distribution. Consequently, this structure makes available the t, *Cauchy*, and *normal* distributions for describing the probability curve of success.

The *Kolmogorov–Smirnov* test statistic distribution for H was suggested by Chen and Dey [8] in the Bayesian parametric approach, resulting in the *logistic* distribution for describing the probability curve, which is deeply desired since it allows one to elicit prior distributions for $\boldsymbol{\beta}$, the vector of regression coefficients, discussing about the odds ratios that imply a great impact on applied research.

Bayesian nonparametric approach through the *Dirichlet* process, proposed by Ferguson [14], allows as a more flexible alternative to previous modeling approaches because it excludes the necessity to define H mixture distribution, which means that it is possible to model heavy-tailed or light-tailed distributions than that prior proposed.

A semiparametric model based on the same structure created by Albert and Chib [1] was introduced by Basu and Mukhopadhyay [4]. Nonetheless, the H mixture distribution is not known and treated as a random quantity such that *Dirichlet* process is considered as the distribution for H.

The computational implementation is based on Polya urn Gibbs sampler algorithm developed by Escobar [12] that is constructed by *Dirichlet* process definition from [5]. Thus, the work presented in [4] resulted in the t, *Cauchy*, and normal distributions as options for the prior expected distribution of the probability curve of success as well as [1].

This work proposes a Bayesian semiparametric general model for binary data. Symmetric probability curves are considered as an extension to the ideas discussed in [4] for *logistic* prior expected distribution through the blocked Gibbs sampler. The blocked Gibbs sampler was introduced by Ishwaran and Zarepour [15] such that it is more general and easy to implement than the Polya urn Gibbs sampler.

Section 27.2 defines the modeling and presents some concepts about the Bayesian nonparametric approach; in Sect. 27.3, the Gibbs sampler is established. Finally, the beetle data from [6] are revisited in Sect. 27.4. Concluding remarks are given in Sect. 27.5.

27.2 The Model

The problem can be described when considering $Y_{il}|p_i \sim B(p_i)$ as independent binary random variables when $p_i = F(\mathbf{x}_i'\beta|\mu, \tau)$ for $i = 1, \ldots, n$ and $l = 1, \ldots, n_i$ such that F is a distribution function of a scale-location family $\mathbb{F} = \{F(.|\mu, \tau) : \mu \in \Re, \tau > 0\}$, $\mathbf{x}_i'$ is a $p \times 1$ covariables vector from i^{th} subject, and β is a $p \times 1$ parameter vector. In order to simplify the notation, a specific $F(.|\mu = a, \tau = b)$ distribution will be indicated by only $F(.|a, b)$.

Following [1],

$$Y_{il}|W_{il} = \mathbb{1}_{(W_{il}>0)}, \tag{27.1}$$

where $W_{il}|\beta,\tau \sim F(.|\mathbf{x}_i{}'\beta,\tau)$ for $i = 1,\ldots,L$, because

$$\begin{aligned} p_i &= E(Y_{il}) \\ &= P(Y_{il} = 1) = P(W_{ij} > 0) \\ &= P(W_{il} - \mathbf{x}_i'\beta > -\mathbf{x}_i'\beta) \\ &= 1 - F(-\mathbf{x}_i'\beta|0,\tau) \\ &= F(\mathbf{x}_i'\beta|0,\tau), \end{aligned} \tag{27.2}$$

if F is symmetric around zero.

Usually, it is assumed $\tau = 1$, although, it is possible to attribute more flexibility to the modeling of W when the F distribution is defined as a mixture of G symmetric distributions in a scale-location family $\mathbb{G}$, that is,

$$f(W|\beta,H) = \int g(W|\mathbf{x}'\beta,\boldsymbol{\tau})dH(\boldsymbol{\tau}). \tag{27.3}$$

In a hierarchical perspective,

$$W_{il}|\beta,\tau_i \overset{ind}{\sim} G(W_{il}|\mathbf{x}_i{}'\beta,\tau_i) \quad \text{for } i = 1,\ldots,L \text{ e } l = 1,\ldots,n_i, \tag{27.4}$$

$$\tau_1,\ldots\tau_L|H \overset{i.i.d.}{\sim} H. \tag{27.5}$$

On the Bayesian paradigm, prior distributions should be attributed to the unknown quantities,

$$\beta|\tau_\beta \sim N_p(0,\tau_\beta I_p), \tag{27.6}$$

$$H|\alpha \sim P(\alpha,H_0), \tag{27.7}$$

$$\alpha \sim \gamma(c_1,c_2).$$

$P(\alpha,H_0)$ being the *Dirichlet* process introduced in [14] such that H_0 is a distribution that indicates the expected distribution for H and α is the parameter that indicates the dispersion around H_0 in relation to the sample size n. Also, α can be seen as a function of the expected value of the number of clusters of τ,

$$C(\alpha,n) = \sum_{i=1}^{n} \frac{\alpha}{\alpha+i-1}. \tag{27.8}$$

Notice that the H_0 distribution should be chosen from the prior expected distribution for W denoted by F_0. In this way, it is important to point out the work of Andrews and Mallows [2], who discussed scale *normal* mixtures and presented the relations used in [1] and [8].

Another point to be highlighted is that the posterior distribution for H is not a simple *Dirichlet* process, but it is a mixture of *Dirichlet* processes defined in [3].

Computational implementation of a mixture of *Dirichlet* processes had been enormously difficult until the Polya urn algorithm was presented in [12], since the previous algorithms could not sample from H posterior distribution adequately.

The Polya urn algorithm is based on the representation of the *Dirichlet* process introduced by Blackwell and McQueen [5]. In spite of the algorithm being appropriate in several situations, it is limited for situations where there is conjugacy between H_0 and G distributions, and it suffers from slow mixing because of the one-at-a-time updates. Solutions for these limitations were presented by Escobar and West [13], MacEachern [16], and MacEachern and Müller [17], among others.

An alternative algorithm was developed by Ishwaran and Zarepour [15] based on the stick-breaking representation introduced by Sethuraman [19],

$$P(.) = \sum_{k=1}^{\infty} q_k \delta_{\tau_k}(.), \tag{27.9}$$

where δ_{τ_k} is a discrete probability measure concentrated on τ_k such that $\tau_k \sim H_0$ and q_k are independent random variables of τ_k which is given by,

$$\begin{aligned} q_1 &= V_1, \\ q_k &= (1 - V_1)(1 - V_2) \times \cdots \times (1 - V_{k-1})V_k \quad \text{for} \quad k \geq 2, \\ \sum_{k=1}^{\infty} q_k &= 1, \end{aligned} \tag{27.10}$$

when V_k for $k = 1, 2, \ldots$ are independent random variables and identically $Beta(1, \alpha)$-distributed.

The algorithm considers an approximation for the *Dirichlet* process by truncating the sum in Eq. (27.9) for a finite sum until N. The quality of this approximation also is established in [15]:

$$||f(\mathbf{W}) - f_N(\mathbf{W})||_1 \leq 4n \exp(-(N-1)/\alpha), \tag{27.11}$$

then, the number of components of *Dirichlet* process, N, should be chosen from the sample size n and the dispersion parameter α.

The implementation of the algorithm requires that the model presented in Eqs. (27.4)–(27.8) be rewritten by considering $\tau_i = Z_{K_i}$ such that K_i for $i = 1, \ldots, L$ are classification variables to identify the variable Z_k associated with each τ. Then, the model can be described as,

$$\begin{aligned} W_{il}|\beta, \mathbf{Z}, \mathbf{K} &\overset{ind}{\sim} \Phi(W_{il}|\mathbf{x}_i'\beta, Z_{K_i}) \quad \text{para } i = 1, \ldots, L \text{ and } l = 1, \ldots, n_i, \\ \beta|\tau_\beta &\sim N_p(0, \tau_\beta I_p), \\ Z_j|v &\overset{i.i.d.}{\sim} H_0(v) \quad \text{para } j = 1, \ldots, N, \end{aligned}$$

$$\begin{aligned}\mathbf{K}|\mathbf{q} &\sim Multinomial(L, q_1, \ldots, q_N),\\ \mathbf{q}|\alpha &\sim GD_N(\alpha, 1),\end{aligned} \tag{27.12}$$

where *GD* is the *generalized Dirichlet* distribution discussed in [9].

Finally, it is possible to present the algorithm from the model.

27.3 Blocked Gibbs Sampler

In this structure, the sampling of H posterior distribution is simplified and divided in the sampling of $\mathbf{Z}|\mathbf{K}, \mathbf{W}, \beta, \nu$, $\mathbf{K}|\mathbf{q}, \mathbf{Z}, \mathbf{W}, \beta$, $\mathbf{q}|\alpha, \mathbf{K}$ and $\alpha|\mathbf{q}, c_1, c_2$. Moreover, there is the sampling of $\mathbf{W}|\mathbf{Z}, \mathbf{K}, \beta, \mathbf{Y}, \mathbf{X}$ and $\beta|\mathbf{W}, \mathbf{Z}, \mathbf{K}$.

The sampling of the posterior complete conditional distribution $\mathbf{Z}|\mathbf{K}, \mathbf{W}, \beta, \nu$ is divided in two parts,

$$Z_k \overset{i.i.d.}{\sim} h_0(Z_k|\nu) \quad \text{for} \quad k \in \{1, \ldots, N\} - K^*, \tag{27.13}$$

$$Z_k \overset{ind}{\sim} h_0(Z_k|\nu) \prod_{\{i:K_i=k\}} \prod_{l=1}^{n_i} \phi(W_{il}|\mathbf{x}_i'\beta, Z_k) \quad \text{for} \quad k \in K^*, \tag{27.14}$$

where $K^* = \{K_1^*, \ldots, K_m^*\}$ consists in the set of unique values of $\mathbf{K}$ vector.

For the sampling of $\mathbf{K}|\mathbf{q}, \mathbf{Z}, \mathbf{W}, \beta$, each component K_i follows the *multinomial* distribution with probabilities given by,

$$q_{i,k} \propto q_k \tau(Z_k)^{n_i/2} \exp\left\{-\frac{\tau(Z_k)}{2} \sum_{l=1}^{n_i} (W_{il} - \mathbf{x}_i'\beta)^2\right\}, \tag{27.15}$$

for $i = 1, \ldots, L$ and $k = 1, \ldots, N$.

The *generalized Dirichlet* conjugates with *multinomial* distribution of $\mathbf{K}$ following [20]; thus, $\mathbf{q}|\alpha, \mathbf{K}$ also is a *generalized Dirichlet* distribution with parameters,

$$\begin{aligned}a_k^W &= 1 + m_k,\\ b_k^W &= \alpha + \sum_{j=k+1}^{N} m_j.\end{aligned} \tag{27.16}$$

m_k being the number of K_i equal to k for $k = 1, \ldots, N-1$.

The $\alpha|\mathbf{q}, c_1^W, c_1^W$ also takes advantage of the conjugacy such that the posterior distribution is still *gamma* with parameters,

$$\begin{aligned}c_1^W &= N + c_1 - 1,\\ c_2^W &= c_2 + ln(q_N).\end{aligned} \tag{27.17}$$

Similarly, in the Bayesian linear regression with *normal* errors, $\beta|\mathbf{W},\mathbf{Z},\mathbf{K}$ is given by the $normal_p\left(\boldsymbol{\mu}^W,\boldsymbol{\Sigma}^W\right)$ distribution such that,

$$\begin{aligned}\boldsymbol{\mu}^W &= \left(\tau_\beta I_p + \mathbf{X}'\boldsymbol{\Sigma}\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Sigma}\mathbf{W},\\ \boldsymbol{\Sigma}^W &= \left(\tau_\beta I_p + \mathbf{X}'\boldsymbol{\Sigma}\mathbf{X}\right).\end{aligned} \tag{27.18}$$

where Σ is a diagonal matrix such that the i-th element corresponds to $\tau(Z_{K_i})$. The sampling of $\mathbf{W}|\mathbf{Z},\mathbf{K},\beta,\mathbf{Y}$ corresponds to the sampling of the $\phi(\mathbf{W}|\mathbf{Z},\mathbf{K},\beta)$ distribution truncated by $\mathbf{Y}$ as defined in (27.1). Considering [18], it is easy to generate univariate truncated *normal*,

$$\pi(\mathbf{W}|\mathbf{Z},\mathbf{K},\beta,\mathbf{Y},\mathbf{X}) \propto \prod_{i=1}^{L}\prod_{l=1}^{n_i}\phi(W_{il}|\mathbf{x}_i'\beta, Z_{K_i})\mathbb{1}_{\left(Y_{il}=\mathbb{1}_{(W_{il}>0)}\right)}. \tag{27.19}$$

The algorithm that samples the posterior distribution $\mathbf{W}, H, \beta|\mathbf{Y},\mathbf{X}$ distribution is defined by,

1. Define initial values for $\mathbf{W}$, $\mathbf{Z}$, $\mathbf{K}$, $\mathbf{q}$, α, and β;
2. Generate H from $\pi(H|\beta,\mathbf{W})$ by parts,
   a) $\mathbf{Z} \sim \pi(\mathbf{Z}|\mathbf{K},\mathbf{W},\beta,v)$ defined in (27.13) and (27.14);
   b) $\mathbf{K} \sim \pi(\mathbf{K}|\mathbf{q},\alpha,\mathbf{Z},\mathbf{W},\beta)$ detailed in (27.15);
   c) $\mathbf{q} \sim DG_N(\mathbf{a}^{\mathbf{W}},\mathbf{b}^{\mathbf{W}})$ with parameters defined in (27.16);
   d) $\alpha \sim gamma(c_1^W, c_2^W)$ with parameters given in (27.17);
3. Generate $\beta \sim N_p\left(\boldsymbol{\mu}^W,\boldsymbol{\Sigma}^W\right)$ with parameters presented in (27.18);
4. Generate $\mathbf{W} \sim \pi(\mathbf{W}|\mathbf{Z},\mathbf{K},\beta,\mathbf{Y},\mathbf{X})$ expressed in (27.19).

The next two sections will present the scale *normal* mixture with *gamma* distribution for the scale parameter as an alternative version of the work present in [4] and the *Kolmogorov–Smirnov* distribution in order to provide the *logistic* distribution for the prior expected distribution of the *Dirichlet* process. Other distributions can be constructed since the algorithm is general enough.

27.3.1 Gamma Distribution

If $\tau(Z) = Z$ such that $Z \sim gamma(v/2, v/2)$ with density,

$$h_0(Z) = \frac{(v/2)^{\frac{v}{2}}}{\Gamma(v/2)}\tau^{\frac{v}{2}-1}\exp\left\{-\frac{v}{2}Z\right\}\mathbb{1}_{(Z>0)}, \tag{27.20}$$

it follows that F_0 is the $t_v(0,1)$ distribution. If $v = 1$, F_0 is a *Cauchy*$(0,1)$ distribution, while, if $v \to \infty$, F_0 is a *normal*$(0,1)$ distribution. Notice that the *t-student* distribution is already considered a reasonable approximation to *normal* distribution with $v > 30$.

Then, the posterior distribution calculated in (27.14) is the *gamma* distribution with parameters,

$$\begin{aligned} v_1^W &= v/2 + m_k/2, \\ v_2^W &= \frac{1}{2} \sum_{\{i:K_i=k\}} \sum_{\{l=1\}}^{n_i} (W_{il} - \mathbf{x}_i'\beta)^2 + \frac{v}{2}. \end{aligned} \tag{27.21}$$

27.3.2 Kolmogorov–Smirnov Distribution

If $\tau(Z)$ is defined as

$$\tau(Z) = \left(\frac{1}{2Z}\right)^2 \quad \text{such that} \quad Z \sim \textit{Kolmogorov–Smirnov}, \tag{27.22}$$

with *Kolmogorov–Smirnov* distribution following [11],

$$h_0(Z) = 8 \sum_{k=1}^{\infty} (-1)^{n+1} n^2 Z \exp\{-2n^2 Z^2\} \mathbb{1}_{(Z>0)}, \tag{27.23}$$

it follows that F_0 is a *logistic*$(0, 1)$ distribution.

It is not simple to evaluate the *Kolmogorov–Smirnov* distribution since the summing limit is not finite, although it is possible to generate these random variables following Devroye [11], who presented an algorithm based on the alternating series algorithm.

Unlike the last section, it is preferable to consider the random variable Z to calculate the posterior distribution which results in,

$$\begin{aligned} &\pi(Z|\mathbf{K}, \mathbf{W}, \beta) \propto \\ &\propto 8 \sum_{k=1}^{\infty} (-1)^{n+1} n^2 Z \exp\{-2n^2 Z^2\} \left(\frac{1}{4Z^2}\right)^{m_k/2} \times \\ &\times \exp\left\{-\frac{1}{8Z^2} \sum_{\{i:K_i=k\}} \sum_{\{l=1\}}^{n_i} (W_{il} - \mathbf{x}_i'\beta)^2\right\}. \end{aligned} \tag{27.24}$$

This posterior distribution is not known. Here, the Blocked Gibbs sampler introduced in [15] becomes extremely interesting because a Metropolis–Hastings step can be added without difficulties to sample from (27.24).

In the Bayesian parametric approach, a proposal distribution for the same distribution in (27.24) is suggested by Chen and Dey [8]. They considered the empirical relation between t_v and *logistic* distributions discussed in [1].

In this way, the proposal distribution is defined by $Z^2 \sim$ *inverse gamma* with the parameters,

$$\begin{aligned} v_1^W &= v/2 + m_k/2, \\ v_2^W &= \frac{1}{8}\left[\frac{v}{b^2} + \sum_{\{i:K_i=k\}}\sum_{\{l=1\}}^{n_i}(W_{il} - \mathbf{x}_i'\beta)^2\right], \end{aligned} \tag{27.25}$$

with acceptance probability of Z^* generated from proposal distribution,

$$\lambda = \frac{h_0(Z^*)/\, h_0^a(Z^*)}{h_0(Z^*)/\, h_0^a(Z^*)}, \tag{27.26}$$

where h_0^a denotes the proposal distribution.

It is necessary to evaluate the *Kolmogorov–Smirnov* distribution to calculate the acceptance probability; therefore, Chen and Dey [8] presented a limit to truncate the infinite sum.

The limit is based on the decomposition of the density in an alternating series,

$$h_0(Z) = cf_d(Z)\sum_{n=0}^{\infty}(-1)^n a_n(Z),$$

where c is a constant, f_d is an easily generating density, and a_n is a monotone decreasing series.

The first decomposition is as follows,

$$\begin{aligned} cf_d(Z) &= 8Z\exp\{-2Z^2\}, \\ a_n(Z) &= (n+1)^2\exp\{-2Z((n+1)^2+1)\}, \end{aligned} \tag{27.27}$$

such that $a_n \searrow 0$ for $Z > \sqrt{1/3}$, while, the second decomposition is given by,

$$\begin{aligned} cf_d(Z) &= \frac{\sqrt{2\pi}\pi^2}{4Z^4}\exp\left\{-\frac{\pi^2}{8Z^2}\right\}, \\ a_n(Z) &= \begin{cases} \frac{4Z^2}{\pi^2}\exp\left\{-\frac{(n-1)^2\pi^2}{8Z^2}\right\} & \text{if odd } n, \\ (n+1)^2\exp\left\{-\frac{((n+1)^2-1)\pi^2}{8Z^2}\right\} & \text{if even } n, \end{cases} \end{aligned} \tag{27.28}$$

with $a_n \searrow 0$ for $Z < \pi/2$. Observe that the convergence interval of both series intersects. Then, we choose $Z = 0.75$ can be chosen as a cutoff to evaluate each series as well as [11].

The limit to truncate the sum in (27.23) is as follows,

$$n^* = inf\{n : cf_d(Z)a_n(Z) < \delta\}, \tag{27.29}$$

where δ is the precision of approximation.

27.4 Predictive Distribution

A direct by-product of the blocked Gibbs sampler is the predictive distribution for a new observation $W_{(n+1)l}$ and, consequently, $Y_{(n+1)l}$ for $l = 1, \ldots, L$. See,

$$f(W_{(n+1)l}|\mathbf{Y}, \mathbf{X}) = \int \phi(W_{(n+1)l}|\mathbf{x}'_{(n+1)}\beta, \boldsymbol{\tau}) d\pi(\beta, \boldsymbol{\tau}|\mathbf{Y}, \mathbf{X})$$
$$= \int \int \phi(W_{(n+1)l}|\mathbf{x}'_{(n+1)}\beta, \boldsymbol{\tau}) d\pi(\boldsymbol{\tau}|H) d\pi(H, \beta|\mathbf{Y}, \mathbf{X}). \quad (27.30)$$

Considering $H \sim P(\alpha, H_0)$, the internal integral of Eq. (27.30) can be approximated by,

$$\int \phi(W_{(n+1)l}|\mathbf{x}'_{(n+1)}\beta, \boldsymbol{\tau}) d\pi(\boldsymbol{\tau}|H) \approx \sum_{k=1}^{N} q_k \phi(W_{(n+1)l}|\mathbf{x}'_{(n+1)}\beta, \tau(Z_k)). \quad (27.31)$$

The approximated predictive distribution $f(W_{(n+1)l}|\mathbf{Y}, \mathbf{X})$ follows,

$$\frac{1}{b} \sum_{b=1}^{B} \sum_{k=1}^{N} q_k^{(b)} \phi(W_{(n+1)l}|\mathbf{x}'_{(n+1)}\beta^{(b)}, \tau(Z_k^{(b)})), \quad (27.32)$$

such that $(\mathbf{q}^{(b)}, \mathbf{Z}^{(b)}, \beta^{(b)})$ is the b^{th} element of the sample generated by Gibbs sampler.

From Eqs. (27.30) and (27.31), the predictive distribution for $Y_{(n+1)l}$ is given by,

$$P(Y_{(n+1)l} = 1|\mathbf{Y}, \mathbf{X}) \approx \frac{1}{b} \sum_{b=1}^{B} \sum_{k=1}^{N} q_k^{(b)} \Phi\left(\mathbf{x}'_{(n+1)}\beta^{(b)}|0, \tau(Z_k^{(b)})\right). \quad (27.33)$$

27.4.1 Conditional Predictive Ordinate

The ith conditional predictive ordinate (CPO) is constructed on $S_i = \sum_{l=1}^{n_i} Y_{il}$ which follows $binomial(n_i, p_i)$ distributed with $p_i = F(\mathbf{x}'_i\beta)$.

Let $\mathbf{S}_{[i]}$ the vector of variables $S_1, \ldots, S_L$ excluding Y_i, then,

$$CPO_i = P(S_i = s_i|\mathbf{S}_{[i]}) = \frac{1}{\frac{1}{b} \sum_{b=1}^{B} \left[P(S_i = s_i|p_i^{(b)})\right]^{-1}}, \quad (27.34)$$

with

$$p_i^{(b)} = \sum_{k=1}^{N} q_k^{(b)} \Phi(\mathbf{x}'_{(n+1)}\beta^{(b)}|0, \tau(Z_k^{(b)}). \quad (27.35)$$

Then, the sum of the logged CPOs (SLCPO) can be an estimator for the logarithm of the marginal likelihood of the model.

Table 27.1 Beetle data—Markov chain Monte Carlo (MCMC) conditions for the Bayesian models

Model	Burn in	Thin
Normal BP	2400	11
Normal BSP fixed α	5125	21
Normal BSP random α	3000	22
Logistic BP	4420	20
Logistic BSP fixed α	15,542	81
Logistic BSP random α	13,829	82

BP Bayesian parametric, *BSP* Bayesian semiparametric

27.5 Beetle Data

A study of beetle mortality after 5 h of exposure to gaseous carbon disulphide is reported in [6]. It is a classical data set which was originally fitted through a *normal* model.

The Bayesian parametric (BP) and Bayesian semiparametric (BSP) approaches were considered for fixed $\alpha = 1$ and random $\alpha \sim gamma(4, 2)$, which confirms that the degree of faith about F_0 is minimum or the expected number of clusters from Eq. (27.8) is 6.75 for τ and 11.51 for fixed and random α, respectively. In this way, the *normal* and *logistic* models were considered to fit the data.

The $\boldsymbol{\beta}$ prior distribution is defined as $normal(0, 1/1000)$ and the number of components for the approximation of *Dirichlet* process is $N = 100$ such that the quality of approximation is verified through Eq. (27.11) in both cases.

The convergence of Markov chain Monte Carlo (MCMC) was evaluated for $\boldsymbol{\beta}$ considering the statistic introduced by [7] with a thin corresponding the autocorrelation function smaller than 0.2. From this strategy, the MCMC conditions are presented in Table 27.1. The estimates for $\boldsymbol{\beta}$ and its 95 % credibility intervals for the models are presented in Table 27.2.

It is possible to see that the Bayesian semiparametric approach requires more computational effort than the Bayesian parametric approach, which is expected since the former is a more general modeling. Moreover, the *logistic* models demand greater burn-in and thin than *normal* models.

The estimates are very similar for each class of models independent of the approach, but the credibility intervals are larger for Bayesian semiparametric models. There is no significant difference between the semiparametric models with fixed and random α parameters.

Finally, the Bayesian semiparametric models present smaller SLCPO in both classes of models. In particular, the *logistic* models take more advantages of the semiparametric approach, as can be seen in Figs. 27.1, 27.2, and 27.3.

Table 27.2 Beetle data—Estimates for the Bayesian logistic models

Model	Estimate	95 % CI	SLCPO
Normal BP			−19.126
β_0	−8.683	(−10.246 ; −7.252)	
β_1	0.146	(0.122 ; 0.171)	
Normal BSP fixed α			−18.482
β_0	−8.699	(−10.669 ; −6.977)	
β_1	0.146	(0.118 ; 0.180)	
Normal BSP random α			−18.558
α	0.541	(0.135 ; 1.268)	
β_0	−8.753	(−10.764 ; −7.078)	
β_1	0.147	(0.119 ; 0.181)	
Logistic BP			−22.846
β_0	−12.028	(−14.481 ; −9.933)	
β_1	0.202	(0.166 ; 0.240)	
Logistic BSP fixed α			−18.618
β_0	−12.236	(−16.526 ; −8.607)	
β_1	0.206	(0.145 ; 0.278)	
Logistic BSP random α			−18.410
α	0.577	(0.146 ; 1.402)	
β_0	−12.335	(−16.869 ; −8.665)	
β_1	0.208	(0.146 ; 0.284)	

CI confidence interval, *SLCPO* sum of the logged conditional predictive ordinate, *BP* Bayesian parametric, *BSP* Bayesian semiparametric

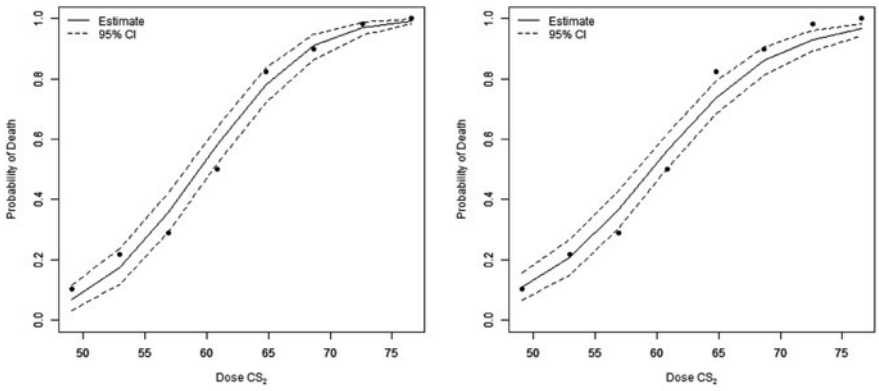

Fig. 27.1 *Normal* and *logistic* Bayesian parametric models

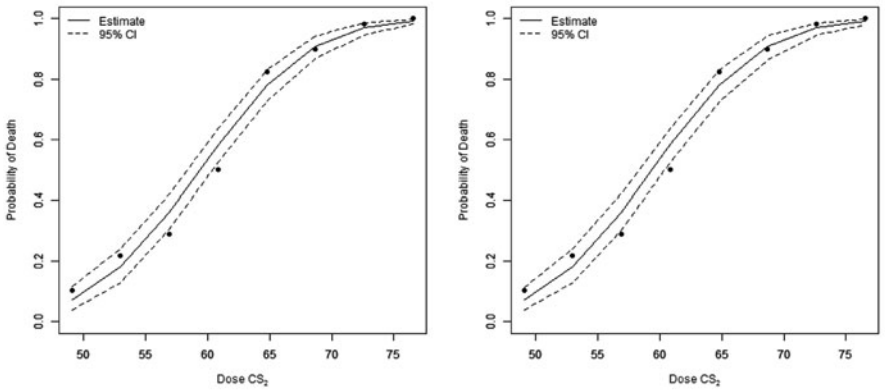

Fig. 27.2 *Normal* and *logistic* Bayesian semiparametric with fixed α models

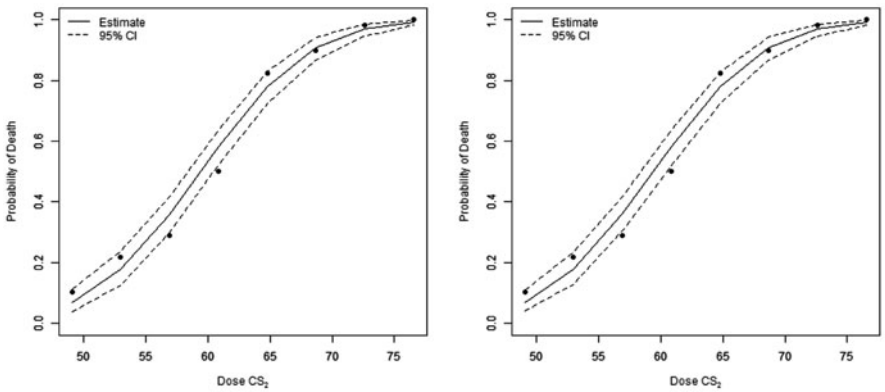

Fig. 27.3 *Normal* and *logistic* Bayesian semiparametric with random α models

27.6 Concluding Remarks

This work presents a Bayesian semiparametric model for binary data that is more interesting than the one in [4] because of the use of blocked Gibbs sampler, which provides a more general framework than previous works based on the parametric approach or using the Polya urn Gibbs sampler.

The semiparametric approach allows us to incorporate the uncertainty around the F distribution of latent data and to model heavy-tailed curves. The *logistic* Bayesian semiparametric model allows one to elicit prior distribution for regression coefficients through odds ratios information without losing the flexibility of modeling heavy-tailed or light-tailed distributions.

Further work should be to expand the modeling to encompass asymmetric distributions.

References

1. Albert, J.H., Chib, S.: Bayesian analysis of binary and polychotomous response data. J. Am. Stat. Assoc. **88**(422), 669–679 (1993)
2. Andrews, D.F., Mallows, C.L.: Scale mixtures of normal distributions. J. R. Stat. Soc. Ser. B (Methodological). **36**(1), 99–102 (1974)
3. Antoniak, C.E.: Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. Ann. Stat. **2**(6), 1152–1174 (1974)
4. Basu, S., Mukhopadhyay, S.: Binary response regression with normal scale mixture links. In: Dey, D.K., Ghosh, S.K., Mallick, B.K. (eds.) Generalized Linear Models: a Bayesian Perspective, pp. 231–241. Marcel Dekker, New York (1998)
5. Blackwell, D., MacQueen, J.B.: Ferguson distributions via pólya urn schemes. Ann. Stat. **1**(2), 353–355 (1973)
6. Bliss, C.I.: The calculation of the dosage-mortality curve. Ann. Appl. Biol. **22**(1), 134–167 (1935)
7. Brooks, S.P., Gelman, A.: General methods for monitoring convergence of iterative simulations. J. Comput Graph. Stat. **7**(4), 434–455 (1998)
8. Chen, M.H., Dey, D.K.: Bayesian modeling of correlated binary responses via scale mixture of multivariate normal link functions. Sankhyā: Indian J. Stat., Ser. A **60**(3), 322–343 (1998)
9. Connor, R.J., Mosimann, J.E.: Concepts of independence for proportions with a generalization of the Dirichlet distribution. J. Am. Stat. Assoc. **64**(325), 194–206 (1969)
10. Cox, D.R.: The Analysis of Binary Data. Chapman and Hall, London (1970)
11. Devroye, L.: Non-uniform Random Variate Generation. Springer, Berlin (1986)
12. Escobar, M.D.: Estimating normal means with a Dirichlet process prior. J. Am. Stat. Assoc. **89**(425), 268–277 (1994)
13. Escobar, M.D., West, M.: Computing nonparametric hierarchical models. In: Dey, D.K., Müller, P., Sinha, D. (eds.) Practical Nonparametric and Semiparametric Bayesian Statistics, pp. 1–22. Springer, New York (1998)
14. Ferguson, T.S.: A Bayesian analysis of some nonparametric problems. Ann. Stat. **1**(2), 209–230 (1973)
15. Ishwaran, H., Zarepour, M.: Markov chain Monte Carlo in approximate Dirichlet and Beta two-parameter process hierarchical models. Biometrika **87**(2), 371–390 (2000)
16. MacEachern, S.N.: Estimating normal means with a conjugate style Dirichlet process prior. Commun. Stat.-Simul. Comput. **23**(3), 727–741 (1994)
17. MacEachern, S.N., Müller, P.: Estimating mixture of Dirichlet process models. J. Comput. Graph. Stat. **7**(2), 223–238 (1998)
18. Robert, C.P.: Simulation of truncated normal variables. Stat. Comput. **5**(2), 121–125 (1995)
19. Sethuraman, J.: A constructive defintion of Dirichlet priors. Stat. Sin. **4**, 639–650 (1994)
20. Wong, T.T.: Generalized Dirichlet distribution in Bayesian analysis. Appl. Math. Comput. **97**(2), 165–181 (1998)

Chapter 28
Assessing a Spatial Boost Model for Quantitative Trait GWAS

Ian Johnston, Yang Jin and Luis Carvalho

Abstract Bayesian variable selection provides a principled framework for incorporating prior information to regularize parameters in high-dimensional large-p-small-n regression models such as genomewide association studies (GWAS). Although these models produce more informative results, researchers often disregard them in favor of simpler models because of their high computational cost. We explore our recently proposed spatial boost model for GWAS on quantitative traits to assess the computational efficiency of a more representative model. The spatial boost model is a Bayesian hierarchical model that exploits spatial information on the genome to uniquely define prior probabilities of association of genetic markers based on their proximities to relevant genes. We propose analyzing large data sets by first applying an expectation–maximization filter to reduce the dimensionality of the space and then applying an efficient Gibbs sampler on the remaining markers. Finally we conduct a thorough simulation study based on real genotypes provided by the Wellcome Trust Case Control Consortium and compare our model to single association tests.

28.1 Introduction

Unique sequences of genetic markers called single nucleotide polymorphisms (SNPs) in a sample of a person's DNA define a fingerprint that reveals information about that person's physical traits. In genomewide association studies (GWAS) researchers collect DNA samples and data about traits from many individuals and apply statistical methods to identify which markers may affect traits of interest related to human health. By knowing which markers significantly affect a particular trait, researchers

I. Johnston (✉)· Y. Jin · L. Carvalho
Boston University, 111 Cummington Mall, Boston, MA 02215, USA
e-mail: ianj@math.bu.edu

Y. Jin
e-mail: yangjin@bu.edu

L. Carvalho
e-mail: lecarval@math.bu.edu

A. Polpo et al. (eds.), *Interdisciplinary Bayesian Statistics,*
Springer Proceedings in Mathematics & Statistics 118, DOI 10.1007/978-3-319-12454-4_28

are able to understand more about that trait and possibly discover ways to control it, for instance, through the use of specialized drugs.

Since successful studies could lead to cures for diseases or even personalized medicine, GWAS is a popular and relevant topic in statistical genetics; however, the search for significant markers is challenging because the number of SNPs for each individual, p, is usually much larger than the sample size, n. Researchers have tried modeling these data jointly by using techniques like penalized regression or Bayesian variable selection to regularize the model parameters (for instance, see [5, 12] and the references therein), but it is computationally intensive to fit these models and thus to search for the optimal penalty term(s) or prior distribution(s).

Although these models offer a better representation of the trait as a function of all of the markers, researchers still most commonly avoid the computational burden of fitting them jointly by instead fitting a simple regression model to each SNP separately and computing a measure of genomewide statistical significance using a multiple testing correction procedure. However, in addition to the problem of ignoring joint effects of markers when applying these single SNP analyses, the threshold for genomewide significance may differ across studies based on the number of markers in each data set, and so it is difficult to compare different sets of results. Ideally, we would like to have a fast way of fitting a representative model to GWAS data that produces informative results which are communicable across studies.

In this work we strive toward that ideal and explore the computational efficiency of our recently proposed pipeline for efficiently analyzing GWAS data using the spatial boost (SB) model on quantitative traits in a simulation study based on real SNP data provided by the Wellcome Trust Case Control Consortium. As a special case of Bayesian variable selection, our model uniquely defines the prior probability of association of each SNP as a function of its proximity to relevant genes. We define our model in Sect. 28.2, outline our method of conducting inference in Sect. 28.3, describe the setup for our simulation study in Sect. 28.4, discuss our results in Sect. 28.5, and then offer some concluding remarks in Sect. 28.6.

28.2 Spatial Boost Model

We model the expected value of the ith individual's quantitative trait, $\mathbb{E}[y_i]$, as a linear combination of the number of alleles present at a set of p SNPs encoded in $x_i^\top \in \{0, 1, 2\}^p$, and model the phenotypic variation that is not attributed to the genotypes as τ^2. Given a vector of coefficients, β, we thus have:

$$y|\beta, \tau^2 \sim \mathrm{N}(X\beta, \tau^2 I_n). \tag{28.1}$$

We consider a continuous version of the spike-and-slab prior distribution [6] for β to derive a simpler expectation–maximization (EM) algorithm that we use to reduce the dimensionality of a data set by sequentially filtering out SNPs that have relatively low posterior probability of being associated with the trait. Each β_j conditional on

an indicator variable, θ_j, a random variance term σ^2, and a tuning parameter, κ, is a priori independently distributed with the following normal distribution:

$$\beta_j|\theta_j, \sigma^2 \sim \mathrm{N}(0, \sigma^2[\theta_j\kappa + 1 - \theta_j]). \tag{28.2}$$

We set κ to be a large number, e.g., $\kappa = 10^3$, to enforce a separation between the variances of the spike and the slab components of this prior distribution. We could also consider choosing κ by controlling a metric such as the Bayesian false discovery rate. In practice, we may add additional covariates to the model; however, we currently only add an intercept column to X with coefficient $\beta_0 \sim \mathrm{N}(0, \sigma^2\kappa)$.

For each of the p genotypes, we assign a unique prior distribution to the indicator variable, θ_j, based on the proximity of the jth SNP to relevant genes as measured in weights given by $w_j^\top(\phi)$ and relevances given by r:

$$\theta_j \sim \mathrm{Bern}(\mathrm{logit}^{-1}[\xi_0 + \xi_1 \cdot w_j^\top(\phi)r]). \tag{28.3}$$

Given the genomic position of the jth SNP, s_j, and the transcription start and end sites of the gth gene, g_l and g_r respectively, we compute the gth gene's weight afforded to the jth SNP as follows:

$$w_{j,g} = \int_{g_l}^{g_r} \frac{1}{\sqrt{2\pi\phi^2}} \exp[-0.5 \cdot (x - s_j)^2/\phi^2]\, dx. \tag{28.4}$$

We normalize all gene weights so that $\max_{j,g}\{w_{j,g}\} = 1$ to enforce consistency in the meaning of ξ_1 across models as the maximum increase in the log odds of a SNP being associated to a trait due to the gene hierarchy. We set ξ_0 to be a value that encodes our prior belief about the percentage of SNPs that are likely to be associated with a particular trait without any gene boost. Researchers have observed that biological phenomena like linkage disequilibrium lead to an inverse relationship between the average correlation between two SNPs and the distance between them [1]. We therefore recommend choosing ϕ by minimizing the mean squared error between the magnitudes of the sample pairwise correlations and $p_{i,j}$ where

$$p_{i,j} = 2 \cdot \int_{|s_i - s_j|}^{\infty} \frac{1}{\sqrt{2\pi\phi^2}} \exp[-0.5 \cdot x^2/\phi^2]\, dx. \tag{28.5}$$

Each $p_{i,j}$ aims to capture an inverse relationship between the genomic distance and the strength of the correlation between the ith and jth SNPs.

In practice we may choose r to be informative; however, for this chapter we simply set r to be the noninformative vector $\mathbf{1}$. Finally we assign independent inverse-gamma prior distributions for τ^2 and σ^2 with respective hyperparameters ν_1, λ_1 and ν_2, λ_2. We recommend choosing ν_1 and λ_1 so that τ^2 has a noninformative prior distribution and choosing ν_2 and λ_2 so that σ^2 has an informative prior distribution that is concentrated around a small value, e.g., 10^{-4}.

28.3 Inference

We want to use the centroid estimator [4] to conduct inference on θ and so we must compute $\mathbb{P}(\theta_j = 1|y)$. However, to speed up the analysis of large data sets, we first treat the θ_j as latent variables and derive an EM algorithm to obtain estimates $\beta_j^*, \sigma^{2^*}, \tau^{2^*}$ and approximate $\mathbb{P}(\theta_j = 1|y) \approx \mathbb{P}(\theta_j = 1|\beta_j^*, \sigma^{2^*}, \tau^{2^*}, y)$. We then filter SNPs by ranking $\mathbb{P}(\theta_j = 1|\beta_j^*, \sigma^{2^*}, \tau^{2^*}, y)$ in descending order and removing the bottom quartile. We repeat this process until we either reach a desired smaller number of SNPs or until the predictive accuracy of our model deteriorates beyond a certain point. Finally, we compute estimates of $\mathbb{P}(\theta_j = 1|y)$ for the remaining SNPs using a Gibbs sampler.

28.3.1 Expectation–Maximization Filter

Our algorithm is similar to a recently proposed EM approach to Bayesian variable selection [10]. Omitting the superscripts (t) to denote the tth iteration of the algorithm, in the E-step we compute $\mathbb{E}[\theta_j|\beta_j, \sigma^2, \tau^2, y] = \text{logit}^{-1}(S_j)$ where:

$$S_j = \xi_0 + \xi_1 \cdot w_j^\top(\phi)r + 0.5 \cdot \beta_j^2(\kappa - 1)/[\sigma^2\kappa] - 0.5 \cdot \log(\kappa). \tag{28.6}$$

In the M-step we optimize the other random variables in the model using the complete data log likelihood and the current values of $\sigma_{\theta_j}^{-2} = [\text{logit}^{-1}(S_j)/\kappa + 1 - \text{logit}^{-1}(S_j)]/\sigma^2$. Letting $\Sigma_\theta^{-1} = \text{Diag}(\sigma_{\theta_j}^{-2})$, we update β as follows:

$$\beta = (\Sigma_\theta^{-1} + \tau^{-2}X^\top X)^{-1}(\tau^{-2}X^\top y). \tag{28.7}$$

We update τ^2 and σ^2 using the modes of their respective posterior distributions:

$$\tau^2 = (\lambda_1 + 0.5 \cdot \sum_{i=1}^{n}(y_i - x_i^\top\beta)^2)/(\nu_1 + n/2 + 1), \tag{28.8}$$

$$\sigma^2 = \frac{\lambda_2 + 0.5 \cdot (\beta_0^2/\kappa + \sum_{j=1}^{p}\beta_j^2[\text{logit}^{-1}(S_j)/\kappa + 1 - \text{logit}^{-1}(S_j)])}{\nu_2 + p/2 + 1}. \tag{28.9}$$

We exploit a truncated singular value decomposition (SVD) to speed up the computation in (28.7) by replacing X with an approximation $\sum_{l=1}^{k} u_{(l)}d_{(l)}v_{(l)}^\top$. By applying the Kailath variant matrix inverse identity, we can substitute the inversion of a p-by-p matrix with the inversion of an k-by-k matrix.

28.3.2 Gibbs Sampler

We derive the conditional posterior distributions of β, τ^2, and σ^2 as follows:

$$\beta|\theta,\sigma^2,y \sim \mathrm{N}[(\Sigma_\theta^{-1}+\tau^{-2}X^\top X)^{-1}(\tau^{-2}X^\top y),(\Sigma_\theta^{-1}+\tau^{-2}X^\top X)^{-1}], \quad (28.10)$$

$$\tau^2|y,\beta \sim \mathrm{IG}(\nu_1+n/2,\lambda_1+0.5\cdot\sum_{i=1}^{n}[y_i-x_i^\top\beta]^2), \quad (28.11)$$

$$\sigma^2|\beta,\theta \sim \mathrm{IG}(\nu_2+p/2, \quad (28.12)$$

$$\lambda_2+0.5(\beta_0^2/\kappa+\sum_{j=1}^{p}\beta_j^2[\mathrm{logit}^{-1}(S_j)/\kappa+1-\mathrm{logit}^{-1}(S_j)])).$$

We then use Eq. (28.6) to compute $P(\theta_j=1|\beta_j,\sigma^2,\tau^2,y)$ and derive the conditional posterior distribution of each θ_j:

$$\theta_j|\beta_j,\sigma^2 \sim \mathrm{Bern}[\mathrm{logit}^{-1}(S_j)]. \quad (28.13)$$

After initializing the values for β, τ^2, σ^2, and θ, we draw samples sequentially from (28.10), (28.11), (28.12), and (28.13) until we have reached a desired total number of samples for each random variable. In practice, we generate several chains of posterior samples and assess convergence using the Brooks and Gelman scale reduction factor [3] on the complete data log likelihood. We compute our final estimates of $\mathbb{P}(\theta_j=1|y)$ for each SNP using N posterior samples as $\hat{\mathbb{P}}(\theta_j=1|y)=\sum_{t=1}^{N}\theta_j^{(t)}/N$.

28.4 Simulation Setup

We generate 100 matrices of size $n=10^2$ and $p=10^3$ by randomly selecting contiguous blocks of genotypes from an overall list of 29,711 SNPs on chromosome 2 in 3503 individuals in a data set provided by the Wellcome Trust Case Consortium. We only consider common variants in our analyses, i.e., SNPs with minor allele frequency >0.05 and variants that do not statistically significantly deviate from Hardy–Weinberg equilibrium [11]. We choose ϕ using (28.5), and set $r=\mathbf{1}$. After normalizing the gene weights given in (28.4) so that the maximum value in each data set is 1, the distribution of all gene weights is heavily left-skewed with 97.2 % of the values occurring below 0.5. In our first simulation study we start by setting $\sigma^2=10^{-4}$ and $\tau^2=10^2$ and then sample values for θ, β, and y for all 100 data sets under 6 different gene boost and heritability combinations. For each replicate s, we highlight the effect of the gene boost by considering both a boostless model with $\xi_0=\mathrm{logit}(10/p_s)$ and $\xi_1=0$ as well as a model with $\xi_0=\mathrm{logit}(1/p_s)$ and $\xi_1=-\mathrm{logit}(1/p_s)$ where p_s is the number of SNPs in the sth data set. We enforce

consistency in the number of true positives across data sets by sampling values for θ such that $\sum_{j=1}^{p_s} \theta_j = 10$. Heritability, h^2, is the proportion of variation in a trait that is explained by the variation in the genotypes. Assuming that $X_{i,j} \sim \text{Binom}(2, \pi_j)$ independently where π_j is the minor allele frequency of the jth SNP, we consider an approximation for h^2 as follows:

$$h^2 \approx \frac{\mathbb{E}_X\left[\kappa\sigma^2 \sum_{j:\theta_j=1} X_{i,j}^2\right]}{\mathbb{E}_X\left[\kappa\sigma^2 \sum_{j:\theta_j=1} X_{i,j}^2 + \sigma^2 \sum_{j:\theta_j=0} X_{i,j}^2 + \tau^2\right]}. \tag{28.14}$$

We select κ for each data set in our simulations using Eq. (28.14) to ensure a desired level of $h^2 \in \{0.1, 0.5, 0.9\}$. In our study this corresponds to choosing average values of $\kappa \in \{1.5 \times 10^4, 1.4 \times 10^5, 1.3 \times 10^6\}$ respectively. After simulating values for β and y we first apply our EM filtering algorithm to reduce the number of SNPs in each data set to a consistent 300 and then run our Gibbs sampler on the retained set of markers to obtain final estimates of $P(\theta_j = 1|y)$ using $N = 1500$. In the EM filtering step we try using X as well as three different truncated SVD approximations to X where the mean squared error (MSE) tolerance is either 1, 10, or 25 %. For comparison we run the usual association tests on our simulated data using the PLINK [9] software.

Since κ explicitly controls the difference in variability of $\beta_j|\theta_j, \sigma^2$ and thus greatly influences our variable selection, we investigate the sensitivity of our model to misspecifications of κ when all other model tuning parameters are ideally set. We use the first 300 consecutive SNPs in each data set and define $\sigma^2 = 10^{-4}$, $\tau^2 = 10^2$, $\xi_0 = \text{logit}(1/300)$, and $\xi_1 = -\text{logit}(1/300)$ and again sample θ such that we have ten true positives. We consider true values of $\kappa \in \{10^3, 10^5\}$ and compute estimates of $\mathbb{P}(\theta_j = 1|y)$ for each SNP after running our Gibbs sampler for $N = 1500$ iterations in 7 different models where we set $\kappa \in \{10^1, 10^2, \ldots, 10^7\}$.

Moreover, since ξ_1 determines the strength of the influence of neighboring genes on θ_j, we also investigate the sensitivity of our model to misspecifications of it. We use the same setup as above but instead set $\kappa = 10^3$, consider true values of $(\xi_0, \xi_1) \in \{(\text{logit}(10/300), 0), (\text{logit}(1/300), -\text{logit}(1/300))\}$, and fit seven different models where we set $\xi_1 \in \{0, 1, \ldots, 6\}$. In each of our simulation studies, we set $\nu_1 = 1.1$, $\lambda_1 = 10$, $\nu_2 = 101$, and $\lambda_2 = 10^{-2}$ and assess the model performance by computing the area under the receiver operating characteristic (ROC) curve, or area under the curve (AUC) [2], using our knowledge of the true and false positives.

28.5 Results

In our first simulation study we observe in Fig. 28.1 that the SB model outperforms the single SNP tests across all heritability profiles when there is a gene boost using either X or one of three SVD approximations to X with MSE tolerances of 1, 10, and 25 %. When there is no gene boost, our model suffers due to the potential sequential loss of true positive weak signals during the EM filtering step, and thus achieves an

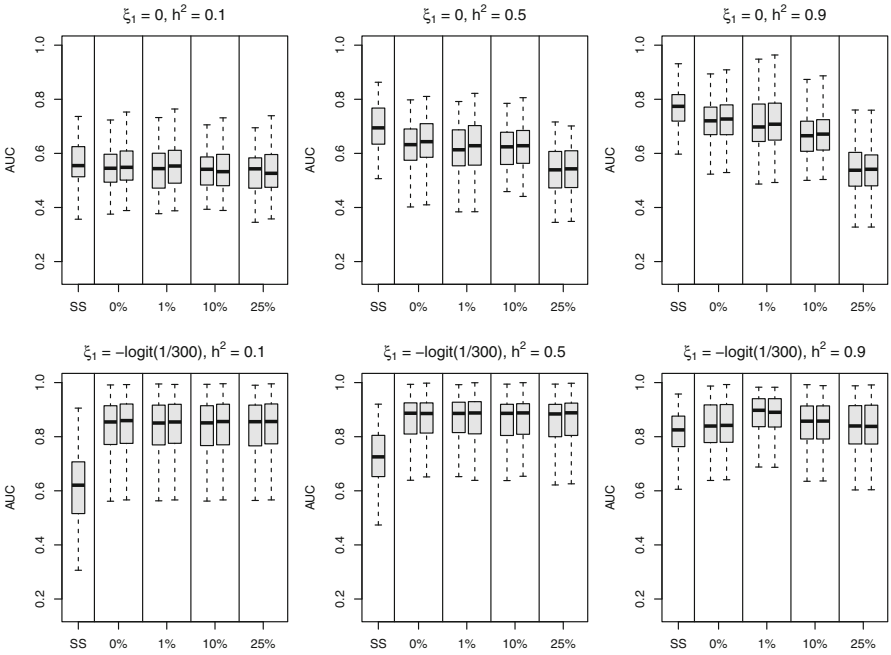

Fig. 28.1 These boxplots depict the performance of the single SNP tests (SS) and the SB model across 6 different gene boost and heritability combinations and 100 different genotype patterns. The percentages indicate the tolerance on MSE that we required when replacing X with an approximation. For each set of SB model results, we present a boxplot (*left*) for the AUC values based on the final estimates of $\text{logit}^{-1}(S_j)$ after running the EM filter and a boxplot (*right*) for the AUC values based on the final estimates of $\mathbb{P}(\theta_j = 1|y)$ after running the Gibbs sampler

average performance similar to the single SNP tests across all heritability profiles when using either X or an approximation with an MSE tolerance of 1 %. Moreover, as expected, the performance deteriorates when using a coarser approximation for traits with moderate and high heritabilities since the variation in the genotypes explains more of the variation in y. Interestingly, we can achieve roughly the same level of performance by computing AUC using the final estimates of $\text{logit}^{-1}(S_j)$ after running the EM filter in place of the final estimates of $\mathbb{P}(\theta_j = 1|y)$ after running our Gibbs sampler. Based on the running times for each aspect of the SB model and the single SNP tests across several different configurations of n and p given in Table 28.1, we see that after computing the SVD of X, it is often *faster* to run a single pass of our EM filter on a coarse approximation to X (MSE tolerance of 25 %) than to fit the single SNP tests. For the largest data size we considered ($n = 10^3$, $p = 10^4$), we see reductions in the time it takes to run the EM filter 5 times by 33.2, 80.7, and 97.3 % when using MSE tolerances of 1, 10, and 25 % respectively. In a few cases, it takes slightly longer to run the EM filter when using a fine approximation to X, e.g., MSE tolerance of 1 %, possibly due to the extra memory needed to store three matrices instead of one.

Table 28.1 We give the mean running times and corresponding standard deviations (in *parentheses*) in minutes for the SB model and the single SNP tests in R using ten replicates

Task (n, p) :	$(10^2, 10^3)$	$(10^2, 10^4)$	$(10^3, 10^3)$	$(10^3, 10^4)$
Compute SVD with irlba [7]	0.35 (0.00)	3.43 (0.08)	1.16 (0.00)	124.90 (4.51)
EM filter on X after running the first pass after retaining 25 % of p	0.02 (0.00) 0.04 (0.00)	10.36 (0.03) 17.81 (0.03)	0.12 (0.00) 0.39 (0.01)	31.23 (0.25) 91.26 (0.49)
EM filter on SVD (1 % MSE) after running the first pass after retaining 25 % of p	0.03 (0.00) 0.15 (0.00)	1.99 (0.01) 3.64 (0.04)	0.13 (0.00) 1.27 (0.01)	33.88 (0.12) 60.95 (1.28)
EM filter on SVD (10 % MSE) after running the first pass after retaining 25 % of p	0.01 (0.00) 0.02 (0.00)	0.77 (0.00) 1.64 (0.01)	0.01 (0.00) 0.04 (0.00)	7.66 (0.00) 17.57 (0.33)
EM filter on SVD (25 % MSE) after running the first pass after retaining 25 % of p	0.00 (0.00) 0.01 (0.00)	0.28 (0.01) 0.83 (0.02)	0.00 (0.00) 0.01 (0.00)	1.47 (0.01) 2.48 (0.01)
Gibbs sampler on X with $N = 1500$	9.00 (0.03)	6626.99 (33.98)	9.32 (0.04)	7612.43 (94.71)
Single SNP tests	0.05 (0.00)	0.53 (0.01)	0.07 (0.00)	0.73 (0.02)

SVD singular value decomposition, *EM* expectation–maximization, *MSE* mean squared error, *SNP* single nucleotide polymorphism

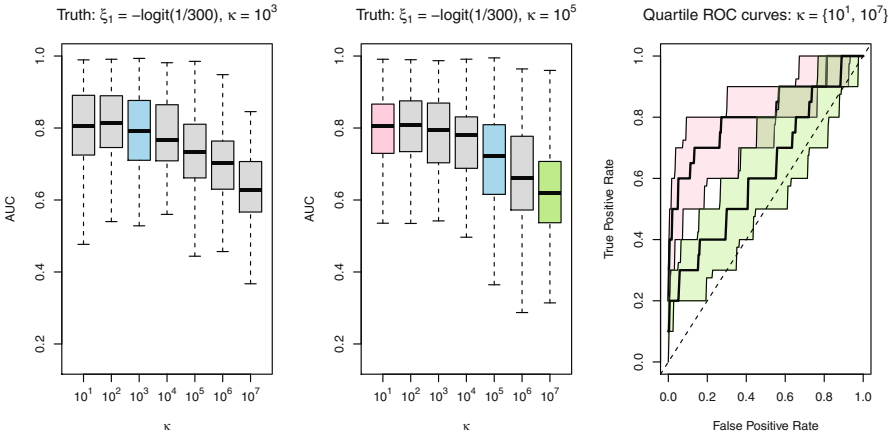

Fig. 28.2 These boxplots depict the performance of the SB model in our second simulation study where we vary κ and fit our model to 100 data sets simulated from 2 different models where $\kappa = 10^3$ (*left*) and $\kappa = 10^5$ (*middle*). The *blue boxplots* show the results when all parameters are ideally set. In the *right plot*, we explore the distribution of ROC curves that generated the AUC values for the first and last boxplots in the *middle plot*

In our second simulation study, we observe better performances from our model in Fig. 28.2 when we choose $\kappa \leq 10^4$ even if the true value of κ is larger. This is likely due to the difficulty in detecting both weak and strong signals simultaneously

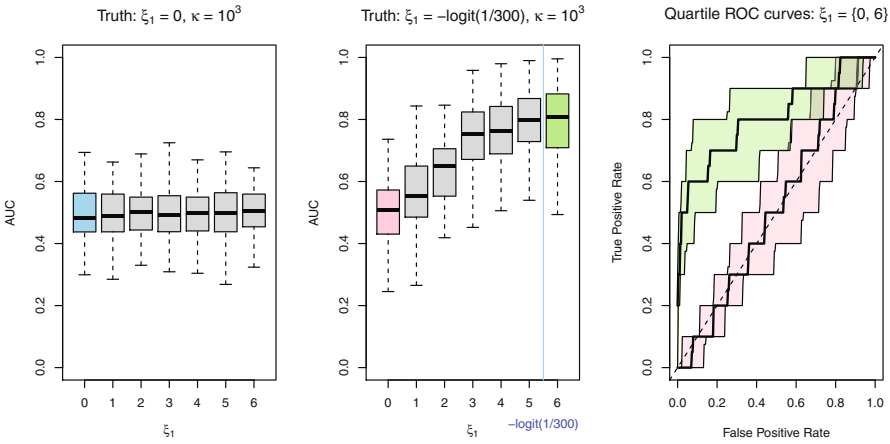

Fig. 28.3 These boxplots depict the performance of the SB model in our third simulation study where we vary ξ_1 and fit our model to 100 data sets simulated from 2 different models where $\xi_1 = 0$ (*left*) and $\xi_1 = -\text{logit}(1/300)$ (*middle*). The *blue boxplot* shows the results when all parameters are ideally set. In the *right plot*, we explore the distribution of ROC curves that generated the AUC values for the first and last boxplots in the *middle plot*

when using a large value for κ. By selecting a relatively smaller value for κ we opt for sensitivity rather than specificity. When viewing the quartiles of the distribution of points on all 100 ROC curves for the two special cases when we select $\kappa \in \{10^1, 10^7\}$ in data sets where $\kappa = 10^5$, we do not see any benefit from being more specific in the early part of the curve by choosing $\kappa = 10^7$. In our third simulation study, we observe in Fig. 28.3 that the SB model is robust to misspecifications of ξ_1 when there is no gene boost, but is sensitive to misspecifications otherwise.

28.6 Conclusions

We find that in a variety of gene boost and heritability configurations, our pipeline for analyzing GWAS data sets using the SB model is an efficient way of fitting a representative model to SNPs *jointly* that exploits proximities to relevant genes to uniquely define prior probabilities of association. Although it takes an impractical amount of time to run our Gibbs sampler, we achieve the same level of performance at a reasonable fraction of that computational cost by settling for the final estimates of $\text{logit}^{-1}(S_j)$ after running our EM filter in place of the final estimates of $\mathbb{P}(\theta_j = 1|y)$ after running the Gibbs sampler. Computing the SVD of X is the next largest computational cost when using our model; however, researchers may already perform such a computation when they apply principal components analysis to genotype data for instance to adjust for population stratification [8] before any subsequent analysis. To maintain a competitive edge when analyzing whole genomes in the future, we may further benefit from analyzing chromosomes in blocks defined based on genomic distance or linkage disequilibrium.

Acknowledgments IJ and LC were partially supported by NSF grant DMS-1107067. This study makes use of data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from `wtccc.org.uk`.

References

1. Ardlie, K.G., Kruglyak, L., Seielstad, M.: Patterns of linkage disequilibrium in the human genome. Nat. Rev. Genet. **3**(4), 299–309 (2002)
2. Bradley, A.P.: The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit. **30**(7), 1145–1159 (1997)
3. Brooks, S.P., Gelman, A.: General methods for monitoring convergence of iterative simulations. J. Comput. Graph. Stat. **7**(4), 434–455 (1998)
4. Carvalho, L.E., Lawrence, C.E.: Centroid estimation in discrete high-dimensional spaces with applications in biology. Proc. Natl. Acad. Sci. **105**(9), 3209–3214 (2008)
5. Guan, Y., Stephens, M.: Bayesian variable selection regression for genome-wide association studies and other large-scale problems. Ann. Appl. Stat. **5**(3), 1780–1815 (2011)
6. Ishwaran, H., Rao, J.S.: Spike and slab variable selection: frequentist and Bayesian strategies. Ann. Stat. **33**(2), 730–773 (2005)
7. Lewis, B.: irlba: Fast partial SVD by implicitly-restarted Lanczos bidiagonalization. R package version 0.1 **1**, 1520 (2009)
8. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., Reich, D.: Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. **38**(8), 904–909 (2006)
9. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., De Bakker, P.I., Daly, M.J., et al.: PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. **81**(3), 559–575 (2007)
10. Ročková, V., George, E.I.: EMVS: The EM approach to Bayesian variable selection. J. Am. Stat. Assoc. **109**(506), 828–846 (2014)
11. Wigginton, J.E., Cutler, D.J., Abecasis, G.R.: A note on exact tests of Hardy-Weinberg equilibrium. Am. J. Hum. Genet. **76**(5), 887–893 (2005)
12. Wu, T.T., Chen, Y.F., Hastie, T., Sobel, E., Lange, K.: Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics **25**(6), 714–721 (2009)

Chapter 29
The Exponential-Poisson Regression Model for Recurrent Events: A Bayesian Approach

Márcia A. C. Macera, Francisco Louzada and Vicente G. Cancho

Abstract In this chapter, we introduce a new regression model for recurrent event data, in which the time of each recurrence is associated to one or multiple latent causes and no information is provided about the cause responsible for the event occurrence. This model is characterized by a fully parametric rate function and it is based on the exponential-Poisson distribution. The time of each recurrence is then given by the minimum lifetime value among all latent causes. Inference aspects of the proposed model are discussed via Bayesian inference by using Markov Chain Monte Carlo (MCMC) method. A simulation study investigates the frequentist properties of the posterior estimators for different sample sizes. A real-data application demonstrates the use of the proposed model.

29.1 Introduction

Recurrent event data are often encountered in longitudinal studies involving multiple subjects. This type of data can be observed in several areas such as biomedicine, public health, engineering and reliability, demography, politics, economics, among others. Recurrent events are predominant in a wide variety of situations, and so innovative probabilistic models and appropriate statistical inference procedures should be developed for analyzing this kind of data. In this regard, several models and methods have been proposed for the analysis of recurrent event data, particularly methods based on counting process [5].

In the literature, two types of time scale are often used to analyze recurrent event data, namely time-to-events (total time) [10, 12] and time-between-events (gap time) [1, 5]. Models that analyze time-to-events data were studied based on point

M. A. C. Macera (✉)
DEs, Federal University of Sao Carlos, Sao Carlos, Brazil
e-mail: marciamacera@gmail.com

F. Louzada · V. G. Cancho
Institute of Mathematics and Computer Science, University of São Paulo—USP,
Avenida Trabalhador São-carlense, 400 - Centro, São Carlos 13566-590, SP, Brazil
e-mail: louzada@icmc.usp.br

V. G. Cancho
e-mail: garibay@icmc.usp.br

A. Polpo et al. (eds.), *Interdisciplinary Bayesian Statistics,*
Springer Proceedings in Mathematics & Statistics 118, DOI 10.1007/978-3-319-12454-4_29

process models, more specifically under a Poisson-type process assumption [3]. In this context, a variety of methods have been proposed, we can refer to [13, 15, 19, 21]. On the other hand, methods that analyze gap-time data are in general based on renewal processes [10, 16]. More general models than the renewal process also have been developed to analyze gap times of recurrent events (see e.g., [5, 8, 10]).

In this chapter, we propose a model for recurrent event data characterized by a fully parametric baseline rate function. The proposed model aims to analyze gap times between consecutive recurrences of an event of interest, and it is based on the two-parameters exponential-Poisson (EP) distribution introduced by [11], namely the distribution of the minimum among a random number (truncated Poisson distributed) of exponential times. Assuming this specific form, our model is stated on a competing risk scenario, and the time of each recurrence is then given by the minimum lifetime value among all latent causes. Hereafter, we will call it as the EP regression model for recurrent events or simply EPre model.

This chapter is organized as follows. The proposed model is described in Sect. 29.2. The inferential procedure based on a full Bayesian approach is presented in Sect. 29.3. Section 29.4 includes the results of a simulation study performed to assess the frequentist properties of the estimation procedure. A real-data analysis on a muscle soreness data set is presented in Sect. 29.5. Section 29.6 concludes the chapter.

29.2 Model Formulation

Suppose that an unit under study may experience consecutive recurrences of a single type of recurrent event. Let $T_1 < T_2 < \ldots < T_j < \ldots$, $j = 1, 2, \ldots$ be the ordered event times, which are measured from the start of the follow-up time. We are interested on the gap times $W_j = T_j - T_{j-1}$, for $j = 1, 2, \ldots$ and $T_0 = 0$.

For continuous recurrent event processes, the rate function of recurrences is defined as [5, p. 12]

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr\{N(t + \Delta t) - N(t) = 1\}}{\Delta t}, \tag{29.1}$$

where $N(t)$ denotes the number of occurrences in $(0, t]$. In addition, the cumulative rate function is given by $H(t) = \int_0^t h(s)ds$.

Following the idea of [20], in which the recurrent events follow a nonhomogeneous Poisson process, the general form of the rate function for our model is expressed as

$$h(w|t_{j-1}) = h_0(w + t_{j-1}), \tag{29.2}$$

where $h(w|t_{j-1})$ is the rate function for the recurrence process up to time $w + t_{j-1}$ and $h_0(t_j)$ is a deterministic function describing the general behavior of an unit over time, where $t_j = w + t_{j-1}$.

Now, we consider a study with n independent units. Assume that the follow-up time is subject to independent right censoring time τ_i, which is noninformative

about the T_{ij} and represents a stopping time as described in [2]. For each unit, we define the variable $K_i = \max\{k \in \mathbb{Z}_0^+ \mid T_{ik} \leq \tau_i\}$, where $\mathbb{Z}_0^+ = \{0, 1, 2, \dots\}$ and $\mathrm{P}(K_i < \infty) = 1$. The variable K_i denotes the number of event occurrences over the monitoring period $[0, \tau_i]$. We note that for the ith unit the quantity $\tau_i - T_{iK_i}$ is the right-censored value of the gap time W_{i,K_i+1}. For the ith unit, we assume that the recurrence process $N_i(w + t_{i,j-1})$ is a nonhomogeneous Poisson-type process with a rate function defined by (29.2). The exponential-Poisson model [11] is used to parametrize the function of time $h_0(w + t_{i,j-1})$.

We assume that the covariates are directly related to one of the model parameters. In addition, the covariate effect is assumed to be a linear combination of fixed variables and is defined by

$$\mathbf{X}^\top \boldsymbol{\gamma} = \gamma_0 + \sum_{p=1}^{r} \mathrm{X}_p \, \gamma_p. \tag{29.3}$$

With a exponential-Poisson form for $h_0(w + t_{i,j-1})$ and the covariates related to a particular parameter by adopting the exponential link function, we have the following rate function for the recurrence process $N_i(w + t_{i,j-1})$ of unit i

$$h_i(w|t_{i,j-1}, \mathbf{x}_i) = \frac{\lambda \beta_i \mathrm{e}^{-\beta_i(w+t_{i,j-1})}}{1 - \mathrm{e}^{-\lambda \mathrm{e}^{-\beta_i(w+t_{i,j-1})}}}, \tag{29.4}$$

where λ is a positive parameter, $\beta_i = \exp(\mathbf{x}_i^\top \boldsymbol{\gamma})$ incorporates the covariate information with $\mathbf{x}_i^\top \boldsymbol{\gamma}$ defined by (29.3) and $\boldsymbol{\gamma}$ is the parameter vector associated to the observed covariates $\mathbf{x}_i$. The proposed model incorporates a special submodel as a particular case in the survival literature. The EPre model reduces to the classical homogeneous Poisson process (HPP) as λ approaches to zero.

The cumulative rate function over $(t_{i,j-1}, t_{i,j-1} + w]$ is defined as

$$\begin{aligned} H(t_{i,j-1}, w) &= H(w + t_{i,j-1}) - H(t_{i,j-1}) = \int_0^w h_i(s|t_{i,j-1}, \mathbf{x}_i) ds \\ &= \int_0^w h_0(s + t_{i,j-1}; \boldsymbol{\theta}) ds = \int_{t_{i,j-1}}^{t_{i,j-1}+w} h_0(s; \boldsymbol{\theta}) ds, \end{aligned} \tag{29.5}$$

where $\boldsymbol{\theta} = (\lambda, \boldsymbol{\gamma})$.

Since the rate function $h_i(w|t_{i,j-1}, \mathbf{x}_i)$ in (29.4) is deterministic and integrable over $(t_{i,j-1}, t_{i,j-1} + w]$, and $H(t_{i,j-1}, w)$ is continuous, the partial survivor function of gap times $W_{ij} = T_{ij} - T_{i,j-1}$ conditional on previous recurrences is given by

$$\begin{aligned} S(w|t_{i,j-1}) &= \exp\left\{-H(t_{i,j-1}, w)\right\} \\ &= \frac{1 - \exp(\lambda \mathrm{e}^{-\beta_i(w+t_{i,j-1})})}{1 - \exp(\lambda \mathrm{e}^{-\beta_i t_{i,j-1}})}. \end{aligned} \tag{29.6}$$

The gap times $(w_{ij})_{j \geq 1}$, $i = 1, \dots, n$, are simulated using the iterative inverse transform algorithm [18]. Thus, the conditional distribution function $F(w|t_{i,j-1}) =$

$1 - S(w|t_{i,j-1})$ can be used for inversion. For a fixed covariate vector $\mathbf{x}_i$, the gap time between $(j-1)$st and jth recurrent events occurring during the subinterval $(t_{i,j-1}, t_{i,j-1} + w]$ is calculated as

$$w_{ij} = \frac{\log(\lambda) - \beta_i t_{i,j-1} - \log[\log\{u_{ij} + (1-u_{ij})e^{\lambda e^{-\beta_i t_{i,j-1}}}\}]}{\beta_i}, \tag{29.7}$$

where $\beta_i = \exp(\mathbf{x}_i^\top \boldsymbol{\gamma})$ and $(u_{ij})_{j\geq 1}$ are sequences of independent realizations of a uniform (0, 1) random variable. The data are then simulated by subsequent iterations, and the recurrent times $(t_{ij})_{j\geq 1}$ can be simulated by $t_{ij} = t_{i,j-1} + w_{ij}$, with $t_{i0} = 0$.

29.3 Statistical Inference

For statistical inference, we adopt a full Bayesian approach. The likelihood function, the prior distributions for the model parameters and the details of the Markov Chain Monte Carlo (MCMC) method are described as follows.

29.3.1 Likelihood Function

Suppose that we have n units, with the ith unit being observed over the time interval $[0, \tau_i]$, $i = 1, \ldots, n$. Moreover, suppose that an unit under study experience K_i consecutive events at times $0 < t_{i1} < \ldots < t_{iK_i} \leq \tau_i$. Let $w_{ij} = t_{ij} - t_{i,j-1}$ $(j = 1, \ldots, K_i$ and $t_{i0} = 0)$ be the observed gap times and $w_{i,K_i+1} = \tau_i - t_{iK_i}$, which is possibly censored. We assume a noninformative censoring mechanism and so, we define a censoring indicator, δ_{ij}, which is equal to zero if the gap time is right-censored and equal to one if the gap time is completely observed, $i = 1, \ldots, n$, $j = 1 \ldots, K_i + 1$. Then, the likelihood function for the vector parameter $\boldsymbol{\theta} = (\lambda, \boldsymbol{\gamma})$ from n independent units is given by

$$L(\boldsymbol{\theta}) = \left\{ \prod_{i=1}^{n} \prod_{j=1}^{K_i+1} h_i(w_{ij}|t_{i,j-1}, \mathbf{x}_i)^{\delta_{ij}} \right\} \exp\left\{ -\sum_{i=1}^{n} \sum_{j=1}^{K_i+1} \int_0^{w_{ij}} h_i(s|t_{i,j-1}, \mathbf{x}_i) ds \right\}, \tag{29.8}$$

where $h_i(\cdot \mid \cdot)$ is the unit's individual failure rate process, and the log-likelihood function is given by

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{j=1}^{K_i+1} \log\left\{ h_i(w_{ij}|t_{i,j-1}, \mathbf{x}_i)^{\delta_{ij}} \right\} - \int_0^{w_{ij}} h_i(s|t_{i,j-1}, \mathbf{x}_i) ds. \tag{29.9}$$

Adapted to the rate function given by (29.4), the log-likelihood function (29.9) can be expressed as

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{j=1}^{K_i+1} \delta_{ij}\left[\log(\lambda\beta_i) - \beta_i(w_{ij} + t_{i,j-1})\right] + \lambda e^{-\beta_i(w_{ij}+t_{i,j-1})} + (1-\delta_{ij})\log(1 - e^{-\lambda\exp(-\beta_i(w_{ij}+t_{i,j-1}))}) - \log(e^{\lambda\exp(-\beta_i t_{i,j-1})} - 1). \tag{29.10}$$

29.3.2 Sampling-Based Inference

The Bayesian method is an alternative statistical approach for the proposed model parameters estimation, $\theta = (\lambda, \boldsymbol{\gamma})$, besides, it allows the incorporation of the previous knowledge of the parameters through a prior distribution. The target distribution for inference is the posterior of the parameters of interest. For this, we need to obtain the marginal posterior densities of each parameter, which are obtained by integrating the joint posterior density with respect to each parameter. We consider a proper joint prior distribution for the model parameters, in order to ensure that the joint posterior distribution is a proper distribution [9]. However, independent of the prior distribution chosen, the joint posterior for the parameters of the proposed model is analytically intractable. Thus, we consider the use of MCMC methods, for example, Gibbs sampling and Metropolis–Hastings algorithm [4].

For simplicity, we assume that the parameters are independent a priori and they have prior distribution according to the parametric space of each one. It means that

$$\pi(\lambda, \boldsymbol{\gamma}) = f_\Gamma(\lambda|a_\lambda, b_\lambda) \times \prod_{p=0}^{r} f_{\mathcal{N}}(\gamma_p|0, \sigma^2_{\gamma_p}), \tag{29.11}$$

where $f_\Gamma(y|a,b) \propto y^{a-1}e^{-by}$, $y > 0$ is the density function of a gamma distribution with shape parameter $a > 0$, scale parameter $b > 0$, mean a/b and variance a/b^2; $f_{\mathcal{N}}(\cdot|0,\sigma^2)$ is the density function of a normal distribution with mean 0 and variance σ^2. The hyperparameters of the prior distribution (29.11) are assumed to be known.

Based on the log-likelihood function (29.9) and the prior specification (29.11), the joint posterior distribution for λ and $\boldsymbol{\gamma}$ is proportional to

$$\pi(\lambda, \boldsymbol{\gamma}|\cdot) \propto \exp\{\ell(\boldsymbol{\theta})\} \times f_\Gamma(\lambda|a_\lambda, b_\lambda) \times \prod_{p=0}^{r} f_{\mathcal{N}}(\gamma_p|0, \sigma^2_{\gamma_p}). \tag{29.12}$$

In order to generate our samples of λ and γ_p, we implemented a Metropolis–Hastings algorithm. We start with $\boldsymbol{\theta}^{(0)} = (\lambda^{(0)}, \gamma_0^{(0)}, \ldots, \gamma_r^{(0)})$ and generate a candidate $\boldsymbol{\theta}^*$ from a jumping distribution $q(\boldsymbol{\theta}^*, \nu)$, mapping $\boldsymbol{\theta}^*$ to ν such that $\nu \leftarrow \boldsymbol{\theta}^* + \sigma z$, with $z \sim \mathcal{N}(0,1)$ and σ a scalar. We generate u from a uniform distribution $U(0,1)$ and then make the following comparison: if $u \leq \min\{1, \pi(\boldsymbol{\theta}^*|\cdot)/\pi(\boldsymbol{\theta}^{(0)}|\cdot)\}$, then update $\boldsymbol{\theta}^{(1)}$ by $\boldsymbol{\theta}^*$. Otherwise, stay with $\boldsymbol{\theta}^{(0)}$, i.e., $\boldsymbol{\theta}^{(1)} = \boldsymbol{\theta}^{(0)}$. We repeat the algorithm steps until a stationary sample can be obtained.

Table 29.1 Summary of the simulated means and respective standard errors of the estimators for λ, γ_0, and γ_1

		λ			γ_0			γ_1		
n	$\hat{\mu}_E$	Mean	SD	RSE	Mean	SD	RSE	Mean	SD	RSE
$\lambda = 0.5$										
30	3.91	0.438	0.166	0.281	0.988	0.096	0.014	−1.005	0.144	0.042
50	3.94	0.438	0.165	0.261	0.994	0.078	0.009	−1.007	0.128	0.028
100	3.94	0.450	0.161	0.256	0.996	0.060	0.004	−0.994	0.104	0.016
$\lambda = 2.0$										
30	4.72	1.985	0.182	0.022	0.987	0.098	0.014	−0.995	0.147	0.042
50	4.71	1.990	0.181	0.020	0.989	0.082	0.010	−1.001	0.133	0.031
100	4.67	1.977	0.174	0.019	0.997	0.063	0.006	−1.004	0.107	0.017

$\hat{\mu}_E$ is the observed mean number of events per unit in all 1,000 replications

29.4 Simulation Study

This section presents the results of a simulation study performed in order to verify the frequentist properties of the estimation procedure based on resamples. To examine the frequentist properties, we focus on the relative square error (RSE), standard error (SD) and on the coverage probability of the 95 % credible intervals (CI) for different samples sizes, $n = 30, 50$ and 100. The rate function of the recurrence process is of the form (29.4), and covariates are linked to the occurrence time of each cause in terms of a binary covariate, i.e., $\beta_i = \exp(\gamma_0 + \gamma_1 x_i)$, $i = 1, \ldots, n$. The vector parameter to be estimated is given by $\boldsymbol{\theta} = (\lambda, \gamma_0, \gamma_1)$. The simulated parameter combinations are $\gamma_0 = 1$, $\gamma_1 = -1$, and $\lambda \in \{0.5, 2\}$. The values of the fixed covariate x_i are generated from a Bernoulli distribution with parameter 0.5. The follow-up time τ_i is generated for each unit from a uniform distribution $U(0, a)$, where a is chosen through a trial-and-error method, such that no unit is allowed to experience more than 20 events. This results in a mean number of events experienced by a unit of approximately four for $\lambda = 0.5$ and five for $\lambda = 2$. For each simulated data set, we obtain the posterior summaries of the parameters. The gamma distribution $\Gamma(0.4, 0.2)$, with mean 2 and variance 10, is considered as the prior distribution of λ. A normal distribution with mean 0 and variance 100 is considered for γ_0 and γ_1. We simulated a chain of 10,500 iterations for each parameter, disregarding the first 500 iterations to eliminate the effect of the initial values. The remaining ones were selected using thinning by 10 to avoid a series correlation, obtaining a sample of size 1,000.

For each set-up, we conducted 1,000 replicates. For these replicates, we averaged the estimates of parameters and calculated the RSE and SD. Also, we have calculated the coverage of the lower CI bound (L), the coverage of the upper CI bound (U), and the coverage of the 95 % CI (C) for λ, γ_0, and γ_1. The results are summarized in Tables 29.1 and 29.2.

Table 29.2 Frequentist coverage of 95 % CI for λ, γ_0, and γ_1, where L, U, and C denote the coverage probabilities of the lower CI bound, upper CI bound, and 95 % CI, respectively

	λ			γ_0			γ_1		
n	L	U	C	L	U	C	L	U	C
$\lambda = 0.5$									
30	0.087	0.201	0.712	0.053	0.073	0.874	0.086	0.104	0.810
50	0.083	0.086	0.831	0.062	0.056	0.882	0.036	0.055	0.909
100	0.037	0.038	0.925	0.041	0.048	0.911	0.029	0.033	0.938
$\lambda = 2.0$									
30	0.123	0.131	0.746	0.055	0.079	0.866	0.097	0.085	0.818
50	0.052	0.090	0.858	0.031	0.041	0.928	0.037	0.039	0.924
100	0.027	0.029	0.944	0.030	0.032	0.938	0.020	0.024	0.956

The empirical RSE and SD decrease as the sample size increases, and the differences between the average estimates and the true values are almost always smaller than one empirical SD.

Also, for large samples, we can observe a balance between lower and upper credible interval bounds. The empirical coverage probabilities for the parameter λ are below 80 % for small sample sizes. However, for the sample size 100, the coverage probabilities are close to the nominal level, with a range from 91 to 96 % for all parameters.

29.5 Muscle Soreness Data

In this section, the methodology is illustrated in a real-data set.[1] We consider the study of two treatment modalities to reduce the occurrence of muscle soreness among middle-aged men beginning weight training [7]. The data set provides the gap times between successive soreness episodes of $n = 400$ participants who joined a health club for the specific purpose of weight training. Subjects were randomized into one of the two programs designed to prevent muscle soreness. The control treatment consisted of the standard written brochures and instructions used by the health club to explain proper technique. The new method includes 1 h with a personal trainer as well as the brochures. The subjects were followed and the dates on which muscle soreness limited the prescribed workout were recorded. The dates were converted into the number of days between soreness episodes. The variables are ID (1-400), AGE (years), TREAT (0=new, 1=control), TIME0 (day of the previous episode), TIME1 (day of new episode), CENSOR (1=muscle soreness episode occurred at

[1] The data can be found on http://www.umass.edu/statdata/statdata/data

Table 29.3 Posterior means and corresponding 95 % credible intervals (*in parentheses*)

	Parameter			
Model	λ	γ_0	γ_1	γ_2
EPre	2.507	0.218	0.292	0.032
	(1.778; 3.102)	(0.109; 0.346)	(0.162; 0.407)	(0.025; 0.038)
HPP		0.273	0.351	0.044
		(0.129; 0.413)	(0.278; 0.435)	(0.040; 0.048)

TIME1, 0=subject left the study or the study ended at TIME1), and EVENT (1-4 muscle soreness episode). The maximum number of episodes per subject was 4 and a total of 1296 records have been observed, with almost 28 % of censoring. Muscle soreness can be caused by excessive amount of exercise, lactic acid buildup in the muscles during strenuous workouts where the oxygen supply in the body is depleted, ultrastructural disruptions of myofilaments, amongst others, which can be regarded as latent competing causes.

Then, the EPre model fitted the data in the sampling-based approach. In addition, we compared the proposed EPre model with its particular case (the HPP model). The covariates x_{i1} : TREAT and x_{i2} : AGE are directly linked to the occurrence time of each cause as $\beta_i = \exp(\gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i2})$, $i = 1, \ldots, n$. We considered as prior distributions: $\lambda \sim \Gamma(1, 0.01)$ with $\mathrm{E}(\lambda) = 100$ and $\mathrm{Var}(\lambda) = 10,000$; $\gamma_0 \sim \mathcal{N}(0, 100)$, $\gamma_1 \sim \mathcal{N}(0, 100)$, and $\gamma_2 \sim \mathcal{N}(0, 100)$ with $\mathrm{E}(\gamma_0) = \mathrm{E}(\gamma_1) = \mathrm{E}(\gamma_2) = 0$ and $\mathrm{Var}(\gamma_0) = \mathrm{Var}(\gamma_1) = \mathrm{Var}(\gamma_2) = 100$. The hyperparameter values were chosen ensuring noninformativeness. A chain of 100, 000 iterations were considered. The first 20, 000 were ignored to avoid the influence of the first values. The remaining ones were selected using thinning by 40 to avoid a series correlation. The convergence of the chain was monitored using the method proposed by [6]. The MCMC computations were implemented in R system [17]. Table 29.3 shows the posterior means and the corresponding 95 % credible intervals (in parentheses) of the parameters.

From Table 29.3, the covariates treatment (γ_1) and age (γ_2) are associated with an increased risk of occurrence of muscle soreness, because have positive values for the mean. The parameter λ is associated to the average number of latent causes, and for the muscle soreness data this average number is equals 2.73. The estimate of γ_1 gives evidence that the new method to prevent muscle soreness is beneficial.

A comparison between the EPre model and its particular case (HPP model) is accomplished with the Akaike information criterion (AIC) and Bayesian information criterion (BIC) [14]. The AIC and BIC criterion values for EPre model are given by -2279.82 and -2259.15, respectively, whereas for HPP model are given by -2205.77 and -2190.27, respectively. The results provide positive evidence for the EPre model, showing the importance of considering a latent competing risk structure acting in the lifetime.

29.6 Remarks

In this chapter, we proposed the EPre model, which is an application of the exponential-Poisson model proposed by [11], for a recurrent event data structure, more specifically for gap time data. Conditional distributions of gap times were obtained from the hazard rate function, which is an attractive formulation for recurrent event data with direct interpretations. We discuss parameter Bayesian inference via MCMC, including a straightforward modeling comparison procedure. Simulation results suggest that the proposed method is accurate. The proposed model can also incorporate event counts and frailty, which may be further investigated by using another appropriate prior distributions.

Acknowledgements This work was partially funded by the Brazilian institutions FAPESP, CAPES, and CNPq.

References

1. Aalen, O.O., Borgan, O., Gjessing, H.K.: Survival and Event History Analysis: A Process Point of View. Springer, New York (2008)
2. Andersen, P.K., Borgan, O., Gill, R.D., Keiding, N.: Statistical Models Based on Counting Processes. Springer, New York (1993)
3. Andersen, P.K., Gill, R.D.: Cox's regression model for counting processes: a large sample study. Ann. Stat. **10**, 1100–1120 (1982)
4. Chib, S., Greenberg, E.: Understanding the Metropolis–Hastings algorithm. Am. Stat. **49**, 327–335 (1995)
5. Cook, R.J., Lawless, J.F.: The statistical analysis of recurrent events. Springer, New York (2007)
6. Gelman, A., Rubin, D.B.: Inference from iterative simulation using multiple sequences. Stat. Sci. **4**, 457–472 (1992)
7. Hosmer, D.W., Lemeshow, S., May, S.: Applied Survival Analysis: Regression Modeling of Time to Event Data. Wiley, New York (2008)
8. Huang, C.Y., Luo, X., Follmann, D.A.: A model checking method for the proportional hazards model with recurrent gap time data. Biostatistics **12**, 535–547 (2011)
9. Ibrahim, J.G., Chen, M.H., Sinha, D.: Bayesian Survival Analysis. Springer, New York (2005)
10. Kalbfleisch, J.D., Prentice, R.L.: The Statistical Analysis of Failure Time Data. Wiley, New Jersey (2002)
11. Kus, C.: A new lifetime distribution. Comput. Stat. Data Anal. **51**, 4497–4509 (2007)
12. Lawless, J.F.: Statistical Models and Methods for Lifetime Data. Wiley, New Jersey (2003)
13. Lawless, J.F., Nadeau, C.: Some simple robust methods for the analysis of recurrent events. Technometrics pp. 158–168 (1995)
14. Paulino, C.D.M., Turkman, M.A.A., Murteira, B.: Estatistica Bayesiana. Fundacao Calouste, Gulbenkian (2003)
15. Pena, E.A., Slate, E.H., Gonzalez, J.R.: Semiparametric inference for a general class of models for recurrent events. J. Stat. Plan. Inference **137**(6), 1727–1747 (2007)
16. Prentice, R.L., Williams, B.J., Peterson, A.V.: On the regression analysis of multivariate failure time data. Biometrika **68**, 373–379 (1981)
17. R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna (2013)

18. Rubinstein, R.Y., Kroese, D.P.: Simulation and the Monte Carlo Method. Wiley-Interscience, New Jersey (2008)
19. Xu, Y., Cheung, Y.B., Lam, K.F., Milligan, P.: Estimation and interpretation of incidence rate difference for recurrent events when the estimation model is misspecified. Biometrical Journal **54**, 750–765 (2012)
20. Zhao, X., Zhou, X.: Modeling gap times between recurrent events by marginal rate function. Comput. Stat. Data Anal. **56**, 370–383 (2012)
21. Zhao, X.B., Zhou, X., Wang, J.L.: Semiparametric model for recurrent event data with excess zeros and informative censoring. J. Stat. Plan. Inference **142**(1), 289–300 (2012)

Chapter 30
Conditional Predictive Inference for Beta Regression Model with Autoregressive Errors

Guillermo Ferreira, Jean Paul Navarrete, Luis M. Castro and Mário de Castro

Abstract In this chapter, we study a partially linear model with autoregressive beta distributed errors [6] from the Bayesian point of view. Our proposal also provides a useful method to determine the optimal order of the autoregressive processes through an adaptive procedure using the conditional predictive ordinate (CPO) statistic [9]. In this context, the linear predictor of the beta regression model $g(\mu_t)$ incorporates an unknown smooth function for the auxiliary time covariate t and a sequence of autoregressive errors ϵ_t, i.e.,

$$g(\mu_t) = x_t^\top \beta + f(t) + \epsilon_t,$$

for $t = 1, \ldots, T$, where x_t is a $k \times 1$ vector of nonstochastic explanatory variable values and β is a $k \times 1$ fixed parameter vector. Furthermore, these models have a convenient hierarchical representation allowing to us an easily implementation of a Markov chain Monte Carlo (MCMC) scheme. We also propose to modify the traditional conditional predictive ordinate (CPO) to obtain what we call the autoregressive CPO, which is computed for each new observation using only the data from previous time periods.

G. Ferreira (✉) · J. P. Navarrete · L. M. Castro
Department of Statistics, Universidad de Concepción, Concepción, Chile
e-mail: gferreir@udec.cl

J. P. Navarrete
e-mail: jnavarretec@udec.cl

L. M. Castro
e-mail: luiscastroc@udec.cl

M. de Castro
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo,
Av. Trabalhador São-carlense, 400 - Centro, São Carlos 13566-590, SP, Brazil
e-mail: mcastro@icmc.usp.br

A. Polpo et al. (eds.), *Interdisciplinary Bayesian Statistics,*
Springer Proceedings in Mathematics & Statistics 118, DOI 10.1007/978-3-319-12454-4_30

30.1 Introduction

The beta distribution is a flexible and useful distribution for modeling data on a bounded interval. In particular, rates and proportions have been successfully represented by a beta random variable on the unit interval $(0, 1)$. Many authors have studied regression models with the response variable following a beta distribution, including Kieschnick and McCullough [11], Ferrari and Cribari-Neto [6], and Espinheira et al. [5], among others. More recently, Figueroa et al. [8] introduced beta regression models with mixed effects under the Bayesian approach. In this approach, the linear predictor is considered as a linear function both in the fixed effects and random effects. However, in practice, the assumption of linearity in the covariates is often violated. In this context, a very good solution to the lack of linearity in the covariates is to consider partially linear models [2, 7]. These models are usually seen as semiparametric models, since they contain both a parametric linear term and a nonparametric component. In fact, in these types of models it is assumed that the linear predictor is linearly dependent on some covariates, whereas its relation to other additional variables is characterized by nonparametric functions. Weihua et al. [17] proposed a partially linear single-index beta regression model and a penalized likelihood function have been employed in order to estimate the parameters. On the other hand, for time series data, some models based on the beta distribution were proposed by Vermaak et al. [16], Rocha and Cribari-Neto [13], da Silva et al. [4], and recently by Jara et al. [10].

The main goal of this chapter is to develop a Bayesian framework to deal with time series data on the unit interval by means of a partially linear beta model with autoregressive (AR hereafter) errors in the linear predictor, i.e.,

$$g(\mu_t) = \mathbf{x}_t^\top \boldsymbol{\beta} + f(t) + \epsilon_t, \quad \text{for} \quad t = 1, \ldots, T, \tag{30.1}$$

where $g : (0, 1) \to \mathbb{R}$ is a twice differentiable strictly monotonic link function, μ_t is the linear predictor [6], $\mathbf{x}_t^\top = (x_{t,1}, \ldots, x_{t,k})$ is a $1 \times k$ vector of nonstochastic regressors (the first element of $\mathbf{x}_t$ is usually taken as 1 to allow an intercept), $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^\top$ is a vector of regression coefficients, f is an unknown smooth function and ϵ_t is an order p AR process.

The chapter is organized as follows. In Sect. 30.2 we present the partially linear beta regression model with AR errors and some of its properties. In Sect. 30.3 the Bayesian implementation of the model is presented. Model selection measures such as CPO statistic, deviance information criterion (DIC), expected Akaike information criterion (EAIC), and expected Bayesian information criterion (EBIC) are discussed in Sect. 30.4. Finally, in Sect. 30.5 a simulation study illustrating the performance of the proposed method is proposed.

30.2 Partially Linear Beta Autoregressive Model

In this section, we introduce the partially linear regression model for the linear predictor of the beta regression model. This model assumes that the relation between the expected value of the response variable and the explanatory variables is represented by (30.1) and that the error term follows an AR model.

30.2.1 The Model

The situations in which data are collected sequentially over a bounded interval, such as rates or proportions, the beta distribution provides a convenient way to model them instead of using transformations, such as the additive log-ratio and Box-Cox transformations, among others. Here, we considered a response variable taking values on the unit interval $(0, 1)$. Situations where the response is limited to a known interval (c, d) are also accommodated through the well-known transformation $y_t^\star = (y_t - c)/(d - c)$.

The conditional distribution of the variable y_t, for $t = 1, \cdots, T$, given the previous information $\mathcal{F}_{t-1}$, follows a beta distribution, parameterized in terms of its mean μ_t and a precision parameter ϕ, with density function given by

$$f(y_t|\mathcal{F}_{t-1}) = \frac{\Gamma(\phi)}{\Gamma(\mu_t\phi)\Gamma((1-\mu_t)\phi)} y_t^{\mu_t\phi-1}(1-y_t)^{(1-\mu_t)\phi-1}, \quad 0 < y_t < 1, \tag{30.2}$$

where $0 < \mu_t < 1, \phi > 0$ and $\Gamma(\cdot)$ denotes the gamma function. Here, $E(y_t|\mathcal{F}_{t-1}) = \mu_t$ and $Var(y_t|\mathcal{F}_{t-1}) = \mu_t(1-\mu_t)/(1+\phi)$.

The specification of a beta regression model requires a transformation of the mean μ_t of y_t that maps the interval $(0, 1)$ onto the real line. A convenient and popular link function is the logit link. Consequently, it is then assumed that $\log\{\mu_t/(1-\mu_t)\} = \mathbf{x}_t^\top \boldsymbol{\beta} + f(t) + \epsilon_t$. Other choices also can be used, such as the probit or complementary log-log link. It is important to stress that for the model in (30.2), all the observations have the same dispersion parameter ϕ. However, it is possible to incorporate a regression structure to model the dispersion parameter. Following [17], we assume that the varying dispersion ϕ_t can be modeled by

$$\phi_t = \phi\, m(\boldsymbol{\omega}_t^\top, \boldsymbol{\delta}), \tag{30.3}$$

where $\boldsymbol{\omega}_t^\top = (\omega_{t,1}, \ldots, \omega_{t,q})$ is a $1 \times q$ vector of covariates, $\boldsymbol{\delta}^\top = (\delta_1, \ldots, \delta_q)$ is a $1 \times q$ vector of unknown regression coefficients and $m(\cdot, \cdot)$ is a known differentiable weight function. In general, although not necessary, $\boldsymbol{\omega}_t$ is a subset of $\mathbf{x}_t$. In this chapter, we restricted our attention to the case when $m(\cdot, \cdot)$ is a loglinear function, i.e., $m(\boldsymbol{\omega}_t^\top, \boldsymbol{\delta}) = \exp(\sum_{j=1}^q \delta_j \omega_{t,j})$. In a forthcoming study, we will deal with the general case in detail.

To incorporate a dependence structure in the errors, we follow [13] considering an AR(p) process for $\{\epsilon_t\}$. In order to do that, first let $\xi_t = g(y_t) - \mathbf{x}_t^\top \boldsymbol{\beta} - f(t)$. Then, $\{\xi_t\}$ is represented as

$$\xi_t = \sum_{i=1}^{p} \varphi_i \xi_{t-i} + r_t, \tag{30.4}$$

where $\{r_t\}$ denotes a random error, satisfying $E(r_t|\mathcal{F}_{t-1}) = 0$.

Let $\boldsymbol{\varphi}^\top = (\varphi_1, \ldots, \varphi_p)$ be the $1 \times p$ AR parameter vector. Furthermore, $\{\xi_{t-i}\}$ is $\mathcal{F}_{t-1}$-measurable and $E(\xi_t|\mathcal{F}_{t-1}) \approx \epsilon_t$. Therefore, taking conditional expectations with respect to the σ-algebra $\mathcal{F}_{t-1}$ in (30.4) and replacing $\{\epsilon_t\}$ in (30.1), we obtained the following general model for the mean μ_t:

$$g(\mu_t) = \mathbf{x}_t^\top \boldsymbol{\beta} + f(t) + \sum_{i=1}^{p} \varphi_i \{g(y_{t-i}) - \mathbf{x}_{t-i}^\top \boldsymbol{\beta} - f(t-i)\}, \tag{30.5}$$

for $t = 1, \cdots, T$.

30.2.2 The Likelihood Function

From (30.2), the likelihood function for the model is given by

$$L(\boldsymbol{\theta}|\mathcal{F}_{t_0:T}) = \prod_{t=t_0}^{T} \frac{\Gamma(\phi_t)}{\Gamma(\mu_t\phi_t)\Gamma(1-\mu_t)\phi_t} y_t^{\mu_t\phi_t - 1}(1-y_t)^{(1-\mu_t)\phi_t - 1}, \tag{30.6}$$

where $\boldsymbol{\theta} = (\phi, \boldsymbol{\delta}, \boldsymbol{\varphi}, \boldsymbol{\beta})^\top$, $\mathcal{F}_{t_0:T} = \{y_{t_0}, \cdots, y_T\}$, and $t_0 = p + 1$. Notice that we consider an approximated likelihood, since we assume that the observations start in $t = p + 1$, so that the first p observations on y_t are discarded. Moreover, in the following, we will assume that the first p initial values of the process are known.

30.3 A Bayesian Approach

To represent the model given by the Eq. (30.1) in the semiparametric context, we proceed as in [14] by considering a mixed model representation of a qth-degree spline, i.e.,

$$g(\mu_t) = \mathbf{w}_t^\top \boldsymbol{\Lambda} + \mathbf{z}_t^\top \mathbf{b} + \epsilon_t,$$

where $\mathbf{w}_t = (\mathbf{x}_t^\top, \mathbf{t}^\top)^\top$ and $\boldsymbol{\Lambda} = (\boldsymbol{\beta}^\top, \boldsymbol{\alpha}^\top)^\top$, with $\boldsymbol{\alpha} = (\alpha_0, \cdots, \alpha_q)^\top$ and $\mathbf{t} = (1, t, \ldots, t^q)^\top$, includes the fixed and polynomial component of the model, $\mathbf{z}_t^\top \mathbf{b} =$

$\sum_{s=1}^{K} b_s(t-\kappa_s)_+^q$ includes the spline basis functions and $\{\epsilon_t\}$ represents the AR error. From Sect. 30.2, we can represent the expected value of the response variable as

$$\mu_t = g^{-1}\left(\eta_t + \sum_{i=1}^{p} \varphi_i \left[g(y_{t-i}) - \eta_{t-i}\right]\right), \quad t = t_0, \ldots, T,$$

where $\eta_t = \mathbf{w}_t^\top \boldsymbol{\Lambda} + \mathbf{z}_t^\top \mathbf{b}$.

Therefore, the Bayesian beta regression model is defined as follows:

$$y_t|\boldsymbol{\Lambda}, \boldsymbol{\varphi}, \phi_t, \mathcal{F}_{t-1}, \mathbf{b} \sim \text{beta}(\mu_t\phi_t, (1-\mu_t)\phi_t), \text{ for } t = t_0, \ldots, T, \tag{30.7}$$

$$\mathbf{b}|\sigma_b^2 \sim \text{N}_K(\mathbf{0}, \sigma_b^2 I_K), \tag{30.8}$$

where $\log(\phi_t) = \log(\phi) + \boldsymbol{\omega}_t^\top \boldsymbol{\delta}$. Before we begin, it is important to note that for $t = 1, \ldots, t_0 - 1$, the conditional distribution of y_t given $\boldsymbol{\Lambda}$, ϕ_t and $\mathbf{b}$ is beta $(\mu_t\phi_t, (1-\mu_t)\phi_t)$, with $\mu_t = g^{-1}(\eta_{1:t_0-1})$.

30.3.1 Prior Distributions

To complete the Bayesian specification of the beta regression model with AR errors, elicitation of prior distributions for all unknown parameters is required. Multivariate normal prior distributions are typically considered for the regression coefficients involved in the mean, i.e., we propose $\boldsymbol{\beta} \sim \text{N}_k(\boldsymbol{\beta}_0, \Omega_0)$ and $\boldsymbol{\varphi} \sim \text{N}_p(\boldsymbol{\nu}, \Upsilon)$ as prior distributions for the parametric components, and $\boldsymbol{\alpha} \sim \text{N}_{q+1}(\mathbf{0}, \Sigma_0)$ as prior distribution for the nonparametric components. In addition, multivariate normal prior distributions are considered for the regression coefficients $\boldsymbol{\delta}$ involved in the precision submodel, i.e., we propose $\boldsymbol{\delta} \sim \text{N}_q(\boldsymbol{\delta}_0, \Psi_0)$. In the Bayesian context under the model with constant dispersion ($\phi_t = \phi$), a natural choice for the prior distribution of ϕ and σ_b^2 would be an inverse gamma distribution. If a slightly informative prior is required, it can be assumed that ϕ and $\sigma_b^2 \sim IG(\varepsilon, \varepsilon)$ with a small fixed positive value for ε. However, Figueroa et al. [8] suggest a less informative prior distribution for ϕ and σ_b^2, given by

$$\phi \overset{d}{=} \sigma_b^2 \overset{d}{=} (aB)^2 \quad \text{and} \quad B \sim \text{beta}(1+\varepsilon, 1+\varepsilon),$$

where $\overset{d}{=}$ represents equality in distribution, with a positive value for a and a small positive value for ε. In the case of a varying dispersion parameter ϕ_t as in (30.3) we have specified a convenient prior for ϕ given by the log-normal distribution, say $\phi \sim LN(\mu_\phi, \sigma_\phi^2)$.

After the prior distributions for the unknown parameters have been specified, the next step is to combine the likelihood function (30.6) with the prior information in order to get the posterior distribution. This procedure is implemented by means of a MCMC scheme using WinBUGS through the R2WinBUGS package in R.

30.4 Model Comparison Tools and Diagnostics

One of the most used methods to compare several competing models fitted to a given data set is derived from the CPO statistic [9]. Let $\mathbf{y} = (y_1, y_2, \cdots, y_T)$ be the full data and $\mathbf{y}_{(i)} = (y_1, y_2, \ldots, y_{i-1}, y_{i+1}, \ldots, y_T)$ denote the data with the ith observation deleted. For the ith observation, the CPO is defined as

$$\text{CPO}_i = f(y_i|\mathbf{y}_{(i)}) = \int_{\Theta} f(y_i|\boldsymbol{\theta}, \mathbf{y}_{(i)})\pi(\boldsymbol{\theta}|\mathbf{y}_{(i)})d\boldsymbol{\theta},$$

The CPO is a cross-validation predictive approach, i.e., it is based on predictive distributions conditioned on the observed data with a single data point deleted. In this chapter, we propose to modify the CPO in order to obtain what we call the beta autoregressive CPO, that is computed for each new observation using only the data from previous time periods. Therefore, the CPO at each time period is given by

$$\text{CPO}_t = f(y_t|\mathcal{F}_{t-1}) = \int_{\Theta} f(y_t|\boldsymbol{\theta}, \mathcal{F}_{t-1})\pi(\boldsymbol{\theta}|\mathcal{F}_{t-1})d\boldsymbol{\theta}.$$

We use the conditional density $(y_t|\boldsymbol{\theta}, \mathcal{F}_{t-1})$ from (30.7), for $t = t_0, \cdots, T$, whereas for $t = 1, \cdots, t_0 - 1$, we use the conditional density in (30.2) with $\mu_t = g^{-1}(\eta_{1:t_0-1})$.

Since for the proposed model a closed form of the CPO_t is not available, a Monte Carlo estimate of the CPO_t is obtained by using the output of the Gibbs sampler. Let $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_Q$ be a sample of size Q drawn from $\pi(\boldsymbol{\theta}|\mathbf{y})$ after the burn-in period. A Monte Carlo approximation of CPO_t is given by the following expression:

$$\widehat{CPO}_t = \left(\frac{1}{Q} \sum_{q=1}^{Q} \frac{f(\mathbf{y}_{t-1}|\boldsymbol{\theta}_q)}{f(\mathbf{y}_T|\boldsymbol{\theta}_q)} \right)^{-1} \left(\frac{1}{Q} \sum_{q=1}^{Q} \frac{f(\mathbf{y}_t|\boldsymbol{\theta}_q)}{f(\mathbf{y}_T|\boldsymbol{\theta}_q)} \right),$$

where $\mathbf{y}_\ell = (y_1, y_2, \ldots, y_\ell)$ is the data vector at time point ℓ. In addition, a summary statistic of the CPO_t's is the log-pseudo marginal likelihood (LPML), defined by $\text{LPML} = \sum_{t=1}^{T} \log(\widehat{\text{CPO}}_t)$. Models with greater LMPL values indicate a better fit. Other Bayesian measure of goodness-of-fit and complexity for model selection is the DIC proposed by [15]. This criterion is based on the posterior mean of the deviance and it can be approximated by $\overline{D} = \sum_{q=1}^{Q} D(\boldsymbol{\theta}_q)/Q$, where $D(\boldsymbol{\theta}) = -2 \sum_{t=t_0}^{T} \log[f(y_t|\mathcal{F}_{t-1}, \boldsymbol{\theta})]$. The DIC criterion can be estimated using the MCMC output by $\widehat{\text{DIC}} = \overline{D} + \sharp(\boldsymbol{\theta}) = 2\overline{D} - \widehat{D}$, where $\sharp(\boldsymbol{\theta})$ is the effective number of parameters, which is defined as $E\{D(\boldsymbol{\theta})\} - D(E\{\boldsymbol{\theta}\})$, where $D(E\{\boldsymbol{\theta}\})$ is the deviance

evaluated at the posterior mean. Finally, $D(E\{\boldsymbol{\theta}\})$ can be estimated as

$$\widehat{D} = D\left(\frac{1}{Q}\sum_{q=1}^{Q}\boldsymbol{\beta}_q, \frac{1}{Q}\sum_{q=1}^{Q}\boldsymbol{\delta}_q, \frac{1}{Q}\sum_{q=1}^{Q}\phi_q, \frac{1}{Q}\sum_{q=1}^{Q}\boldsymbol{\alpha}_q, \frac{1}{Q}\sum_{q=1}^{Q}\boldsymbol{\varphi}_q\right).$$

Given the comparison of two alternative models, the model that better fits a data set is the model with the smallest value of the DIC. Note that it is important to integrate out all latent variables in the deviance calculation as this yields a more appropriate penalty term $\sharp(\boldsymbol{\theta})$ [12]. It is important to stress that for all these criteria, the evaluation of the likelihood function $L(\boldsymbol{\theta}|\mathbf{y})$ is a key aspect. However, for our beta autoregressive partial linear model, this function can be easily computed from the result given in Sect 30.2.2.

Finally, the EAIC and the EBIC can be estimated by means of $\widehat{EAIC} = \overline{D} + 2\sharp(\boldsymbol{\theta})$ and $\widehat{EBIC} = \overline{D} + \sharp(\boldsymbol{\theta})\log(T)$ [1].

30.5 Simulation Studies

30.5.1 Frequentist Properties

In this section, we study through some simulation experiments, the behavior of the Bayesian estimates, based on the frequentist bias squared error (MSE) and the frequentist mean (Mean). We performed the simulation with partially linear beta autoregressive model for different scenarios. For each scenario we consider different values for the parameters and different autoregressive orders.

We assume the following predictor linear structure

$$g(\mu_t) = \mathbf{x}_t^\top\boldsymbol{\beta} + f(t) + \sum_{i=1}^{p}\varphi_i[g(y_{t-i}) - \mathbf{x}_{t-i}^\top\boldsymbol{\beta} - f(t-i)],$$

where $f(t) = -0.15t + 0.5\sin(t+t^2+t^3) - 0.15\log(t+t^2) - 0.15\cos(t+t^2+t^3)$, $\boldsymbol{\beta} = (\beta_1, \beta_2)$, and $t = 1, 2, ..., 200$.

The covariates x_{t1} and x_{t2} are simulate from a random sample Uniforme as $x_{t1} \sim U(0, 0.5)$ and $x_{t2} \sim U(0.1, 0.3)$. The parameter vector $\boldsymbol{\varphi}$ used in this study are given in Table 30.1. Once 100 Monte Carlo generated data sets, and once the data are simulated we fit a beta regression model with autoregressive errors. The following independent priors are considered to perform the Gibbs sampler. (a) *parametric component*, $\beta_k \sim N(0, 1.0^{-6})$, for $k = \overline{1,2}$, $\phi \sim \mathrm{U}^2$, with $\mathrm{U} \sim U(1, 20)$ and $\boldsymbol{\varphi} \sim N_p(\boldsymbol{\nu}, \Upsilon)$, where $\boldsymbol{\nu} = (p+1)^{-1}\boldsymbol{\ell}$ ($\boldsymbol{\ell}$ being the unit vector of dimension $p+1$), $\Upsilon = 100I_p$, with I_r denotes the $r \times r$ identity matrix for $p = 1, 2, 3.$; (b) *non-parametric component*, $\mathbf{b} \sim N_{15}(0, \sigma_b^2 I_{15})$, $\sigma_b \sim IGamma(0.001, 0.001)$ and $\boldsymbol{\alpha}_{q+1} \sim N(0, 10^6)$, for $q = 0, 1, 2, 3$. For each generated data set we simulate one chain of size 10,000 for each parameter, disregarding the first 2000 iterations to eliminate the effect of the initial values and to avoid correlation problems, we consider a

Table 30.1 The parameter settings employed in the MCMC experiments, AR(3) = $\{\beta_1 = 0.5, \beta_2 = 0.5, \phi = 50, \varphi_0 = 0.1, \varphi_1 = 0.25, \varphi_2 = 0.35, \varphi_3 = 0.15\}$, AR(2) = $\{\beta_1 = 0.5, \beta_2 = 1.5, \phi = 20, \varphi_0 = 0.2, \varphi_1 = 0.3, \varphi_2 = 0.25\}$, AR(1) = $\{\beta_1 = 1, \beta_2 = 0.45, \phi = 10, \varphi_0 = 0.25, \varphi_1 = 0.15\}$

Model	Parameter estimates							
		$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\phi}$	$\widehat{\varphi}_0$	$\widehat{\varphi}_1$	$\widehat{\varphi}_2$	$\widehat{\varphi}_3$
AR(3)	MC mean	0.502	0.332	46.233	0.263	0.220	0.303	0.179
	MC sd	0.155	0.339	4.650	0.085	0.056	0.055	0.055
	RelBias	0.004	−0.335	−0.075	1.632	−0.118	−0.135	0.197
	MSE	4.6e-06	2.8e-02	1.4e+01	2.7e-02	8.8e-04	2.2e-03	8.7e-04
AR(2)	MC mean	0.498	1.381	19.738	0.318	0.291	0.338	
	MC sd	0.212	0.580	1.956	0.092	0.052	0.054	
	RelBias	−0.003	−0.080	−0.013	0.589	−0.031	0.350	
	MSE	2.9e-06	1.4e-02	6.8e-02	1.4e-02	8.9e-05	7.7e-03	
AR(1)	MC mean	0.934	0.512	9.759	0.337	0.185		
	MC sd	0.348	0.791	1.026	0.127	0.088		
	RelBias	−0.066	0.138	−0.024	0.348	0.233		
	MSE	4.3e-03	3.8e-03	5.8e-03	1.1e-03	1.0e-03		

spacing of size 40, obtaining a effective sample of size 200 upon which the posterior inference is based on. We compute the MSE and the "Relative Bias" (RelBias) as follows: $\text{RelBias}(\theta) = \frac{1}{200}\sum_{i=1}^{200}\left(\widehat{\theta}^i/\theta - 1\right)$ and $\text{MSE}(\theta) = \frac{1}{200}\sum_{i=1}^{200}\left(\widehat{\theta}^i - \theta\right)^2$, where $\boldsymbol{\theta} = (\beta_1, \beta_2, \phi, \boldsymbol{\varphi})$ and $\widehat{\theta}^i$ is the posterior estimates of $\boldsymbol{\theta}$ for the ith sample. From Table 30.1, it is observed that, the model indicates that it is efficient in the Bayesian estimation of the parameters. Observe that, ϕ plays the role of a precision parameter, then the larger value of ϕ, the smaller variance of the response variable y for a fixed μ. Therefore, in a forthcoming work we should study the behavior of the estimates $\widehat{\boldsymbol{\theta}}$ for different values either of ϕ as sample size T in order to check which are the values that minimize RealBias.

30.5.2 Model Selection

In this section, we study the efficiency of the CPO in the model selection for autoregressive processes for up to order $p = 5$. We assume the following autoregressive partial linear model of order $p = 2$ for the linear predictor μ_t

$$g(\mu_t) = \mathbf{x}_t^\top \boldsymbol{\beta} + f(t) + \varphi_1[g(y_{t-1}) - \mathbf{x}_{t-1}^\top \boldsymbol{\beta} - f(t-1)] + \varphi_2[g(y_{t-2}) - \mathbf{x}_{t-2}^\top \boldsymbol{\beta} - f(t-2)],$$

where $f(t) = \sin(t + t^2 + t^3) + 0.15\log(t + t^2 + t^3) - t + 0.2t^2 + t^3$, $\boldsymbol{\beta} = (0.5, 1.5)$ and $t = 1, 2, ..., 200$. The covariates x_{t1} and x_{t2} are simulate from a random

Table 30.2 Comparison of autoregressive models. MC indicate the arithmetic average of the respective criterion

Model	MC LPML	MC DIC	MC EAIC	MC EBIC
AR(1)	161.3318	−1002.510	−321.8012	−305.3096
AR(2)	168.2525	−1043.162	−333.1018	−313.3119
AR(3)	167.1811	−1041.854	−330.4004	−307.3122
AR(4)	166.4065	−1039.243	−327.3144	−300.9278
AR(5)	166.2056	−1039.446	−325.1673	−295.4824

sample Uniforme as $x_{t1} \sim U(0, 0.5)$ and $x_{t2} \sim U(0.1, 0.3)$.Once 100 Monte Carlo generated data sets, and once the data are simulated we fit a beta regression model with autoregressive errors.

The MCMC computations were done in a similar way to those given in the last section. In order to monitor the convergence of the Markov chains we have used some of the methods recommended by Cowles and Carlin [3].

Table 30.2 presents the arithmetic mean of the measurement used for model comparison, i.e., LPML, DIC, EAIC, and EBIC for each autoregressive model with $p = 1, 2, 3, 4, 5$. We notice that all these measures favored the AR(2) model for our simulated data showing the ability of these Bayesian selection methods to detect an obvious departure from regression beta with autoregressive errors.

Acknowledgments Guillermo Ferreira would like to thank the support from ECOS-CONICYT C10E03 and partial financial support from DIUC grant 213.014.021-1.0, established by the Universidad de Concepción. Luis M. Castro acknowledges funding support from FONDECYT (Grant 1130233) form the Chilean government and Grant 2012/19445-0 from Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP-Brazil). Mário de Castro is partially supported by CNPq, Brasil.

Appendix

The R-project code for the calculating CPO is presented below.

```
 out<- bugs(); V<-out$sims.array
Q_mu<-matrix(rep(0,n*length(V[,,1])),ncol=n,nrow=length(V[,,1]))
for(i in 1:n){
Q_mu[,(i)]<-V[,,i]
}
Q_phi<-V[,,n+p]
Y<-matrix(rep(0,n*length(V[,,1])),ncol=n,nrow=length(V[,,1]))
for(i in 1:n){
 for(j in 1:length([,,1])){
Y[j,i]<-(dbeta(simulate[i],Q_mu[j,i]*Q_phi[j],
(1-Q_mu[j,i])*Q_phi[j] )) }}
FN=matrix(0, ncol=n, nrow=length(V[,,1]))
FN[,n]<-1/Y[,n]
for(k in 2:n){
```

```
   for(i in 1:M){
FN[i,k-1]<-1/prod(Y[i,k:n])}
}
GN=matrix(0, ncol=n, nrow=length(V[,,1]))
for(k in 1:n){
   for(i in 1:M){
GN[i,k]<-1/(prod(Y[i,k:n]))}
}
CPO<-apply(FN,2,mean)/apply(GN,2,mean); LPML<-sum(log(CPO))
```

References

1. Brooks, S.P.: Discussion on the paper by Spiegelhalter, Best, Carlin, and van der Linde. J. R. Stat. Soc. B **64**, 616–618 (2002)
2. Castro, L.M., Lachos, V.H., Ferreira, G., Arellano-Valle, R.B.: Partially linear censored regression models using heavy-tailed distributions: a Bayesian approach. Stat. Methodol. **18**, 14–31 (2014)
3. Cowles, M.K., Carlin, B.P.: Markov chain Monte Carlo convergence diagnostics: a comparative review. J. Am. Stat. Assoc. **91**, 883–904 (1996)
4. da Silva, C.Q., Migon, H.S., Correia, L.T.: Dynamic Bayesian beta models. Comput. Stat. Data Anal. **55**, 2074–2089 (2011)
5. Espinheira, P.L., Ferrari, S.L.P., Cribari-Neto, F.: On beta regression residuals. J. Appl. Stat. **35**, 407–419 (2008)
6. Ferrari, S., Cribari-Neto, F.: Beta regression for modeling rates and proportions. J. Appl. Stat. **31**, 799–815 (2004)
7. Ferreira, G., Castro, L.M., Lachos, V.H., Dias, R.: Bayesian modeling of autoregressive partial linear models with scale mixture of normal errors. J. Appl. Stat. **40**, 1796–1816 (2013)
8. Figueroa, J., Arellano-Valle, R., Ferrari, S.: Mixed beta regression: a Bayesian perspective. Comput. Stat. Data Anal. **61**, 137–147 (2013)
9. Geisser, S., Eddy, W.: A predictive approach to model selection. J. Am. Stat. Assoc. **74**, 153–160 (1979)
10. Jara, A., Nieto-Barajas, L.E., Quintana, F.: A time series model for responses on the unit interval. Bayesian Anal. **8**, 723–740 (2013)
11. Kieschnick, R., McCullough, B.D.: Regression analysis of variates observed on (0,1): percentages, proportions and fractions. Stat. Model. **3**, 193–213 (2003)
12. Kim, S., Chen, M.H., Dey, D.K.: Flexible generalized t-link models for binary response data. Biometrika **95**, 93–106 (2008)
13. Rocha, V.A., Cribari-Neto, F.: Beta autoregressive moving average models. TEST **18**, 529–545 (2009)
14. Ruppert, D., Wand, M.P., Carroll, R.J.: Semiparametric Regression. Cambridge University Press, New York (2003)
15. Spiegelhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A.: Bayesian measures of model complexity and fit. J. R. Stat. Soc. B **64**, 583–639 (2002)
16. Vermaak, J., Andrieu, C., Doucet, A., Godsil, S.J.: Reversible jump Markov chain Monte Carlo strategies for Bayesian model selection in autoregressive processes. J. Time Ser. Anal. **25**, 785–809 (2004)
17. Weihua, Z., Riquan, Z., Zhensheng, H., Jingyan, F.: Partially linear single-index beta regression model and score test. J. Multivar. Anal. **103**, 116–123 (2012)