Springer Proceedings in Mathematics & Statistics

Raffaele Argiento
Daniele Durante
Sara Wade *Editors*

Bayesian Statistics and New Generations

BAYSM 2018, Warwick, UK, July 2–3
Selected Contributions

Springer

Springer Proceedings in Mathematics & Statistics

Volume 296

Springer Proceedings in Mathematics & Statistics

This book series features volumes composed of selected 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.

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

Raffaele Argiento · Daniele Durante ·
Sara Wade
Editors

Bayesian Statistics and New Generations

BAYSM 2018, Warwick, UK, July 2–3
Selected Contributions

Springer

Editors
Raffaele Argiento
Department of Statistical Sciences
Università Cattolica del Sacro Cuore
Milan, Italy

Daniele Durante
Department of Decision Sciences
Bocconi University
Milan, Italy

Sara Wade
School of Mathematics
University of Edinburgh
Edinburgh, UK

ISSN 2194-1009 ISSN 2194-1017 (electronic)
Springer Proceedings in Mathematics & Statistics
ISBN 978-3-030-30610-6 ISBN 978-3-030-30611-3 (eBook)
https://doi.org/10.1007/978-3-030-30611-3

Mathematics Subject Classification (2010): 62C10, 62F15, 62-06, 62Fxx, 62Gxx, 62Mxx, 62Pxx, 65C60, 62Hxx

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Organization

Chair

Sara Wade, University of Edinburgh, UK

Scientific and Local Organizers

Raffaele Argiento, Università Cattolica del Sacro Cuore, Italy
Martine Barons, University of Warwick, UK
Daniele Durante, Bocconi University, Italy
Dario Spanò, University of Warwick, UK

Preface

This book presents theoretical, computational, and applied contributions by young researchers in the field of Bayesian statistics, addressing new challenges and providing recent advancements. The volume is structured in three parts. The opening section on *Theory and Methods* is devoted to mathematical statistics, model building, and theoretical studies of Bayesian methodology. The second section, on *Computational Statistics*, develops and compares computational tools for inference. Finally, a section on *Applied Statistics* deals with the applications of complex methods to real-world problems and data. Examples of the problems addressed in this section include the estimation of slow and fast diffusion from magnetic resonance images, a simulation study of HIV temporal patterns, and the study of past ice sheet shapes.

All of the contributions have been peer reviewed and were among the presentations delivered at the fourth Bayesian Young Statisticians Meeting (BAYSM 2018). This conference was organized and hosted by the Department of Statistics, University of Warwick in Coventry, UK on July 2 and 3, 2018 and was patronized and supported by the International Society of Bayesian Analysis (ISBA) and the Junior Section of ISBA (j-ISBA).

BAYSM provides an opportunity for early-career researchers interested in Bayesian statistics to connect with their scientific community. The goal is to stimulate collaborations, encourage discussion, and establish networks with colleagues at a similar career stage as well as with senior researchers, thereby providing valuable support for the young researchers and also promoting research across the wide spectrum of fields, where Bayesian methods can be employed.

Contributions from young researchers highlighted the diversity in applications of Bayesian methodology, ranging from clinical trials and genomics to sports, climate change, and dark matter.

We acknowledge all participants of BAYSM 2018, whose attendance and contributions made the conference an outstanding scientific event and an enjoyable experience. We thank the speakers, both junior and senior, and in particular we appreciate the valuable work of the discussants—Deborah Ashby, Bärbel Finkenstädt, Jim Griffin, Michele Guindani, Amy Herring, Jim Smith, and Mark Steel.

We also express our sincere gratitude to the referees, who thoroughly reviewed the contributions in this volume and provided helpful comments for the young researchers. Finally, we give credit to Dario Spanò and Martine Barons for their valuable work on the organizing committee and to the sponsors for their generous support: the "de Castro" Statistics Initiative, Google, ISBA, and j-ISBA. For the organizers, hosting this meeting was an exciting and rewarding experience. We expect that BAYSM conferences will continue with the same success as the first four editions, providing inspiration for new generations of Bayesian statisticians. The next meeting will be in held in Kunming, China in 2020, as a satellite to the ISBA World Meeting in Kunming, China. Additionally, we were delighted to announce that BAYSM is now the official meeting of j-ISBA.

Milan, Italy Raffaele Argiento
Milan, Italy Daniele Durante
Edinburgh, UK Sara Wade
April 2019

Contents

Theory and Methods

Computational Statistics

Applied Statistics

About the Editors

Raffaele Argiento is an Associate Professor of Statistics at the Department of Statistical Sciences, Università Cattolica del Sacro Cuore of Milano, Italy. He is a member of the board for the Ph.D. in Modeling and Data Science at the University of Turin (Italy) and affiliated to the "de Castro" Statistics initiative hosted by the Collegio Carlo Alberto, Turin. His research focuses on Bayesian parametric and nonparametric methods from both theoretical and applied viewpoints. He is the Executive Director of the Applied Bayesian Summer School (ABS) and a member of the BAYSM board.

Daniele Durante is an Assistant Professor of Statistics at the Department of Decision Sciences, Bocconi University, Italy, and a Research Affiliate at the Bocconi Institute for Data Science and Analytics (BIDSA). His research is characterized by its use of an interdisciplinary approach at the intersection of Bayesian methods, modern applications, and statistical learning to develop flexible and computationally tractable models for handling complex data. He was the chair of the Junior Section of the International Society for Bayesian Analysis (j-ISBA) in 2018.

Sara Wade is a Lecturer in Statistics and Data Science at the School of Mathematics, University of Edinburgh, UK. Prior to this, she was a Harrison Early Career Assistant Professor of Statistics at the University of Warwick, UK, where she organized and chaired the fourth BAYSM. Her research focuses on Bayesian nonparametrics and machine learning, especially the development of flexible nonparametric priors and efficient inference for complex data.

Theory and Methods

A Polya Tree Based Model for Unmarked Individuals in an Open Wildlife Population

Alex Diana, Jim Griffin and Eleni Matechou

Abstract Many ecological sampling schemes do not allow for unique marking of individuals. Instead, only counts of individuals detected on each sampling occasion are available. In this paper, we propose as novel approach for modelling count data in an open population where individuals can arrive and depart from the site during the sampling period. A Bayesian nonparametric prior, known as Polya Tree, is used for modelling the bivariate density of arrival and departure times. Thanks to this choice, we can easily incorporate prior information on arrival and departure density while still allowing the model to flexibly adjust the posterior inference according to the observed data. Moreover, the model provides great scalability as the complexity does not depend on the population size but just on the number of sampling occasions, making it particularly suitable for data-sets with high numbers of detections. We apply the new model to count data of newts collected by the Durrell Institute of Conservation and Ecology, University of Kent.

Keywords Polya tree · Statistical ecology · Bayesian nonparametrics · Count data

1 Introduction

Monitoring wildlife populations presents particular challenges. For example, it is typically not possible to perform a census of the population of interest by encountering all of the individuals. One of the most cost and time effective ways to monitor a wildlife population is to collect counts of the population on repeated sampling

A. Diana (✉) · J. Griffin · E. Matechou
School of Mathematics Statistics and Actuarial Science, University of Kent,
Canterbury, England
e-mail: ad603@kent.ac.uk

J. Griffin
e-mail: J.E.Griffin-28@kent.ac.uk

E. Matechou
e-mail: E.Matechou@kent.ac.uk

R. Argiento et al. (eds.), *Bayesian Statistics and New Generations*,
Springer Proceedings in Mathematics & Statistics 296,
https://doi.org/10.1007/978-3-030-30611-3_1

occasions (SO). This protocol is considerably easier to perform than a capture-recapture (CR) scheme as it does not require physical capture or unique identification of the individuals in the population. The most popular model for analysing count data in a frequentistic setting is the N-mixture model introduced in [6], which allows the estimation of population size and capture probability when the population is closed, that is the same individuals are present throughout the study period. However, when the data are sparse or detection probability is low, N-mixture models are known to suffer from parameter identifiability issues and may give rise to infinite estimates for population size [2]. In a Bayesian setting, the natural way to solve issues of parameter identifiability is to assume informative prior distributions on detection probability or on population size in order to obtain sensible posterior distributions.

In this paper we work in a Bayesian framework and we relax the assumption of population closure, allowing for individuals to enter and leave the site (and thus become available or unavailable for detection) at random times, but still assuming that emigration is permanent. These random arrival and departure times are sampled from a distribution with unknown parameters. However, the absence of closure makes it more challenging to separately estimate capture probability, population size and density of arrival/departure times. Hence, it is of primary importance to assume informative prior distributions in order to obtain ecologically sensible posterior distributions.

In order to allow for the posterior distribution to correctly adjust to the data without relying on parametric assumptions, we use a Bayesian nonparametric (BNP) approach to choose the prior distribution of the bivariate distribution of arrival and departure times. In particular, we work with Polya Trees, which in the BNP framework are the main alternative to Dirichlet process mixture models for modelling continuous distributions. More information on other nonparametric priors can be found in [4].

2 The Polya Tree Prior

We model the joint density of arrival and departure times, using the Polya tree (PT) prior, defined in [5]. A PT has two parameters: the first is a sequence of nested partitions Π of the sample space Ω ($\Omega = \mathbb{R}^2$ in our case), while the second parameter, α, is a sequence of positive numbers associated with each set of each partition.

The partition at the first level, π_1, is obtained by splitting the sample space in two sets, B_0 and B_1. Then for the partition at the second level, π_2, we split each of the two sets in two additional sets B_{00}, B_{01} and B_{10}, B_{11}, respectively.

$$\pi_2 = \{B_{00}, B_{01}, B_{10}, B_{11}\}, \qquad B_{00} \cup B_{01} = B_0, \quad B_{10} \cup B_{11} = B_1.$$

The same process is repeated to generate the partitions at the remaining level. A visual representation of the scheme for $\Omega = [0, 1]$ is given in Fig. 1.

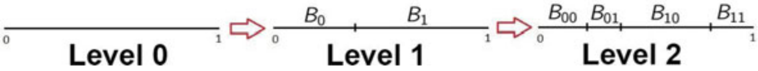


Fig. 1 Scheme of the partitions at the first three levels of the Polya tree

The PT prior is defined in terms of the (random) mass associated with each set of each partition. By defining $\varepsilon_1 \dots \varepsilon_m$ as a generic sequence of 0 s and 1 s, $B_{\varepsilon_1 \dots \varepsilon_m}$ as a generic set of the partition and $\alpha_{\varepsilon_1 \dots \varepsilon_m}$ as the associated parameter, the mass associated to $B_{\varepsilon_1 \dots \varepsilon_m}$ by the Polya Tree is

$$G(B_{\varepsilon_1 \dots \varepsilon_m}) = \prod_{i=1}^{m} Y_{\varepsilon_1 \dots \varepsilon_i} \tag{1}$$

where $Y_{\varepsilon_1 \dots \varepsilon_{i-1} 0}$ is a Beta$(\alpha_{\varepsilon_1 \dots \varepsilon_{i-1} 0}, \alpha_{\varepsilon_1 \dots \varepsilon_{i-1} 1})$ random variable and $Y_{\varepsilon_1 \dots \varepsilon_{i-1} 1} = 1 - Y_{\varepsilon_1 \dots \varepsilon_{i-1} 0}$. For example, $G(B_{01}) = Y_0(1 - Y_{00})$ where $Y_0 \sim$ Beta(α_0, α_1) and $Y_{00} \sim$ Beta$(\alpha_{00}, \alpha_{01})$.

A conjugate scheme for a PT can be constructed if we assume a PT prior for a distribution G, and we have observations $y_1, \dots, y_n \sim G$, since the posterior distribution $G \mid y_1, \dots, y_n$ is still a PT. The parameters $\alpha^\star_\varepsilon$ of the posterior distribution can be computed as $\alpha^\star_\varepsilon = \alpha_\varepsilon + n_\varepsilon$ where n_ε is the number of observations falling into set B_ε.

A common choice is to center the PT on a pre-specified distribution G_0, which means that, for every set B of the partition, $\mathbb{E}[G(B)] = G_0(B)$. In this paper we will set the $\alpha_{\varepsilon_1 \dots \varepsilon_{j-1} 0}$ and $\alpha_{\varepsilon_1 \dots \varepsilon_{j-1} 1}$ associated with the sets $B_{\varepsilon_1 \dots \varepsilon_{j-1} 0}$ and $B_{\varepsilon_1 \dots \varepsilon_{j-1} 1}$ to be proportional to the mass assigned to these sets from G_0, that is $\alpha_\varepsilon = c_\varepsilon \times G_0(B_\varepsilon)$, where c_ε is a scaling parameter tuning the overall variance around the mean distribution. Finally, we assume that G_0 has random parameter η and we let this vary by placing an additional prior on it, leading to what is known as Mixtures of Polya Trees (MPT), as defined in [3].

3 Model

The data consist of the number of individuals, D_k, detected on SO k, with $k = 1, \dots, K$. We denote by N_k the (latent) number of individuals available for detection at SO k and by p the detection probability, assumed to be constant for each individual and each SO. Clearly $D_k \sim$ Binomial(N_k, p).

We do not assume that individuals are present throughout the study period but we instead assume that their arrival and departure times are random. These times are assumed to be sampled from a Poisson process, with intensity that can be written as $\omega \times \tilde{\nu}$ where ω is the overall mass of the process and $\tilde{\nu}$ is a probability density function. The MPT is employed as a prior for $\tilde{\nu}$ and we call P_0 the prior distribution on the hyparameters of the centring distribution G_0, which we define later. As departure is always greater than arrival, $\tilde{\nu}$ is defined on $\{(x, y) \in \mathbb{R}^2 | x < y\}$.

Although the data depend only on the latent number of individuals N_k, we introduce additional latent variables to make the inference of the PT more efficient. Let $\{t_k\}_{k=1,\ldots,K}$ be the times when samples are collected and take by convention $t_0 = -\infty$ and $t_{K+1} = \infty$. Additionally, let n_{ij} be the number of individuals having arrival times between t_i and t_{i+1} and departure times between t_j and t_{j+1} (with $n_{ij} = 0$ for $i > j$). The N_k can easily be obtained from the n_{ij} as $N_k = \sum_{j=k}^{K} \sum_{i=0}^{k-1} n_{ij}$.

We make the standard choices of a Beta prior distribution for detection probability and a Gamma prior distribution for the overall intensity of the process. The hierarchical structure of the model is the following:

$$\begin{aligned}
D_k &\sim \text{Binomial}(N_k, p), & N_k &= \sum_{j=k}^{K} \sum_{i=0}^{k-1} n_{ij}, \qquad k = 1, \ldots, K, \\
n_{ij} &\sim \text{Poisson}(\omega \times \omega_{ij}), & & i = 0, \ldots, K, \quad j = 0, \ldots, K \quad i > j, \\
\omega_{ij} &= \int_{t_i}^{t_{i+1}} \int_{t_j}^{t_{j+1}} \tilde{\nu}(x, y)\, dx\, dy, & & i = 0, \ldots, K, \quad j = 0, \ldots, K, \quad i > j, \\
\omega &\sim \text{Gamma}(a_\omega, b_\omega), & p &\sim \text{Beta}(a_0, b_0), \\
\tilde{\nu} &\sim \text{PT}(\Pi, \alpha^\eta), & \eta &\sim P_0.
\end{aligned}$$

In order to center the PT on a pre-specified distribution, we use the approach explained in Sect. 2 of using a fixed partition and choosing the α according to the value η of the parameters of the centring distribution. The dependence on η is thus only in the α.

The process used to create the partition is explained in Fig. 2. The last level corresponds to the partition defined by the sampling occasions. Since we use the latent variables n_{ij} and not the exact, on a continuous scale, individual arrival and departure times, it is not meaningful to build a finer level of the partitions, as no information is available about them.

We center the PT on a bivariate distribution with independent double exponential marginal distributions, with probability density function (pdf)

$$G_0(x_1, x_2; \mu_1, \mu_2, \lambda_1, \lambda_2) = \frac{1}{2\lambda_1} \exp\left(-\frac{|x_1 - \mu_1|}{\lambda_1}\right) \frac{1}{2\lambda_2} \exp\left(-\frac{|x_2 - \mu_2|}{\lambda_2}\right),$$

with the constraint that $G_0(x_1, x_2; \mu_1, \mu_2, \lambda_1, \lambda_2) = 0$ if $x_1 > x_2$.

The sets of the partition are squares and triangles, as shown in Fig. 2. The choice of the double exponential is motivated by the fact that integrals of this distribution on squares and triangles can be computed analytically, without resorting to numerical techniques. The hyperparameters (μ_1, μ_2) are given a bivariate normal prior distribution and λ_1 and λ_2 two independent Gamma prior distributions.

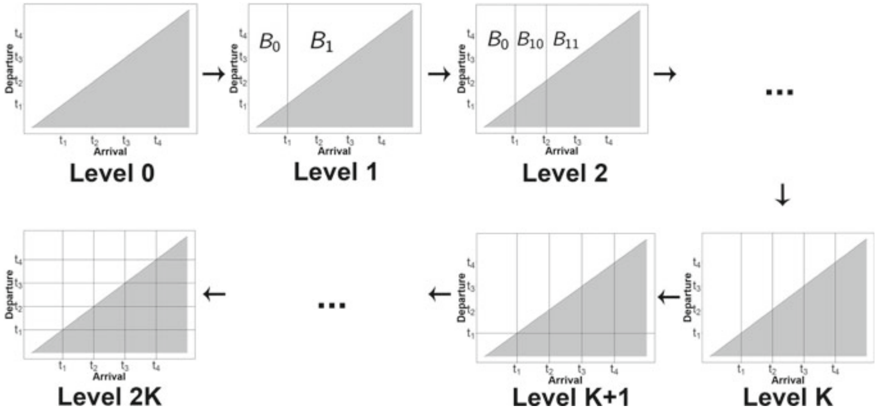


Fig. 2 Partitions of the Polya Tree. Observation occur only in the region above the diagonal. At step k, the set of individuals departing after the kth sampling occasion is split into the individuals departing between the kth and $(k+1)$th sampling occasion and the individuals departing after the $(k+1)$th sampling occasion. After having reached level K, each set is split according to the same procedure but with respect to the other dimension (arrival times)

4 Computational Notes

Posterior inference is performed using a Gibbs sampler. While this is straightforward for some parameters, such as the detection probability p and the intensity ω of the Poisson process thanks to conjugacy, for other parameters posterior inference is not straightforward given that we are working with a PT.

The conditional distribution of the parameters $(\{n_{ij}\}, \{\omega_{ij}\}, \eta)$ given p and ω can be written as

$$p(\{n_{ij}\}, \{\omega_{ij}\}, \eta|\{D_k\}, p, \omega) \propto p(\{D_k\}|\{N_k\}, p)\ p(\{n_{ij}\}, \omega, \{\omega_{ij}\}\ p(\{\omega_{ij}\}|\alpha_\eta)\ p(\eta)$$

$$\propto \prod_{k=1}^{K} \text{Binomial}(D_k|N_k, p)\ \prod_{i=1}^{K+1}\prod_{j=1}^{i} \text{Poisson}(n_{ij}|\omega \times \omega_{ij})\ p(\{\omega_{ij}\}|\Pi, \alpha_\eta)\ p(\eta)$$

where the distribution $p(\{\omega_{ij}\}|\alpha_\eta)$ is given from the PT.

When writing the posterior distribution of the parameter η, we can integrate out the parameter ω_{ij}, by employing a different parametrisation of the PT. First, we define as n_ε the number of observations in set B_ε and as $q_{\varepsilon 0}$ the probability of assigning an observation in set $B_{\varepsilon 0}$ given that we are in B_ε which, according to the structure of the PT, has a Beta$(\alpha_{\varepsilon 0}, \alpha_{\varepsilon 1})$ prior distribution. The marginal likelihood of the n_ε can be expressed as

$$p(\{n_\varepsilon\}|\{\alpha_\varepsilon\}) = \prod_{\varepsilon} \int \text{Bin}(n_{\varepsilon 0}|n_\varepsilon, q_{\varepsilon 0})\text{Beta}(q_{\varepsilon 0}|\alpha_{\varepsilon 0}, \alpha_{\varepsilon 1})dq_{\varepsilon 0}.$$

Integrating out the probabilities $q_{\varepsilon 0}$ gives as a result a beta-binomial distribution. Since the probability mass function of the beta-binomial is

$$f(k|n, \alpha, \beta) = \binom{n}{k} \frac{B(k + \alpha, n - k + \beta)}{B(\alpha, \beta)},$$

the marginal likelihood of the latent variable n_ε given the hyperparameter η is

$$p(\{n_\varepsilon\} |\eta) \propto \prod_\varepsilon \frac{B(\alpha_{\varepsilon 0} + n_{\varepsilon 0}, \alpha_{\varepsilon 1} + n_{\varepsilon 1})}{B(\alpha_{\varepsilon 0}, \alpha_{\varepsilon 1})}.$$

The posterior distribution for the latent variable n_{ij} can be written as

$$p(n_{ij}|\{D_k\}, p, \omega, \omega_{ij}) \propto \text{Poisson}(n_{ij}|\omega \times \omega_{ij}) \prod_{k=1}^{K} \text{Binomial}(D_k|N_k, p).$$

The parameter is updated with a random walk with uniform proposal over $(n_{ij} - K_{ij}, \dots, n_{ij} + K_{ij})$, where K_{ij} is chosen according to the value of the n_{ij} chosen as a starting point for the MCMC. In our case, we choose K_{ij} to be $1/5$ of the starting point of n_{ij}.

The parameters ω_{ij} correspond to the masses assigned by the distribution $\tilde{\nu}$ to the sets in the partition of the last level of the Polya tree. Hence, they can be sampled as a product of Beta distributions as in (1). The parameters of the PT are updated at each iteration conditional on the latent variable n_{ij}, using the standard update explained in Sect. 2.

5 Application

The data used in our application consist of weekly detections of great crested newts (GCN) (*Triturus cristatus*). This species generally start to migrate to ponds in late winter in order to breed. Subsequently, they leave the breeding site at the end of the summer and hibernate on land. Sampling took place in ponds located at the University of Kent with data collected between the end of February 2016 until the beginning of September of the same year, which covers a large part of the breeding period of the newts. Samples were collected on weeks 1–22, 24–27 and 29 of the season.

GCN are uniquely identifiable and hence individual capture histories of this population exist. A total of 69 individuals were captured during the study. However, in this case, the individual CR data have been collapsed to simple count data, which are obtained by recording the number of individual newts caught on each of the sampling weeks. It is believed that the population size is close to the sample size, and we choose a prior distribution for ω with mean 76 to represent our belief that around 90% of the individuals have been detected at least once. The variance has

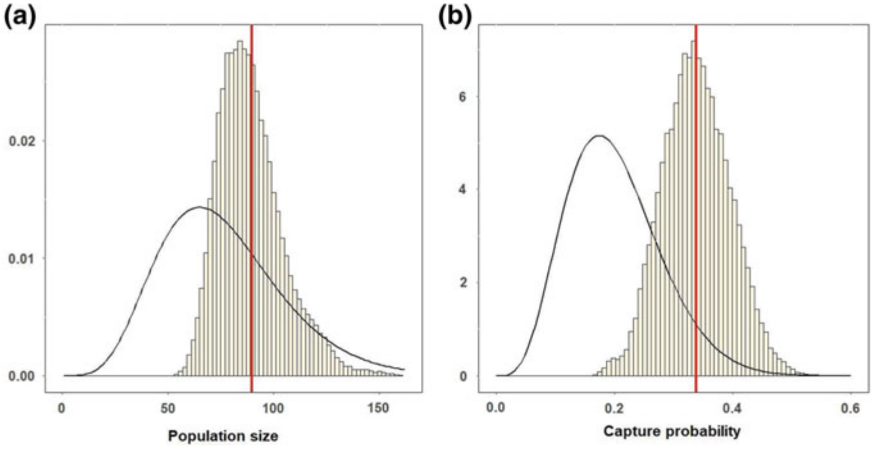


Fig. 3 Posterior distributions of population size (**a**) and detection probability (**b**), with the red vertical lines showing the posterior means and the black line showing the prior distribution in each case

been chosen in order to have a relative weakly informative prior, as the 95% prior mass includes up to 135 individuals. Finally, 95% of the prior mass for detection probability is placed on the (0.05, 0.35) interval, based on previous analyses of data on the same population.

Prior knowledge suggests that a considerable number of individuals tend to arrive at the site between the beginning of March and the end of April. Additionally, individuals depart between the end of May and the end of July. In order to translate this knowledge into our prior distributions, we choose hyperpriors for $\eta = (\mu, \lambda)$ such that 95% of the prior mass of the arrival and departure density is in the aforementioned ranges.

The posterior mean estimate of the population size is 89, while the posterior mean of the detection probability is 0.33. The two posterior distributions are shown in Fig. 3. The posterior distribution of the population size is different from the prior distribution, as more individuals are estimated to be at the site than expected by the ecologists. Moreover, in Fig. 4 we display the posterior mean of the latent number of individuals available at each sampling occasion together with the number of individuals counted. For some sampling occasions, the empirical estimated detection probability, estimated as the ratio between the estimated number of individuals available and the counted individuals, is outside the 95% posterior credible interval for detection probability. This suggests that detection probability is not constant across sampling occasions, as we have assumed in our model. According to expert knowledge, changes in detection probability might be due to differences in environmental conditions between sampling occasions, which affect behaviour of newts.

The posterior cumulative distribution functions of arrival and departure are also shown in Fig. 4. As 95% of the individuals are estimated to arrive before any

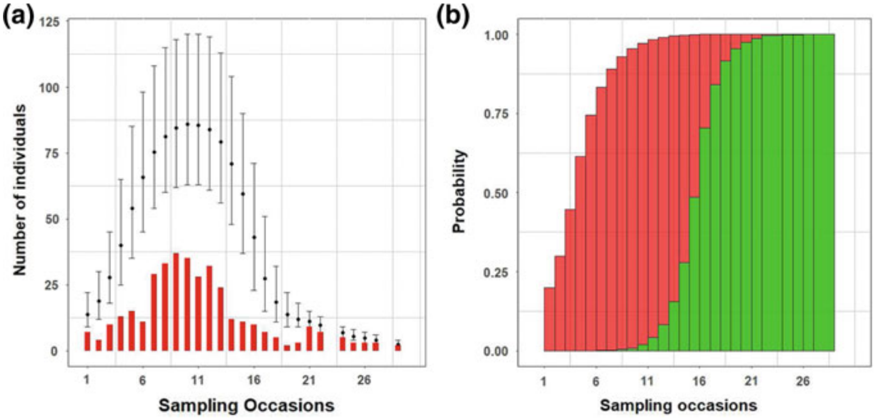


Fig. 4 **a** 95% posterior credible interval of the latent number of individuals available for detection each week, shown in black, and the number of individuals detected each week shown in red. **b** Posterior mean of the cumulative distribution function of arrival times (in red) and departure times (in green)

individual has departed, the number of individuals is estimated to be fairly constant between sampling occasions 9 and 14.

6 Conclusion

In this paper we have presented a BNP model for count data on an open wildlife population consisting of individuals entering and exiting the site at random times. By assuming a PT prior, we make no parametric assumptions on the shape of the arrival and departure distribution. Moreover, the implementation is fast as the computational complexity does not depend on the number of individuals but on the levels of the PT, which depends on the number of sampling occasions. However, given the small amount of information provided in count data, it is important to assume meaningful and informative prior distributions in order to have sensible posterior distributions. In this paper, we assume informative prior distributions for detection probability and for population size, available thanks to expert knowledge.

As we mentioned in the introduction, another common sampling protocol is CR which, as opposed to count data, provides individual information that can improve estimation. Hence, a possible extension is to model count data and CR data jointly. Another useful extension is to model data collected at different sites, by replacing the Polya tree prior with a hierarchical Polya tree prior, defined in [1].

References

1. Christensen, J., Ma, L.: A Bayesian hierarchical model for related densities using Polya trees. arXiv:1710.01702v2
2. Dennis, E.B., Morgan, B.J., Ridout, M.S.: Computational aspects of N-mixture models. Biometrics **71**(1), 237–246 (2015)
3. Hanson, T.: Inference for mixtures of finite Polya tree models. J Am Stat Assoc **101**, 1548–1565 (2006)
4. Hjort, N.L., Holmes, C., Muller, P., Walker, S.G.: Bayesian Nonparametrics, vol. 28. Cambridge University Press, Cambridge (2010)
5. Lavine, M.: Some aspects of Polya tree distributions for statistical modelling. Ann. Stat. 1222–1235 (1992)
6. Royle, J.A.: N-mixture models for estimating population size from spatially replicated counts. Biometrics **60**(1), 108–115 (2004)

Bias Estimation and Correction Using Bootstrap Simulation of the Linking Process

Shovanur Haque and Kerrie Mengersen

Abstract Record linkage involves a number of different linking methods to link records from one or more data sources. Linkage error that occurs in the linking methods due to erroneous entries or missing identifying information can lead to biased estimates. It is essential to focus on the impact of bias and techniques for the bias correction. This paper finds an expression for the bias of simple estimators of cross-products of variables across linked files and constructs a bias-corrected estimator. To derive the expressions for bias of the estimators, different scenarios of linked files are considered. It is assumed that linkage is independent of linking variable values. The situation is also considered where this independence assumption does not hold. An expression of bias is defined where the product of variable values for a true matched record pair are considered as a random and also a fixed value. For the bias correction, this paper also proposes bootstrap simulation for the estimation of match and non-match probabilities.

Keywords Record linkage · Markov chain Monte Carlo · Bias estimation · Bias correction · Bootstrap

1 Introduction

Record linkage is the process of linking records from one or more data sources that belong to the same entity. Different linking methods to find matches and link records from different data sources, may have impact on the accuracy of the results [3]. Errors can occur in the linking process due to missed-matches or false-matches. Missed-matches are records which belong to the same individual or entity but fail to be matched. False-matches are records that are erroneously matched but belong to

S. Haque (✉) · K. Mengersen
Queensland University of Technology, Brisbane, Australia
e-mail: shovanur.haque@hdr.qut.edu.au

K. Mengersen
e-mail: k.mengersen@qut.edu.au

R. Argiento et al. (eds.), *Bayesian Statistics and New Generations*,
Springer Proceedings in Mathematics & Statistics 296,
https://doi.org/10.1007/978-3-030-30611-3_2

two different individuals or entities. These two possible linkage errors can produce biased estimates. Many authors show that bias increases as linkage error increases (see for example, [8, 9, 17]). It is difficult to measure the extent of this bias with the formal measures of linkage errors such as sensitivity, specificity or match rate [2, 3, 7, 15, 16].

Linkage error that occurs due to erroneous entries or missing identifying information can lead to biased estimates. Harron et al. [11] realizes the importance of determining the potential effect of linkage error on an outcome when linked data are to be used in health research. They assessed the impact of linkage error on estimated infection rates in paediatric intensive care based on linking a national audit dataset (PICA-Net, the Paediatric Intensive Care Audit Network) and infection surveillance data (Paediatric Intensive Care Audit Network National Report, 2009–2011) using two different methods: highest-weight (HW) classification and prior-informed imputation (PII). Their study found that the bias was greater when the match rate was low or the error rate in variable values was high. In their analysis, they assumed that both the match weight and match probabilities were calculated accurately as they are based on the true match status of record pairs. However, this would not be the case in a real linkage situation. In regression analysis, the possible linkage error affects the estimation of the relationships between variables of the two files. The presence of ‘false matches’ reduces the observed level of association between variables [16]. Scheuren and Winkler [18, 19] and Lahiri and Larsen [13] also demonstrate this problem in detail. They introduce bias when estimating the slope of the regression line.

In the absence of a unique identifier, records belonging to two different people may be incorrectly linked. As a quality measure, Christen [6] describes Precision which is the proportion of links that are matches. The proportion of links that are matches help measure linkage error. Incorrect links create measurement error and consequently can bias analysis. The impacts of the bias and techniques for the bias correction have been studied in the literature (see for example, [4, 5]). Chipperfield and Chambers [4] develop a parametric bootstrap method for making inferences on binary variables using a probabilistic method to link files under the 1-1 constraint (every record from one file is linked to a distinct record in another file). Their method is valid as long as the linkage process can be replicated. They show that the analytic estimates of Precision in Lahiri and Larsen [13] are poor for 1-1 probabilistic linkage where every record from one file is linked to a distinct record in another file. Larsen and Rubin [14] estimate true match status by using the posterior probability of a match and improve the classification of matches and non-matches through clerical review. However, clerical review can be expensive and time consuming. Linking datasets and analysing linked dataset are usually performed separately with different individuals to protect data confidentiality. Thus, researchers who are involved in analysing linked datasets often lack sufficient information to correctly assess the impact of errors on results [1, 12]. This paper finds expressions for the bias of estimators of cross-products of variables across linked files and constructs bias-corrected estimators using bootstrap simulation of the linking process.

The paper is organised as follows. Section 2 briefly describes the linking process in the assessment method, *MaCSim*. A derivation of bias estimation and correction of simple estimators of cross-products of variables across linked files is provided in Sect. 3. Section 4 provides details of estimation of match and non-match probabilities using bootstrap simulation of the linking process.

2 Linking Process

In our first paper [10], we described the method *MaCSim* for assessing linkage accuracy. *MaCSim* is a Markov chain based Monte Carlo simulation method for assessing linking process. *MaCSim* utilizes two linked files to create an agreement array or agreement matrix from all linking fields across all records in the two linked files and then simulates the agreement array using a defined algorithm which maintains internal consistency in the patterns of agreement while preserving the underlying probabilistic linking structure. From the agreement array we calculate necessary parameter values and create an observed link using a defined linking process. Then we simulate the agreement array using a defined algorithm developed for generating re-sampled versions of the agreement array. In each simulation with the simulated data, records are re-linked using the same linking method that has been used before simulation. Then the simulated link is compared with the observed link and the accuracy of the individual link is calculated, which ultimately provides an estimate of the accuracy of the linking method that has been followed to link the records.

The linkage of the two files, X and Y say, is undertaken by calculating composite weights W_{ij} for each pair of records, i and j, by summing individual weights over all linking variables for that pair. An individual weight, w_{ijl} is calculated using match and non-match probabilities for any record pair (i, j) that agree or disagree on the lth linking variable value. In particular, we have:

$$W_{ij} = \sum_{l=1}^{L} \log_2 \left(Pr\{A_{ijl} \mid i, j \text{ a match}\} / Pr\{A_{ijl} \mid i, j \text{ not a match}\} \right)$$

where $\mathbf{A} = (A_{ijl})$; $i = 1, \ldots, R_X, j = 1, \ldots, R_Y, l = 1, \ldots, L$, is a three-dimensional array denoting the agreement pattern of all linking variables across all records in the two files. The agreement array $\mathbf{A}$ contains three outcomes, i.e. 1 (when values agree), -1 (when values disagree) and 0 (when either or both the values are missing), of the comparison between record i of file X and record j of file Y for the lth linking variable value. According to these codes, each linking field is given a weight using probabilities m_l, u_l and g_l, where $m_l = Pr\{A_{ijl} = 1 \mid i, j \text{ a match}\}$, $u_l = Pr\{A_{ijl}{=}1 \mid i, j \text{ not a match}\}$ and $g_l = Pr\{A_{ijl}{=}0 \mid \text{either or both } i, j \text{ are missing}\}$. These probabilities are estimated from the linked dataset. For any $(i{,}j)$th record pair and any linking variable l, if the agreement value is 1 (i.e. $A_{ijl} = 1$) then the weight is calculated using $w_{ijl} = \log_2(m_l/u_l)$; if the value is -1 (i.e. $A_{ijl} = -1$), the weight is calculated using $w_{ijl} = \log_2(1 - m_l - g_l)/(1 - u_l - g_l)$ and for a missing value (i.e. $A_{ijl} = 0$), the weight formula is $w_{ijl} = \log_2(g_l/g_l) = \log_2(1)$. Once weights of all record pairs, W_{ij} are calculated, the record pairs are then sorted according to their weights in descending order. The first record pair in the ordered list is linked if it has a weight greater than the chosen threshold cut-off value. All the other record pairs that contain either of the records from the associated record pair that has been linked are removed from the list. Thus, possible duplicate links are discarded. The same procedure is then followed for the second record pair in the list and so on until no more records can be linked.

3 Bias Estimation and Correction for Simple Estimators of Cross-Products of Variables Across Linked Files

We consider two linked files, X and Y, each with $R_X = R_Y = n$ records. Further, we assume that each record in file X has a unique true match in file Y and that we are able to uniquely link each record in file X to a partner in file Y. Since there is error in the linkage process, the records may not be correctly linked (i.e., two records may be erroneously linked). Note that we will eventually move to the more realistic case of $R_X \leq R_Y$ and some of the records in file X remaining un-linked.

Let $C = (C_1, \ldots, C_n)^T$ be the values of a variable of interest for each record in file X and similarly let $D = (D_1, \ldots, D_n)^T$ be the values of a related variable for each record in file Y. We are interested in the expectation of the total product:

$$N_{CD} = E\big(C^T D_C\big) = E\left\{\sum_{i=1}^{n} C_i D_{j(i)}\right\} = \sum_{i=1}^{n} E\big\{C_i D_{j(i)}\big\} = n\mu_{CD}$$

where $j(i)$ is the index in file Y corresponding to the true match of record i from file X, $D_C = (D_{j(1)}, \ldots, D_{j(n)})^T$ and $\mu_{CD} = E\{C_i D_{j(i)}\}$ is the expected product for a true matched record pair. Note that the vector D_C has the elements from file Y re-ordered to ensure that the corresponding elements of C and D relate to their true matches. Further, note that in this scenario, we are considering the observed pairs $(C_i, D_{j(i)})$ to be independent and identically distributed replicates from some larger population.

A simple, but potentially biased estimate of N_{CD} is constructed as

$$\hat{N}_{CD} = C^T \hat{D}_C = \sum_{i=1}^{n} C_i D_{\hat{j}(i)}$$

where $\hat{D}_C = (D_{\hat{j}(1)}, \ldots, D_{\hat{j}(n)})^T$ and $\hat{j}(i)$ is the index of the record in file Y which is linked to record i from file X.

Our goal is to find an expression for the bias of our estimator in terms of the values:

$$q_{ii} = Pr\{\hat{j}(i) = j(i)\}$$

and then construct a bias-corrected estimator.

In determining the bias, we start by making two simplifying assumptions:

1. the event $\{\hat{j}(i) = j(i)\}$ is independent of the values of C_i and $D_{j(i)}$; and,
2. the values of C_i and D_j are independent for any $j \neq j(i)$.

To calculate the bias of our simple estimator, we start by noting:

$$E\{\hat{N}_{CD}\} = \sum_{i=1}^{n} E\{C_i D_{\hat{j}(i)}\}$$

and

$$E\{C_i D_{\hat{j}(i)}\} = E\{C_i D_{\hat{j}(i)} \mid \hat{j}(i) = j(i)\} q_{ii} + E\{C_i D_{\hat{j}(i)} \mid \hat{j}(i) \neq j(i)\}(1 - q_{ii})$$

$$= E\{C_i D_{j(i)}\} q_{ii} + E\{C_i D_{\hat{j}(i)} \mid \hat{j}(i) \neq j(i)\}(1 - q_{ii}) = \mu_{CD} q_{ii} + \mu_C \mu_D (1 - q_{ii})$$

where we have used our assumption that linkage is independent of the C and D variables and $\mu_C = E(C_i)$ and $\mu_D = E(D_j)$. So, the bias of $\hat{N}_{CD}$ is:

$$Bias(\hat{N}_{CD}) = E\{\hat{N}_{CD}\} - N_{CD} = \sum_{i=1}^{n} E\{C_i D_{\hat{j}(i)}\} - N_{CD}$$

$$= \sum_{i=1}^{n} [\{\mu_{CD} - \mu_C \mu_D\} q_{ii} + \mu_C \mu_D] - N_{CD}$$

$$= \left(\sum_{i=1}^{n} \mu_{CD} q_{ii} - N_{CD} \right) + \mu_C \mu_D \sum_{i=1}^{n} (1 - q_{ii}) = \{N_{CD} - n\mu_C \mu_D\}(\bar{q} - 1)$$

where $\bar{q} = \frac{1}{n}\sum_{i=1}^{n} q_{ii}$. Or, equivalently, we have $E\{\hat{N}_{CD}\} = N_{CD}\bar{q} + n\mu_C\mu_D$ $(1 - \bar{q})$.

Hence, if we define

$$\hat{N}_{CD}^{BC,1} = \frac{\hat{N}_{CD} - n\hat{\mu}_C \hat{\mu}_D (1 - \bar{q})}{\bar{q}}$$

where $\hat{\mu}_C = \frac{1}{n}\sum_{i=1}^{n} C_i$ and $\hat{\mu}_D = \frac{1}{n}\sum_{j=1}^{n} D_j$, then $E\{\hat{N}_{CD}^{BC,1}\} = N_{CD}$. In other words, $\hat{N}_{CD}^{BC,1}$ is unbiased.

3.1 When the Values of C_i and D_j Are Not Independent for Any $j \neq j(i)$

As an alternative approach, which does not rely on assumption (2), we note that

$$E\{\hat{N}_{CD}\} = N_{CD}\bar{q} + nP_{CD}(1-\bar{q})$$

where $P_{CD} = E\{C_i D_j \mid j \neq j(i)\}$. So, if we define a new estimator as:

$$\hat{N}_{CD}^{BC,2} = \sum_{i=1}^{n} w_1 C_i D_{\hat{j}(i)} + \sum_{i=1}^{n}\sum_{j\neq i} w_2 C_i D_{\hat{j}(j)} = C^T W \hat{D}_C$$

where W is an $n \times n$ matrix with w_1s on the diagonal and w_2s on the off-diagonals, then

$$E\big(\hat{N}_{CD}^{BC,2}\big) = \sum_{i=1}^{n} w_1 E\{C_i D_{\hat{j}(i)}\} + \sum_{i=1}^{n}\sum_{j\neq i} w_2 E\{C_i D_{\hat{j}(j)}\}$$

$$= nw_1\{\mu_{CD}\bar{q} + P_{CD}(1-\bar{q})\} + \sum_{i=1}^{n}\sum_{j\neq i} w_2 E\{C_i D_{\hat{j}(j)}\}.$$

Now, since $\hat{D}_C$ is simply a permutation of D, we have

$$\sum_{i=1}^{n}\sum_{j\neq i} w_2 E\{C_i D_{\hat{j}(j)}\} = \sum_{i=1}^{n}\sum_{j=1}^{n} w_2 E\{C_i D_{\hat{j}(j)}\} - \sum_{i=1}^{n} w_2 E\{C_i D_{\hat{j}(i)}\}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} w_2 E\{C_i D_j\} - \sum_{i=1}^{n} w_2 E\{C_i D_{\hat{j}(i)}\}$$

$$= \sum_{i=1}^{n}\sum_{j\neq j(i)} w_2 E\{C_i D_j\} + \sum_{i=1}^{n} w_2 E\{C_i D_{j(i)}\} - \sum_{i=1}^{n} w_2 E\{C_i D_{\hat{j}(i)}\}$$

$$= n(n-1)w_2 P_{CD} + nw_2\mu_{CD} - nw_2\{\mu_{CD}\bar{q} + P_{CD}(1-\bar{q})\}$$

Thus,

$$E\big(\hat{N}_{CD}^{BC,2}\big) = n(w_1 - w_2)\{\mu_{CD}\bar{q} + P_{CD}(1-\bar{q})\} + n(n-1)w_2 P_{CD} + nw_2\mu_{CD}$$

$$= n\{(w_1 - w_2)\bar{q} + w_2\}\mu_{CD} + n\{(w_1 - w_2)(1-\bar{q}) + (n-1)w_2\}P_{CD}.$$

So, to ensure that $\hat{N}_{CD}^{BC,2}$ is unbiased, we need to solve the following two equations:

$$(w_1 - w_2)\bar{q} + w_2 = 1 \text{ and } (w_1 - w_2)(1 - \bar{q}) + (n - 1)w_2 = 0.$$

Solving these equations results in

$$w_1 = \frac{1}{q} + \frac{(1 - \bar{q})^2}{\bar{q}(n\bar{q} - 1)} \text{ and } w_2 = \frac{\bar{q} - 1}{n\bar{q} - 1}$$

provided $\bar{q} \neq 1/n$. From this, it is readily seen that $W = Q^{-1}$, where Q is a matrix with diagonal elements $\bar{q}$ and off-diagonal elements $(1 - \bar{q})/(n - 1)$.

3.2 When Considering the Total Product $C^T D_C$ as a Fixed Value

All the above discussions are based on the idea that $C^T D_C$ was random and its components are independent of the linking process. However, if we consider $C^T D_C$ as a fixed value (i.e., condition on its value), we can define the bias of $\hat{N}_{CD}$ as follows:

$$Bias\big(\hat{N}_{CD}\big) = E\big(\hat{N}_{CD}\big) - C^T D_C = E\big(C^T \hat{D}_C\big) - C^T D_C = C^T\big\{E\big(\hat{D}_C\big) - D_C\big\}$$

Further, since there is now no randomness in $C^T D_C$, we have:

$$E\big\{\hat{N}_{CD}\big\} = \sum_{i=1}^{n} C_i E\big\{D_{\hat{j}(i)}\big\}$$

and

$$\begin{aligned} C_i E\big\{D_{\hat{j}(i)}\big\} &= C_i E\big\{D_{\hat{j}(i)} \mid \hat{j}(i) = j(i)\big\} q_{ii} + C_i E\big\{D_{\hat{j}(i)} \mid \hat{j}(i) \neq j(i)\big\}(1 - q_{ii}) \\ &= C_i D_{j(i)} q_{ii} + C_i E\big\{D_j \mid j \neq j(i)\big\}(1 - q_{ii}) \\ &= C_i D_{j(i)} q_{ii} + \sum_{i=1}^{n} \sum_{j \neq i} C_i D_{j(j)} q_{ij} \end{aligned}$$

so that $E\{\hat{N}_{CD}\} = C^T Q D_C$, where Q is now a matrix with diagonal entries q_{ii} and off-diagonal elements $q_{ij} = Pr\{\hat{j}(i) = j(j)\}$ for $j \neq i$. Thus, to create a bias corrected estimator, we could use $\hat{N}_{CD}^{BC,3} = C^T W \hat{D}$, where $W = Q^{-1}$, provided Q is non-singular.

We can see that Q is right stochastic from its definition. In the current scenario, where all records in each file are linked, Q is doubly stochastic, but this will not be true in the case of less than 100% linkage. In this situation, we still cannot say Q is invertible. However, if Q is strictly diagonally dominant, it will be invertible.

Since the structure of Q has $q_{ii} = Pr\{\hat{j}(i) = j(i)\}$, we see that the diagonal elements are simply the probabilities of correct linkage. Thus, given its right stochastic property (i.e., the rows sum to unity and the entries are all non-negative), the diagonal dominance property amounts to:

$$| q_{ii} | > \sum_{j \neq i} | q_{ij} | = 1 - q_{ii} \Rightarrow q_{ii} > 1/2 \ \forall i$$

In other words, as long as the correct linkage has greater than a 50% chance of occurring for all records, Q will be invertible, and thus bias correction in this context is guaranteed to be possible.

4 Bootstrap Simulation of the Linking Process to Estimate Match and Non-match Probabilities

The above discussions of bias correction required the use of the values q_{ii} and q_{ij}, which would generally need to be estimated. One possibility for estimating these quantities would be bootstrap simulation of the linking process. Suppose we have a bootstrap procedure to simulate a large number, B, of values $\hat{D}_b^* = \left(D_{\hat{j}_b^*(1)}, \ldots, D_{\hat{j}_b^*(n)}\right)^T$, where $b = 1, \ldots, B$. Standard bootstrap bias correction would then create the adjusted estimator:

$$\hat{N}_{CD}^{BC,4} = C^T \hat{D}_C - \widehat{Bias}_{BOOT}\left(\hat{N}_{CD}\right) = C^T \hat{D}_C - \left(\frac{1}{B} \sum_{b=1}^{B} C^T \hat{D}_b^* - C^T \hat{D}_C\right)$$

$$= 2C^T \hat{D}_C - \frac{1}{B} \sum_{b=1}^{B} C^T \hat{D}_b^*.$$

Now, in the scenario where we have assumed our linkage process will create a complete 1-1 connection between the two files, we see that $\hat{D}_b^*$ is simply a permutation of $\hat{D}_C$. So, we can write, $\hat{D}_b^* = P_b^* \hat{D}_C$ where P_b^* is the $n \times n$ permutation matrix associated with the bth replication of the linkage process. Thus, we can further rewrite our bootstrap bias correction estimator as

$$\hat{N}_{CD}^{BC,4} = 2C^T \hat{D}_C - \frac{1}{B} \sum_{b=1}^{B} C^T P_b^* \hat{D}_C = 2C^T \hat{D}_C - C^T \bar{P}^* \hat{D}_C = C^T W_{BOOT} \hat{D}_C$$

where $\bar{P}^* = \frac{1}{B} \sum_{b=1}^{B} P_b^*$ and $W_{BOOT} = 2I_n - \bar{P}^*$, with the $n \times n$ identity matrix, I_n.

5 Conclusion

This paper derives expressions for the bias of simple estimators of cross-products of variables across two linked files and formulates bias corrected estimators considering three different scenarios. The product for a true matched record pairs is considered as both random and as a fixed value. We assume that for any record pair, the agreement of linking variable values is independent among matches and non-matches for the construction of the bias estimators and also proposed bias correction estimators when the independence assumption does not hold for any non-matched pair. The bias corrected estimators require the estimation of correct match and non-match probabilities. We suggest estimating these quantities using bootstrap simulation of the linking process. This paper covers the theoretical idea of bias correction method in multiple scenarios. The practical implementation of the proposed method with data would be a future work.

References

1. Baldi, I., Ponti, A., Zanetti, R., Ciccone, G., Merletti, F., Gregori, D.: The impact of record linkage bias in the Cox model. J. Eval. Clin. Pract. **16**(1), 92–96 (2010)
2. Bohensky, M., Jolley, D., Sundararajan, V., Evans, S., Pilcher, D., Scott, I., Brand, C.: Data linkage: a powerful research tool with potential problems. BMC Health Serv. Res. **10**(1), 346–352 (2010)
3. Campbell, K.M.: Impact of record-linkage methodology on performance indicators and multivariate relationships. J. Subst. Abus. Treat. **36**(1), 110–117 (2009)
4. Chipperfield, J.O., Chambers, R.: Using the Bootstrap to Analyse Binary Data Obtained Via Probabilistic Linkage. National Institute for Applied Statistics Research Australia, Australian Bureau of Statistics (2015)
5. Chipperfield, J.O., Bishop, G.R.: Maximum likelihood estimation for contingency tables and logistic regression with incorrectly linked data. Surv. Methodol. Stat. Can. **37**(1), 13–24 (2011)
6. Christen, P.: Data Matching. Springer, New York (2012)
7. Christen, P., Goiser, K.: Assessing deduplication and data linkage quality: what to measure? In: Proceedings of the fourth Australasian Data Mining Conference, Proceedings of the 4th AusDM, pp. 37–52, Sydney (2005)
8. Fair, M., Cyr, M., Allen, A.C., Wen, S.W., Guyon, G., MacDonald, R.C.: An assessment of the validity of a computer system for probabilistic an assessment of the validity of a computer system for probabilistic record linkage of birth and infant death records in Canada. Chronic Dis. Can. **21**(1), 8–13 (2000)
9. Gomatam, S., Carter, R., Ariet, M., Mitchell, G.: An empirical comparison of record linkage procedures. Stat. Med. **21**(10), 1485–1496 (2002)
10. Haque, S., Mengersen, K., Stern, S.: Assessing the accuracy of record linkages with Markov chain based Monte Carlo simulation approach (2019). Submitted for publication, http://arxiv.org/abs/1901.04779
11. Harron, K., Wade, A., Gilbert, R., Muller-Pebody, B., Goldstein, H.: Evaluating bias due to data linkage error in electronic healthcare records. BMC Med. Res. Methodol. **14**, 36 (2014)
12. Kelman, C.W., Bass, A.J., Holman, C.D.J.: Research use of linked health data–a best practice protocol. Aust. N. Z. J. Public Health **26**(3), 251–255 (2002)
13. Lahiri, P., Larsen, M.D.: Regression analysis with linked data. J. Am. Stat. Assoc. **100**, 222–230 (2005)

14. Larsen, M.D., Rubin, D.B.: Iteratative automated record linkage using mixture models. J. Am. Stat. Assoc. **96**, 32–41 (2001)
15. Leiss, J.K.: A new method for measuring misclassification of maternal sets in maternally linked birth records: true and false linkage proportions. Matern. Child Health J. 2007 **11**(3), 293–300 (2007)
16. Neter, J., Maynes, E.S., Ramanathan, R.: The effect of mismatching on the measurement of response error. J. Am. Stat. Assoc. **60**, 1005–1027 (1965)
17. Nitsch, D., DeStavola, B.L., Morton, S.M.B., Leon, D.A.: Linkage bias in estimating the association between childhood exposures and propensity to become a mother: an example of simple sensitivity analyses. J. R. Stat. Society. Ser. A (Statistics in Society) **169**(3), 493–505 (1965)
18. Scheuren, F., Winkler, W.E.: Regression analysis of data files that are computer matched. Surv. Methodol. **19**, 39–58 (1993)
19. Scheuren, F., Winkler, W.E.: Regression analysis of data files that are computer matched- part II. Surv. Methodol. **23**, 157–165 (1997)

Non-parametric Overlapping Community Detection

Nishma Laitonjam and Neil Hurley

Abstract In this paper, we present a non-parametric overlapping community detection method based on the affiliation graph model using an Indian Buffet Process to determine the number of communities. We compare this model with a full stochastic blockmodel using the same prior, as well as two other models, a blockmodel and a community model, that employ non-parametric priors based on a Gamma Process. We ask two questions; firstly, whether community models are sufficient to model the overlapping structure of real networks, without resorting to blockmodels that entail significantly more parameters; secondly, which is the better non-parametric approach of the two analysed? Measuring performance in terms of predicting missing links, we find that all models obtain similar performance, but in general, the Indian Buffet Process prior results in simpler models, with fewer blocks or communities. We argue that, when obtaining the latent structure is the purpose of the analysis, the simpler affiliation graph model, with Indian Buffet Process is preferred.

Keywords Affiliation graph model · Generative model · Indian buffet process · Non-parametric model · Overlapping community detection

1 Introduction

Much work has been carried out in the identification of community structure in social networks. Communities are sub-sets of highly interconnected vertices. Two broad and distinct categories of community identification are studied, namely non-overlapping and overlapping community detection. In the latter case, which is the focus of this paper, vertices are allowed to belong to more than one community and this would seem the best model of real-world social communities. After all,

N. Laitonjam (✉) · N. Hurley
Insight Centre for Data Analytics, University College Dublin, Dublin, Ireland
e-mail: nishma.laitonjam@insight-centre.org

N. Hurley
e-mail: neil.hurley@insight-centre.org

R. Argiento et al. (eds.), *Bayesian Statistics and New Generations*,
Springer Proceedings in Mathematics & Statistics 296,
https://doi.org/10.1007/978-3-030-30611-3_3

people tend to belong to many social groups such as leisure and work groups. A wide range of approaches to overlapping community detection have been taken in the state-of-the-art such as local seed-expansion [10], global optimization [3] and other heuristic approaches, including an extension of the label propagation method to the overlapping case [7]. Statistical generative approaches propose a model through which the graph is generated, by initially selecting the community labels for the nodes, followed by the generation of edges, whose probability of existence depends on the community labels of their end-points. Computational inference techniques are used to determine the model parameters from an observed graph and hence to infer the community labels. In this paper, we focus on such generative methods.

We adopt the following notation. Let $G = (V, E)$ be an undirected graph, such that V is a set of nodes or vertices and $E \subseteq V \times V$ is a set of edges. An edge can be represented as the undirected tuple (v, w), where $v, w \in V$. Let, $C = \{1, \ldots, K\}$ be a set of community labels, where $K > 0$ is the total number of communities. Write $n = |V|$. Assuming some numbering such that the vertices can be written as $\{v_1, \ldots, v_n\}$, we will sometimes refer to a vertex by its index i under this numbering. An unweighted graph may be represented by its binary *adjacency matrix* $\mathrm{A} = \{a_{ij}\}$ where $a_{ij} = 1$ if $(i, j) \in E$ and $a_{ij} = 0$ otherwise. We sometimes consider integer-valued positive *weights* associated with an edge (i, j), and we write $\mathrm{X} = \{x_{ij}\}$ for the weighted graph with $x_{ij} \in \mathbb{Z}^+$. A *community-assignment* is an association of a set of communities $C_i \subseteq C$, with each $v_i \in V$. It is useful to describe the community assignment as the binary $n \times K$ matrix $\mathrm{Z} = \{z_{ik}\}$, such that $z_{ik} = 1$ if $i \in C_k$ and $z_{ik} = 0$ otherwise. When a positive real value is used to indicate the affinity of a node i to a community k, $\Phi = \{\phi_{ik}\}$ is used with $\phi_{ik} \in \mathbb{R}^+$.

2 Models

From the ground truth of real world networks, it has been found that many real world network communities exhibit *pluralistic homophily* [18], that is, that the likelihood of an edge existing between any pair of nodes is directly proportional to the number of shared communities. The focus of this paper is on models that adhere to pluralistic homophily. In the following sub-sections we gather such models into a framework which we refer to as "Overlapping Weighted Stochastic Blockmodels" and from this general framework, we define four models that are of particular interest in our analysis. Two of these models are full blockmodels that allow for any type of block structure in the network, while the other two are constrained to just community structure. Our hypothesis is that community models, containing significantly fewer parameters than full blockmodels, should be easier to learn and sufficiently accurate to model real-world overlapping community structure. As part of this analysis, we combine for the first time, the *affiliation graph model* (AGM) with an Indian Buffet Process and compare this with the other three variants from the state-of-the-art.

Overlapping Weighted Stochastic Blockmodel: Many statistical methods for overlapping community detection proposed to date are based on a common overlapping weighted stochastic blockmodel (OWSB). The general model assumes that edges in a graph G are generated independently at random, given the community assignment parameters, which we write as Ψ. The likelihood of an observed weighted graph is

$$\mathrm{P}(\mathrm{X}|\Psi) = \prod_{i=1}^{n}\prod_{j=1}^{n}\mathrm{P}(X_{ij} = x_{ij}|\Psi) \tag{1}$$

where $\mathrm{P}(X_{ij} = w)$ is the probability that the edge weight w is observed on edge (i, j). In particular, $\mathrm{P}(X_{ij} = w)$ is Poisson distributed with rate depending on the community assignments of nodes i and j, such that, the rate may be written as $\sum_{k=1}^{K}\sum_{\ell=1}^{K}\phi_{ik}\phi_{j\ell}\lambda_{k\ell}$, where $\lambda_{k\ell}$ represents the rate of edge generation between nodes in communities k and ℓ and $\phi_{ik} \in \mathbb{R}^{+}$ is a real-valued *affiliation* strength between node i and the community k. Thus, the community assignment parameters consist of the set $\Psi = (\phi_{ik}, \lambda_{k\ell}, K)$. The rates between communities are symmetric i.e. $\lambda_{k\ell} = \lambda_{\ell k}$. We refer to the above model as a *blockmodel* because between-community edges as well as within-community edges are modelled, i.e. we have $\lambda_{k\ell} > 0 \forall k, \ell$. It is referred to in [20] as an Edge Partition Model (EPM).

The weighted model may be reduced to an unweighted model, with likelihood

$$\mathrm{P}(\mathrm{A}|\Psi) = \prod_{i=1}^{n}\prod_{j=1}^{n} p_{ij}^{a_{ij}}(1 - p_{ij})^{1-a_{ij}}\ , \tag{2}$$

where $p_{ij} = \mathrm{P}(X_{ij} > 0)$. Hence, $p_{ij} = 1 - \prod_{k}\prod_{\ell}(1 - \pi_{k\ell})^{\phi_{ik}\phi_{j\ell}}$ and $\pi_{k\ell} \equiv 1 - \exp(-\lambda_{k\ell})$. Note that $\pi_{k\ell} \in [0, 1]$ and represents the probability that nodes in communities k and ℓ are joined by an edge. Thus the unweighted model may be generated, by starting with rates $\lambda_{k\ell}$, or directly from probabilities $\pi_{k\ell}$.

For a fully Bayesian treatment of this model, we require the posterior distribution, which is proportional to the product of the likelihood and prior probabilities, $\mathrm{P}(\Psi)$, i.e. $\mathrm{P}(\Psi|\mathrm{X}) \propto \mathrm{P}(\mathrm{X}|\Psi) \times \mathrm{P}(\Psi)$. Different specialisations of the model differ in their choice of prior. In particular, the value of K may be given as an input parameter, in which case a model selection method must be applied to choose among a range of K's. A more sophisticated approach is to use a *non-parametric* prior.

Affiliation Graph Model: The Affiliation Graph Model (AGM) [18] is the specialisation of the OWSB in which the between-community edge rate $\lambda_{k\ell}$ for $k \neq \ell$, is set to zero and the edges are unweighted, with ϕ_{ik} constrained as $\phi_{ik} \equiv z_{ik} \in \{0, 1\}$. In this case, the $K(K+1)/2$ rate parameters reduce to just K parameters which we write as $r_k \equiv \lambda_{kk}$, with $\pi_k = 1 - \exp(-r_k)$. To allow for the possibility of an edge between any pair of nodes, a *null* community is introduced, consisting of *all* the nodes in the network, to capture any noisy edges that are not otherwise explained by community membership. We refer to this as the ϵ community, with associated parameters r_ϵ and π_ϵ. State-of-the-art methods that follow this model, and fit it using a heuristic optimization technique include [11, 18, 19].

Non-parametric Models: Non-parametric models do not assume that the structure of the model is fixed. In the context of overlapping community detection, this corresponds to allowing the number of communities K to be inferred from the data and potentially grow to infinity. Two approaches are proposed in the state-of-the-art. The Infinite Edge Partition Model [20] uses a Hierarchical Gamma Process on $\lambda_{k\ell}$, and hence we refer to it as HGP-EPM. This model has been constrained to the AGM likelihood, in which case Gamma Process priors on r_k suffice and we refer to this constrained model as GP-AGM.

The IMRM [13] is an alternative non-parametric model, that constrains $\phi_{ik} \equiv z_{ik} \in \{0, 1\}$ and imposes an Indian Buffet process (IBP) prior on z_{ik}, allowing for an arbitrary number of communities. The IBP can be defined in terms of a community weight w_k, such that $z_{ik}|w_k \sim \text{Bernoulli}(w_k)$, with $w_k|\alpha \sim \text{Beta}(\alpha/K, 1)$. Allowing the number of communities $K \to \infty$, we obtain $Z \sim \text{IBP}(\alpha)$. The parameter α controls the number of active communities, the number of communities a node belongs to and the expected number of entries in Z. Our contribution is to constrain the IMRM to the AGM, by setting inter-community rates to zero. We refer to this model as an IBP-AGM model and its generative process is given in Algorithm 1.

The Bernoulli-Beta process in the IBP is a rich-get-richer process in which the probability that a node is assigned to any community is proportional to the number of nodes already in that community. On the other hand, the community assignment parameter in the HGP-EPM and GP-AGM depends only on node-specific parameters, so we do not expect to see a strong preferential attachment phenomenon in the community assignments in this case.

3 Related Work

In recent years, a number of generative models of networks with latent structure have been proposed in the state-of-the-art. The stochastic blockmodel (SBM) [8] posits that each node belongs to a single latent cluster and that interactions between pairs of nodes are governed by parameters, conditional on the cluster assignments of the nodes. The mixed membership stochastic blockmodel (MMSB) [1] extends this idea by assuming that each node has a distribution over the set of latent clusters and that pairs sample their cluster labels before selecting their interaction parameters. The assortative MMSB [5], specialises the MMSB by disallowing between-cluster interactions. However, MMSB models do not explicitly model for pluralistic homophily. The infinite relational model [9] is a non-parametric version of the SBM that places a Dirichlet Process prior on the community assignment of nodes. Another work on detecting latent overlapping clusters in networks is the mixtures of Dirichlet Network Distributions (MDND) model [17], that focuses on clustering links, rather than nodes. Other work focuses on the link prediction task, rather than on community identification. The work of [2] for instance, presents a model for sparse network generation, which is not explicitly defined in terms of latent clusters, but which can be related to stochastic blockmodels, in so far as the probability of a link between two

Algorithm 1 Generative process model of IBP-AGM

```
procedure GenerateNetwork
    $\alpha \sim \text{Gamma}(1, \gamma)$
    $Z \sim \text{IBP}(\alpha)$
    for $k = 1 : K$ do
        $\pi_k \sim \text{Beta}(a, b)$
    for $i < j$ such that $j \in 1 : n$ and $i \geq 1$ do
        $a_{ij} \sim \text{Bernoulli}(p_{ij})$ where $p_{ij} = 1 - (1 - \pi_\epsilon) \prod_{k=1}^{K} (1 - \pi_k)^{z_{ik} z_{jk}}$
```

nodes depends on the nodes' parameters through a link function. These parameters do not correspond to explicit cluster assignments. In summary, while there are many works on stochastic blockmodels or similar models, our focus is on models that are based on the OWSB likelihood that explicitly models pluralistic homophily.

4 Inference Techniques

We have used standard MCMC algorithms for inference in IBP-AGM. The pseudocode for the MCMC of IBP-AGM is given in Algorithm 2. For sampling the IBP parameter α, we use Gibbs sampling with a $\text{Gamma}(1, \gamma)$ prior on α. This results in a Gamma posterior given in step 14 of Algorithm 2. For sampling the community assignment Z, due to the non-conjugacy of the likelihood $P(A|Z, \pi)$ in (2) and its prior $Z \sim \text{IBP}(\alpha)$, we use Auxiliary Gibbs Sampling (AGS) to sample from a non-conjugate IBP [6]. This is similar to AGS [14] for the non-conjugate Dirichlet process mixture model. If $m_{-ik} = \sum_{j \neq i}(z_{jk}) > 0$, we sample z_{ik} from its posterior in step 4 of Algorithm 2. To sample new communities, we consider K^* auxiliary communities with $\pi^* \sim \text{Beta}(a, b)$. A node i can be assigned to any of 2^{K^*} combinations of these K^* communities. With l running over all 2^{K^*} possibilities, each Z_l^* has joint posterior consisting of the Bernoulli probabilities of assigning the node i to each of the auxiliary communities with probability $\frac{\alpha/K^*}{N+\alpha/K^*}$. For sampling the edge probability parameter π, we use Hamiltonian Monte Carlo (HMC) [4]. This method utilizes the gradient of the log posterior to reduce the correlation between successive sampled states, thus targeting higher probability states and resulting in fast convergence. It requires a discretisation of the Hamiltonian equations, for which we use the leapfrog method. Similar to the HMC for IMRM [13], π is sampled using HMC by transforming $\pi_k \in [0, 1]$ to $s_k \in (-\infty, \infty)$ with $\pi_k = \frac{1}{1+\exp(-s_k)}$, $\pi_k \sim \text{Beta}(a, b)$ and computing the log posterior, $\log P(s|A, Z)$ and its gradient, $\nabla \log P(s|A, Z)$.

Algorithm 2 MCMC algorithm for IBP-AGM

```
Input: A, $K^*$
Initialize: $a, b, \pi$, Z, $\alpha$
For each iteration, repeat :
for $i = 1 : n$ do {Sample Z using Auxiliary Gibbs Sampling}
    for $k = 1 : K$ do
        if $m_{-ik} > 0$ then
            Sample $z_{ik}$ from $\mathrm{P}(z_{ik} = z|\mathrm{A}, \mathrm{Z}_{-ik}, \pi) \propto ((1-z)\mathrm{n} - (-1)^z m_{-ik})\mathrm{P}(\mathrm{A}|\mathrm{Z}, \pi)$
        else{$m_{-ik} = 0$}
            Set one of the auxiliary community edge probability $\pi^*$ to be $\pi_k$ and $z_{ik} = 0$
    Sample the remaining $\pi^*$ from Beta$(a, b)$
    for $l = 1 : 2^{K^*}$ do
        Compute the posterior $\mathrm{P}(\mathrm{Z}_l^*|\mathrm{A}, \mathrm{Z}, \pi, \pi^*) \propto \mathrm{P}(\mathrm{Z}_l^*)\mathrm{P}(\mathrm{A}|\mathrm{Z}_l^*, \mathrm{Z}, \pi, \pi^*)$
    Sample $\mathrm{Z}_l^*$ from $2^{K^*}$ possibilities with probabilities $\mathrm{P}(\mathrm{Z}_l^*|\mathrm{A}, \mathrm{Z}, \pi, \pi^*)$
    Set $\pi = \{\pi, \pi^*\}$ and $\mathrm{Z} = \{\mathrm{Z}, \mathrm{Z}_l^*\}$
    Remove the zero columns from Z and its corresponding $\pi$
Sample $\pi$ using Hamiltonian Monte Carlo using $\nabla \log \mathrm{P}(s|\mathrm{A}, \mathrm{Z})$, where $\pi = \frac{1}{1+\exp(-s)}$
Sample $\alpha$ using Gibbs Sampling from its posterior $\mathrm{P}(\alpha|\mathrm{Z}) = \mathrm{Gamma}(1 + K, \gamma + \sum_{j=1}^{n} \frac{1}{j})$
```

5 Experiments

The main purpose of our empirical analysis is to compare the blockmodels with the community models, with the two choices of non-parametric prior (GP and IBP). The inference algorithm for IBP-AGM is described in Sect. 4. For IMRM, we use HMC for IMRM [13] and Auxiliary Gibbs sampling for IBP. For EPM models, we used Gibbs sampling described in [20], using a truncated Gamma Process and hence requires to input a maximum number of communities. On the other hand, inference with the IBP can generate any number of communities (truly non-parametric).

The following settings are used throughout the experiments, unless otherwise specified. In IBP-AGM and IMRM, in each Gibbs update for AGS, we choose the number of auxiliary communities $K^* = 3$ and for the HMC, a step size of 0.01 and a number of leapfrogs of 10 and set $\pi_\epsilon = 0.00005$. For IBP-AGM, we consider $\gamma = 100$ and $a = b = 1$. Similarly to HGP-EPM, we choose the prior on the edge probability for IMRM in such a way that interaction within a community is greater than between communities, i.e. $a = 1, b = 5 \forall k \neq \ell$ and $a = 5, b = 1 \forall k = \ell$ as given in [13] and consider $\gamma = 1$. For EPM, we have considered the same settings and initialization as given in [20]. In all the four non-parametric models, the number of communities is initialised, similarly to [20], with the condition $K = \min(100, n)$ if $n < 2{,}000$ and $K = \min(256, n)$, otherwise. We have taken the same initial settings for IBP-AGM and GP-AGM in terms of initial community assignments and edge probability parameters i.e. $z_{ik} = 1 \forall i, k$ and $\pi_k = 1 - \exp(-1/K)$ for IBP-AGM or $r_k = 1/K$ for GP-AGM. For IMRM, each edge probability is assigned to $1 - \exp(-1/K)$ initially. All the experiments are done on Intel core i5, 4 cores.

Similarly to the evaluation in [12, 13, 20], we have used missing link prediction as a measure to compare the performance between methods. The test set consists of 20% of the available pairs of nodes, whose between-edge existence or otherwise is withheld from the training data. AUC-ROC values are calculated for this test set.

While the AUC-ROC is a useful measure to detect convergence of the models, it is worth noting, as seen in the experiments later in the paper, that the different models are able to predict missing edges with similar performance. We emphasise that the purpose of our work is to detect the community structure as an output, and therefore, given similar AUC-ROC performance, we contend that simpler latent structures are preferable over more complex ones. This implies a preference towards community models, rather than blockmodels, since overlapping block structure is difficult to comprehend and a preference towards fewer, rather than more communities. Initially, we test their ability to recover the communities from networks generated from the AGM. These generated networks are simple with a known number of communities yet sufficiently complex to allow for different community structures to be found. Then, we test their performance in detecting real world communities.

Networks Generated by AGM: Networks with two communities are generated using the generative process of AGM. We choose the network size from $n = \{30, 100, 500\}$ and set $K = 2$ with $\pi_k = 0.8 \forall k$ and $\pi_\epsilon = 0.00005$. We choose community assignment Z, such that in each network, 20% of the nodes belong to the overlapping region of the two communities and 40% of the nodes belong to each community only. An edge between nodes i and j is generated with probability $p_{ij} = 1 - (1 - \pi_\epsilon) \prod_k (1 - \pi_k)^{z_{ik} z_{jk}}$. We run the different models with these networks as input, expecting to find $K = 2$ communities. After burn in of 15,000 iterations, 15,000 samples are collected, and the average result of 5 random runs is reported.

From Table 1, we can see that GP-AGM converges to a model with greater than two communities, whereas IBP-AGM converges to 2 communities. HGP-EPM and IMRM also fit the networks with more complex block structure rather than having two blocks with $\pi_{k\ell} \approx 0, \forall k \neq \ell$. As the number of edges increases in the network, the blockmodels scale less well (in time) compared with the community models. From the trace plots of the number of communities and log likelihood in Figs. 1 and 2, respectively, IBP-AGM converges in fewer iterations to GP-AGM. Moreover, as GP-AGM converges to a higher number of communities, this results in a smaller likelihood than IBP-AGM, see Fig. 2. In terms of predicting the missing links, all models seem to perform equally.

Real World Networks: We take 3 small networks ($n < 250$) i.e. Football [15] with a ground truth of 12 communities, Protein230 [20], NIPS234 [20], and 4 large networks ($n > 2{,}000$) i.e. NIPS12 [20], Yeast [13], UsPower [13], Erdos [13]. 1,500 samples are collected after burn-in of 1, 500 iterations and the average result of 5 random runs is reported. The first notable finding of Table 2 is that all models achieve similar AUC-ROC scores, with HGP-EPM and IBP-AGM marginally outperforming the others. Examining the found latent structure, we see that generally, the models based on the Gamma Process prior, HGP-EPM and GP-AGM, sample a greater number of blocks or communities than IMRM and IBP-AGM. The shrinkage

Table 1 Comparison between HGP-EPM, IMRM, GP-AGM, IBP-AGM on generated networks

#Nodes	HGP-EPM			IMRM			GP-AGM			IBP-AGM		
	AUC-ROC	Time (h)	#Block	AUC-ROC	Time (h)	#Block	AUC-ROC	Time (h)	#Comm	AUC-ROC	Time (h)	#Comm
n = 30	0.9166	0.0279	3.1508	0.9054	0.0952	2.5641	0.8608	0.0104	7.4751	0.8968	0.0605	2.0596
n = 100	0.8629	0.2907	5.5419	0.8697	0.5634	2.4125	0.8650	0.0604	21.8374	0.8732	0.1084	2.0593
n = 500	0.8643	6.9062	8.0207	0.8635	9.8146	4.5546	0.8656	1.2149	43.5966	0.8649	0.6952	2.0829

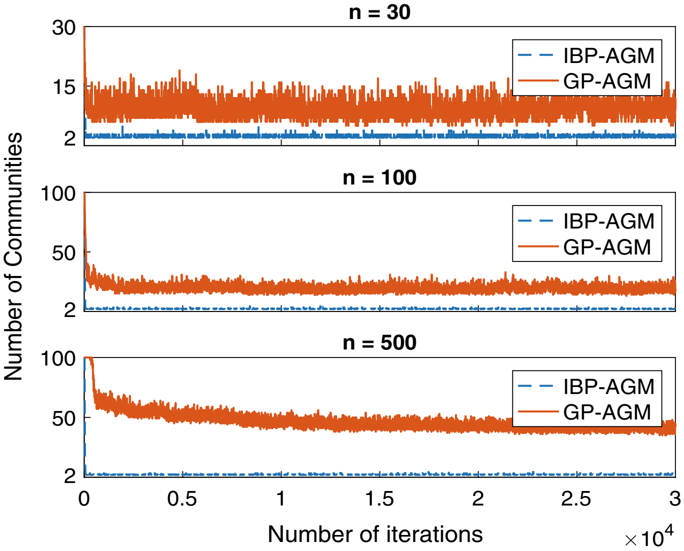


Fig. 1 Trace plots of the number of communities for a single run of the IBP-AGM and GP-AGM on generated networks

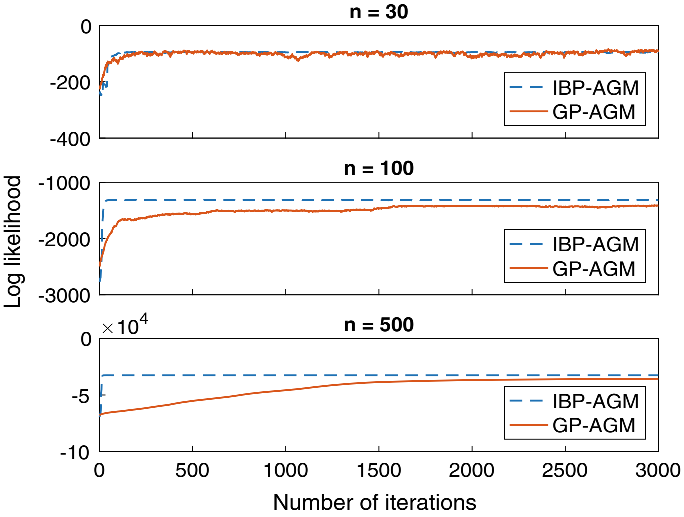


Fig. 2 Trace plots of log likelihood of first 3000 iterations for a single run of the IBP-AGM and GP-AGM on generated networks

Table 2 Comparison between HGP-EPM, IMRM, GP-AGM, IBP-AGM on real world networks

Real networks	HGP-EPM			IMRM			GP-AGM			IBP-AGM		
	AUC-ROC	Time (h)	#Block	AUC-ROC	Time (h)	#Block	AUC-ROC	Time (h)	#Comm	AUC-ROC	Time (h)	#Comm
Football	0.8411	0.0157	22.8625	0.8082	0.0335	7.9711	0.7651	0.0052	49.6084	0.8357	0.0167	11.2261
Protein230	0.9532	0.0251	31.9112	0.8958	0.0558	6.5621	0.8919	0.0119	44.4279	0.9190	0.0250	7.6931
NIPS234	0.9444	0.0282	60.0447	0.8284	0.0521	6.3777	0.8961	0.0127	63.3365	0.9522	0.0295	14.1815
NIPS12	0.9763	1.0088	242.7089	0.7635	0.9446	6.9140	0.9395	2.5234	234.2423	0.9611	1.4542	67.1507
Yeast	0.9329	1.5899	205.0551	0.8889	1.6602	8	0.8956	4.4883	238.5441	0.9178	1.3865	41.1652
UsPower	0.6175	4.2743	117.2415	0.5634	3.2992	3.9888	0.5278	22.8958	33.7568	0.6581	6.6673	35.0416
Erdos	0.9721	5.3497	91.0867	0.9601	7.5348	8.6455	0.9638	29.9171	190.1051	0.9749	11.1775	60.2116

problem of Gamma Process prior on EPM model resulting into more number of communities has been studied in [16]. This suggests that the IBP prior can recover simpler latent structure, particularly for the larger networks. For example, the GP-AGM finds 190 communities, while IBP-AGM finds 60 in the Erdos network.

While HGP-EPM generally obtains the best AUC-ROC performance, the structure it uncovers, consisting of 91 overlapping blocks in Erdos, is extremely complex. Each edge is explained, either by its membership of multiple blocks or by the fact that it straddles different blocks with non-zero between-block edge probability. It is very difficult to interpret such a model. The corresponding IBP-AGM uncovers just 60 overlapping communities, with between community edges explained entirely as noisy links, and in this case, a marginally higher AUC-ROC score.

6 Conclusion

In this paper, we have presented an overlapping community detection algorithm, which is a non-parametric version of the affiliation graph model (AGM) that uses an Indian Buffet Process (IBP) prior and exhibits pluralistic homophily. We have compared the model with another non-parametric model of AGM that uses a Gamma Process (GP) prior. In terms of convergence, IBP-AGM performs better than the GP-AGM with fewer communities in general, though there is not much difference in AUC scores for missing link prediction. Hence, empirically, IBP-AGM has better shrinkage while GP-AGM tends to overfit the data with larger number of parameters. Comparing with blockmodels, we find that the community models obtain similar performance to blockmodels and generally return simpler structures. Our future work will focus on making IBP-AGM more scalable.

References

1. Airoldi, E.M., Blei, D.M., Fienberg, S.E., Xing, E.P.: Mixed membership stochastic blockmodels. J. Mach. Learn. Res. **9**, 1981–2014 (2008)
2. Caron, F., Fox, E.B.: Sparse graphs using exchangeable random measures. J. R. Stat. Soc. Ser. B Stat. Methodol. **79**(5), 1295–1366 (2017)
3. Crampes, M., Plantié, M.: Overlapping community detection optimization and Nash equilibrium. In: Proceedings of the 5th International Conference on Web Intelligence, Mining and Semantics, p. 2. ACM (2015)
4. Duane, S., Kennedy, A.D., Pendleton, B.J., Roweth, D.: Hybrid Monte Carlo. Phys. Lett. B **195**(2), 216–222 (1987)
5. Gopalan, P.K., Gerrish, S., Freedman, M., Blei, D.M., Mimno, D.M.: Scalable inference of overlapping communities. In: Advances in Neural Information Processing Systems, pp. 2249–2257 (2012)
6. Görür, D., Jäkel, F., Rasmussen, C.E.: A choice model with infinitely many latent features. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 361–368. ACM (2006)

7. Gregory, S.: Finding overlapping communities in networks by label propagation. New J. Phys. **12**(10), 103018 (2010)
8. Holland, P.W., Laskey, K.B., Leinhardt, S.: Stochastic blockmodels: first steps. Soc. Netw. **5**(2), 109–137 (1983)
9. Kemp, C., Tenenbaum, J.B., Griffiths, T.L., Yamada, T., Ueda, N.: Learning systems of concepts with an infinite relational model. In: Proceedings of the 21st National Conference on Artificial Intelligence, pp. 381–388 (2006)
10. Lee, C., Reid, F., McDaid, A., Hurley, N.: Seeding for pervasively overlapping communities. Phys. Rev. E **83**, 066107 (2011)
11. McDaid, A., Hurley, N.: Detecting highly overlapping communities with model-based overlapping seed expansion. In: Proceedings of the International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 112–119. IEEE (2010)
12. Mørup, M., Schmidt, M.N.: Bayesian community detection. Neural Comput. **24**(9), 2434–2456 (2012)
13. Mørup, M., Schmidt, M.N., Hansen, L.K.: Infinite multiple membership relational modeling for complex networks. In: Proceedings of the International Workshop on Machine Learning for Signal Processing (MLSP), pp. 1–6. IEEE (2011)
14. Neal, R.M.: Markov chain sampling methods for Dirichlet process mixture models. J. Comput. Graph. Stat. **9**(2), 249–265 (2000)
15. Newman, M.E.: The structure and function of complex networks. SIAM Rev. **45**(2), 167–256 (2003)
16. Ohama, I., Sato, I., Kida, T., Arimura, H.: On the model shrinkage effect of gamma process edge partition models. In: Advances in Neural Information Processing Systems, pp. 397–405 (2017)
17. Williamson, S.A.: Nonparametric network models for link prediction. J. Mach. Learn. Res. **17**(1), 7102–7121 (2016)
18. Yang, J., Leskovec, J.: Community-affiliation graph model for overlapping network community detection. In: IEEE 12th International Conference on Data Mining (ICDM), pp. 1170–1175. IEEE (2012)
19. Yang, J., Leskovec, J.: Overlapping community detection at scale: a nonnegative matrix factorization approach. In: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, pp. 587–596. ACM (2013)
20. Zhou, M.: Infinite edge partition models for overlapping community detection and link prediction. In: Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS), pp. 1135–1143 (2015)

Posterior Consistency in the Binomial Model with Unknown Parameters: A Numerical Study

Laura Fee Schneider, Thomas Staudt and Axel Munk

Abstract Estimating the parameters from k independent $\text{Bin}(n, p)$ random variables, when both parameters n and p are unknown, is relevant to a variety of applications. It is particularly difficult if n is large and p is small. Over the past decades, several articles have proposed Bayesian approaches to estimate n in this setting, but asymptotic results could only be established recently in Schneider et al. (arXiv:1809.02443, 2018) [11]. There, posterior contraction for n is proven in the problematic parameter regime where $n \to \infty$ and $p \to 0$ at certain rates. In this article, we study numerically how far the theoretical upper bound on n can be relaxed in simulations without losing posterior consistency.

Keywords Bayesian estimation · Binomial distribution · Discrete parameter · Posterior contraction · Simulation study

1 Introduction

We consider estimating the parameter n of the binomial distribution from k independent observations when the success probability p is unknown. This situation is relevant in many applications, for example in estimating the population size of a species [10] or the total number of defective appliances [4]. Another recent application is quantitative nanoscopy, see [11]. There, the total number of fluorescent markers (fluorophores) attached to so-called DNA-origami is estimated from a time series of microscopic images. The number of active fluorophores counted in each

L. F. Schneider (✉) · T. Staudt · A. Munk
Institute for Mathematical Stochastics, University of Göttingen, Göttingen, Germany
e-mail: laura-fee.schneider@mathematik.uni-goettingen.de

T. Staudt
e-mail: thomas.staudt@uni-goettingen.de

A. Munk
e-mail: amunk1@gwdg.de

R. Argiento et al. (eds.), *Bayesian Statistics and New Generations*,
Springer Proceedings in Mathematics & Statistics 296,
https://doi.org/10.1007/978-3-030-30611-3_4

image is modeled as a binomial observation, where the probability p that a fluorophore is active in the respective image is very small (often below 5%).

This setting, where the success probability p is small (and n potentially large), is very challenging. The difficulties that arise can be understood by considering the following property of the binomial distribution: if n converges to infinity, p converges to zero, and the product np converges to $\lambda > 0$, then a $\mathrm{Bin}(n, p)$ random variable converges in distribution to a Poisson variable with parameter λ. Thus, the binomial distribution converges to a distribution with a single parameter. This suggests that it gets harder to derive information about the two parameters separately when n is large and p small.

In this context, it is instructive to look at the sample maximum M_k as an estimator for n, which was suggested by Fisher in 1941 [5]. Although it turns out to be impractical, see [3], the sample maximum is consistent and converges in probability for fixed parameters (n, p) exponentially fast to the true n, as $k \to \infty$. This can be seen from

$$\mathbf{P}\,(M_k = n) = 1 - (1 - p^n)^k, \tag{1}$$

which implies, by Bernoulli inequality and since $1 - x \leq e^{-x}$, that

$$1 - e^{-kp^n} \leq \mathbf{P}\,(M_k = n) \leq kp^n.$$

In an asymptotic setting where $n \to \infty$ and $p \to 0$ such that $kp^n \to 0$, the probability in (1) no longer converges to one. Thus, the sample maximum is a consistent estimator for n only as long as $kp^n \to \infty$. The condition $e^n = O(k)$ is necessary for this to hold.

Estimating n in this difficult regime becomes more manageable by including prior knowledge about p. We therefore consider random N and P, and variables $X_1, \ldots, X_k$ that are independently $\mathrm{Bin}(n, p)$ distributed given that $N = n$ and $P = p$. Various Bayesian estimators have been suggested over the last 50 years, see [1, 4, 6, 7, 10]. In all these works, a product prior for (N, P) is used, and the prior Π_P on P is chosen as beta distribution $\mathrm{Beta}(a, b)$ for some $a, b > 0$. Since this is the conjugate prior, it is a natural choice. In contrast, there is quite some discussion about the most suitable prior Π_N for N, see for example [1, 8, 9, 13]. Therefore, the asymptotic results in [11] are not restricted to a specific Π_N and only require a condition that ensures that enough weight is put on large values of n (see Eq. (4) in Sect. 2).

In [11], we also introduce a new class of Bayesian point estimators for n, which we call scale estimators. We choose $\Pi_P = \mathrm{Beta}(a, b)$ and set $\Pi_N(m) \propto m^{-\gamma}$ for a positive value γ. If $\gamma > 1$, the prior Π_N is a proper probability distribution, but it is sufficient to ensure $\gamma + a > 1$ in order to obtain a well-defined posterior distribution, as discussed in [8]. The scale estimator is then defined as the minimizer of the Bayes risk with respect to the relative quadratic loss, $l(x, y) = (x/y - 1)^2$. Following [10], it is given by

$$\hat{n} := \frac{\mathbb{E}\left[\frac{1}{N}|\mathbf{X}^k\right]}{\mathbb{E}\left[\frac{1}{N^2}|\mathbf{X}^k\right]} = \frac{\sum_{m=M_k}^{\infty} \frac{1}{m} L_{a,b}(m)\Pi_N(m)}{\sum_{m=M_k}^{\infty} \frac{1}{m^2} L_{a,b}(m)\Pi_N(m)}, \tag{2}$$

where $\mathbf{X}^k = (X_1, \ldots, X_k)$ denotes the sample, M_k is the sample maximum, and $L_{a,b}$ is the beta-binomial likelihood, see [2]. We refer to [11] for a detailed discussion and numerical study of this estimator.

The present article is structured as follows. In Sect. 2, the main theorem (proven in [11]) is presented, which shows uniform posterior contraction in the introduced Bayes setting for suitable asymptotics of n and p. The theorem states that $n^{6+\epsilon} = O(k)$ for $\epsilon > 0$ is already sufficient for consistency of the Bayes estimator. This is a significant improvement over the sample maximum, where the sample size must be much larger for consistent estimation of n. In Sect. 3, we then conduct a simulation study to better investigate the restrictions for the parameters n and p needed to ensure consistency. Our findings indicate that estimation of n is still consistent if $n^5 = O(k)$, but that it becomes inconsistent for $n^3 = O(k)$. It is hard to pin down the exact transition from consistency to inconsistency when $n^\alpha = O(k)$. However, our results suggest that, if $\alpha \approx 4$, n already grows too fast to be consistently estimated from a sample of size k. We discuss our results and provide several remarks in Sect. 4.

2 Posterior Contraction for *n*

To study posterior contraction in the binomial model we consider the Bayesian setting described in Sect. 1. For fixed parameters n and p that are independent of the number of observations k, posterior consistency follows from Doob's theorem, see, e.g., [12]. We extend this result to the class of parameters

$$\mathcal{M}_\lambda := \left\{(n_k, p_k)_k : 1/\lambda \leq n_k p_k \leq \lambda,\ n_k \leq \lambda \sqrt[6]{k/\log(k)}\right\} \tag{3}$$

for fixed $\lambda > 1$. Since we want to handle a variety of suitable prior distributions for N, we only require that Π_N is a proper probability distribution on $\mathbb{N}$ that fulfills the condition

$$\Pi_N(m) \geq \beta e^{-\alpha m^2} \tag{4}$$

for all $m \in \mathbb{N}$ and some positive constants α and β.

Theorem 1 (c.f. [11]) *Conditionally on $N = n_k$ and $P = p_k$, let $X_1, \ldots, X_k \overset{i.i.d.}{\sim}$ $\mathrm{Bin}(n_k, p_k)$. For any prior distribution $\Pi_{(N,P)} = \Pi_N \Pi_P$ on (N, P) with $\Pi_P =$ $\mathrm{Beta}(a, b)$ for $a, b > 0$, and where Π_N satisfies (4), we have uniform posterior contraction over the set $\mathcal{M}_\lambda$ of sequences $(n_k, p_k)_k$ defined in (3) for any $\lambda > 1$, i.e.,*

$$\sup_{(n_k,p_k)_k\in\mathcal{M}_\lambda} \mathbb{E}_{n_k,p_k}\big[\Pi\big(N\neq n_k\,|\,\mathbf{X}^k\big)\big]\to 0,\ \ as\ k\to\infty.$$

This result directly implies consistency of the scale estimator (2) for parameter sequences in $\mathcal{M}_\lambda$. The mild restrictions on the prior distribution allow applying the result to the estimators derived in [6, 7] as well. Furthermore, it is possible to apply the statement of Theorem 1 to estimators as in [1, 4] by restricting the improper prior to a compact support with increasing upper bound, as done in Theorem 2 in [11]. This upper bound needs to be greater than n_k and small enough such that condition (4) holds for $m = n_k$.

3 Simulation Study

The theorem presented in the previous section states that the asymptotic behavior $n_k = O\big(\sqrt[6]{k/\log(k)}\big)$ leads to posterior contraction of N for suitable priors, as long as $n_k p_k$ stays in a compact interval bounded away from zero. In this section we try to answer the question by how much the constraints on $\mathcal{M}_\lambda$ in Theorem 1 can be relaxed. We address this problem by studying the relation between posterior contraction and the order $\alpha > 0$ when $n_k = O\big(\sqrt[\alpha]{k}\big)$. More precisely, we are interested in the smallest $\alpha = \alpha^*$ such that posterior consistency,

$$\mathbb{E}_{n_k,p_k}\left[\Pi\left(N\neq n_k\,|\,\mathbf{X}^k\right)\right]\to 0,\quad \text{as}\quad k\to\infty, \tag{5}$$

remains valid. Tackling this problem analytically turns out to be extremely challenging, see the proof of Theorem 1 in [11].

In our simulations, we consider sequences $(n_k, p_k)_k$ defined by $n_k = w\sqrt[\alpha]{k}$ and $p_k = \mu/n_k$ for parameters $w, \mu > 0$. The values of w and μ should, ideally, not matter for the asymptotics and thus for the pursuit of α^*. Suitable choices of w and μ for given α are still necessary for practical reasons to ensure that the asymptotic behavior becomes visible for the values of k covered by the simulations. For any selection (α, w, μ), we calculate the posterior probability of the true parameter n_k and the MSE of different estimators for values of k up to 10^{11}. In order to achieve these extremely large observation numbers, we take care to minimize the number of operations when expressing the beta-binomial likelihood $L_{a,b}$ in our implementation. Since $L_{a,b}$ does not depend on the order of the observations but only on the frequencies of each distinct outcome x_i, the runtime depends on n_k (the number of different values that x_i can take) instead of k itself.

Figure 1a–b show the (empirical) mean posterior probability in (5) and the (empirical) mean square error (MSE) between $\hat{n}$ and n_k for different scale estimators $\hat{n}$ in several scenarios (α, w, μ). The number of samples was set to 200. It is clearly visible that the choice $\alpha = 6$ leads to posterior consistency (which is in good agreement with Theorem 1), since the posterior probability approaches 1 while the MSE converges to 0. However, the simulations indicate that this also holds true for $\alpha = 5$.

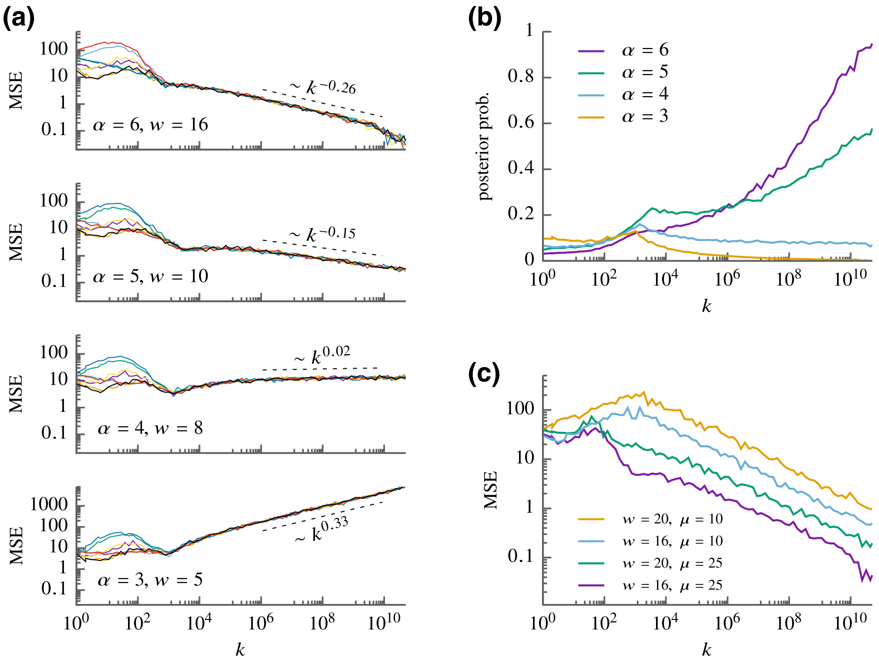


Fig. 1 Asymptotic behavior of the scale estimator and posterior contraction. **a** shows log-log plots of the MSE of several scale estimators in different asymptotic scenarios (α, w, μ). The value μ was set to 25 in each simulation, and the parameters for the scale estimators were picked as all possible combinations of $\gamma \in \{0.5, 1\}$, $a \in \{1, 5\}$, and $b \in \{1, 5\}$. **b** shows the empirical mean of the posterior probabilities $\Pi(N = n_k \mid \mathbf{X}^k)$ for the same four settings depicted in (**a**). **c** shows the MSE of the scale estimator with parameters $\gamma = a = b = 1$ for constant $\alpha = 6$ and varying values of w and μ

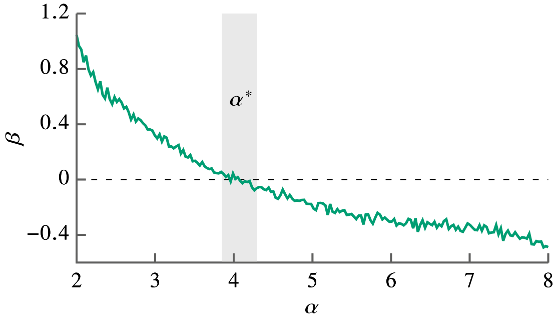
For $\alpha = 4$, it becomes questionable whether posterior contraction will eventually happen. The choice $\alpha = 3$, in contrast, leads to a clear increase of the MSE with increasing k, and posterior contraction evidently fails.

An interesting observation is the power law behavior $\sim k^{-\beta}$ of the MSE, which is revealed by linear segments in the respective log-log plots. Figure 1a shows that the slope β is independent of the chosen estimator, and Fig. 1c suggests that it might also be independent of w and μ. We can therefore consider β as a function $\beta(\alpha)$ of α alone. A numerical approximation of α^* is then given by the value of α where β changes sign, i.e.,

$$\beta(\alpha^*) = 0.$$

Since $\beta(\alpha)$ is strictly monotone, as a higher number k of observations will lead to better estimates, such an α^* is uniquely defined. Figure 2 displays an approximation of the graph of $\beta(\alpha)$ for values between $\alpha = 2$ and $\alpha = 8$. The respective slopes are estimated by linear least squares regressions for k between 10^7 and 10^9. Even though

Fig. 2 Relation between α and β. For a given order α, the corresponding value of β was determined by conducting simulations like in Fig. 1a and fitting the slope for k between 10^7 and 10^9. The graph shows that the zero point α^* of the conjectured function $\beta(\alpha)$ has to be in the vicinity of 4



our numerical results do not allow us to establish the precise functional relation between α and β, it becomes clear that α^* indeed has to be close to 4.

For comparison, we additionally conducted simulations that target other asymptotic regimes. First, we keep p_k constant and let n_k again increase with the sample size, $n_k = w\sqrt[\alpha]{k}$. In this scenario, a properly rescaled binomial random variable converges to a standard normal distribution. Our simulations confirm that estimation of n_k is easier in this case: the MSE in Fig. 3a decreases faster when $\alpha = 6$ and $p_k = 0.05$ is fixed compared to $\alpha = 6$ and $p_k \to 0$. The rate of convergence β seems to be independent of the specific choice of $p_k = p$ in this alternative setting, see Fig. 3b. Thus, it makes sense to look at the smallest order α that still exhibits consistency for fixed p. Figure 3c reveals that the estimation of n_k remains consistent over a larger range of values for α in this setting, approximately as long as $\alpha > 2$ (compared to $\alpha > 4$ in the original setting).

The last asymptotic regime we consider is the classical one for parameter estimation, where $n_k = n$ and $p_k = p$ both stay constant as k grows to infinity. Figure 3a covers this regime in the last plot. It affirms that estimating n is easiest in this setting, and we obtain the expected rate $\sim k^{-1}$ for the convergence of the MSE towards zero.

4 Discussion

Theorem 1 (see [11]) shows posterior contraction under diverging parameters n_k and p_k as long as $(n_k, p_k) \in \mathcal{M}_\lambda$, which implies $n_k = O\big(\sqrt[6]{k/\log(k)}\big)$. The aim of our simulation study in Sect. 3 was to explore the minimal rate $\sqrt[\alpha]{k}$ for n_k such that posterior consistency remains valid. The difference in the permissible rates turns out to be rather small, since our investigation suggests that $\alpha = 5$ still allows for consistent estimation, whereas $\alpha = 3$ clearly leads to inconsistency. Figure 2 shows that the true boundary α^* is likely close to 4, indicating that Theorem 1 cannot be improved fundamentally.

Several aspects of our simulations and findings deserve further commentary. First, Fig. 1c reveals that the slope β is not strongly affected by the parameters w and μ in

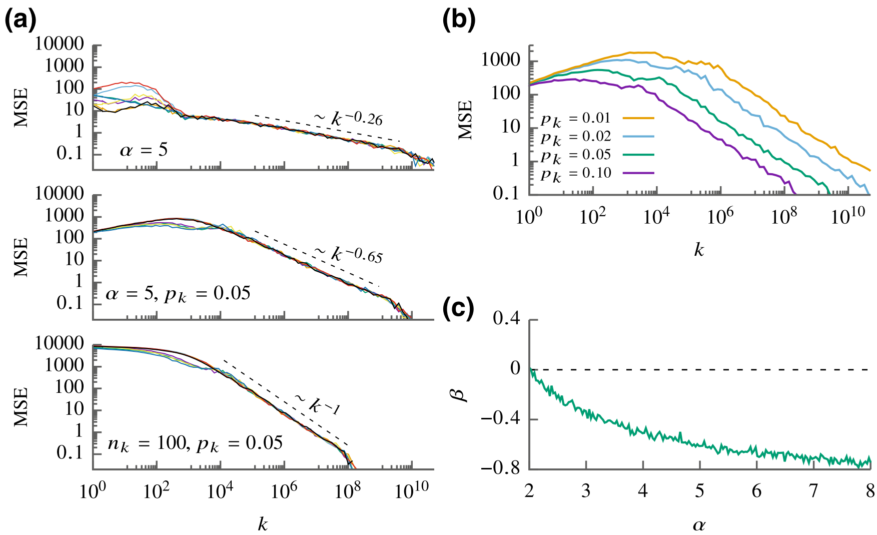


Fig. 3 Comparison of alternative asymptotic settings. **a** shows the MSE for three different asymptotic scenarios. In the first plot, n_k and p_k behave like in Fig. 1 with $w = 16$ and $\mu = 25$. In the second plot, p_k is fixed to the value 0.05, while n_k still increases with k ($w = 16$). The third plot addresses the scenario where both n_k and p_k are held fixed. **b** shows the scenario of growing n_k (with $\alpha = 6$ and $w = 16$) and different fixed values p_k. The graph shows that the slope β in the linear segment does not depend on p_k. **c** shows the relation between β and α for the scenario with fixed p_k and growing n_k. The values of the slopes β are determined as described in Fig. 2, with adapted ranges for k

the settings that we tested. However, our numerical approach is not suitable to verify questions like this with a high degree of confidence. For example, our numerics become instable for values $k > 10^{11}$.

Secondly, we additionally conducted simulations for other estimators than the scale estimator (2) that are not shown in the article. For example, we tested various versions of the Bayesian estimator given in [4]. While their performance for $k \leq 10^3$ varies quite much—similar to the different estimators shown in Fig. 1a—their asymptotic performance is exactly the same as for the scale estimator. Notably, the maximum likelihood estimator also exhibits the very same asymptotic behavior, even though it performs poorly in the regime of smaller k. The sample maximum, in contrast, shows a completely different behavior: the MSE diverges even for $n_k \sim \log(k)$. This illustrates the sharpness of the assumptions for Lemma 5 in [11], which states that the sample maximum is consistent if $n_k \log(n_k) < c \log(k)$ for $c < 1$.

Finally, we consistently observed a phase transition in all simulations when the MSE drops below a value of about 0.1. There, the MSE changes its behavior and begins to decreases faster than $\sim k^{\beta}$. Indeed, it seems to decay exponentially from that point on. We conjecture that this happens due to the discreteness of n, which implies that the MSE cannot measure small deviations $|\hat{n} - n| < 1$ without dropping to zero.

Rather, if the posterior contracts so much that we estimate n correctly most of the time, the MSE essentially captures the probability that $\hat{n}$ lies outside of the interval $(n-1, n+1)$, and such probabilities usually decay exponentially fast. For applications, the rate of the MSE before the exponential decay is often much more interesting. One instructive example in this context is the sample maximum in the setting of fixed n and p, for which we know from Sect. 1 that it converges exponentially fast. However, as argued above, this only takes place when the MSE is already very small, and simulations suggest that the rate of convergence is much slower if the MSE is larger than 0.1. For instance, if $p = 0.2$ and $n = 25$, we find $\beta \approx -0.13$. Thus, even though the true asymptotic behavior of the sample maximum is exponential, the practically meaningful rate of convergence is considerably worse than the rate of the Bayesian estimators, where $\beta = -1$.

Acknowledgements Support of the DFG RTG 2088 (B4) and DFG CRC 755 (A6) is gratefully acknowledged.

References

1. Berger, J.O., Bernardo, J.M., Sun, D.: Objective priors for discrete parameter spaces. J. Am. Stat. Assoc. **107**, 636–648 (2012)
2. Carroll, R.J., Lombard, F.: A note on n estimators for the binomial distribution. J. Am. Stat. Assoc. **80**, 423–426 (1985)
3. DasGupta, A., Rubin, H.: Estimation of binomial parameters when both n, p are unknown. J. Stat. Plan. Inference **130**, 391–404 (2005)
4. Draper, N., Guttman, I.: Bayesian estimation of the binomial parameter. Technometrics **13**, 667–673 (1971)
5. Fisher, R.: The negative binomial distribution. Ann. Hum. Genet. **11**, 182–187 (1941)
6. Günel, E., Chilko, D.: Estimation of parameter n of the binomial distribution. Commun. Stat. Simul. Comput. **18**, 537–551 (1989)
7. Hamedani, G.G., Walker, G.G.: Bayes estimation of the binomial parameter n. Commun. Stat. Theory Methods **17**, 1829–1843 (1988)
8. Kahn, W.D.: A cautionary note for Bayesian estimation of the binomial parameter n. Am. Stat. **41**, 38–40 (1987)
9. Link, W.A.: A cautionary note on the discrete uniform prior for the binomial n. Ecology **94**, 2173–2179 (2013)
10. Raftery, A.E.: Inference for the binomial n parameter: a hierachical Bayes approach. Biometrika **75**, 223–228 (1988)
11. Schneider, L.F., Schmidt-Hieber, J., Staudt, T., Krajina, A., Aspelmeier, T., Munk, A.: Posterior consistency for n in the binomial (n, p) problem with both parameters unknown - with applications to quantitative nanoscopy. arXiv:1809.02443 (2018)
12. van der Vaart, A.W.: Asymptotic Statistics. Cambridge University Press, Cambridge (1998)
13. Villa, C., Walker, S.G.: A cautionary note on using the scale prior for the parameter n of a binomial distribution. Ecology **95**, 2674–2677 (2014)

Learning in the Absence of Training Data—A Galactic Application

Cedric Spire and Dalia Chakrabarty

Abstract There are multiple real-world problems in which training data is unavailable, and still, the ambition is to learn values of the system parameters, at which test data on an observable is realised, subsequent to the learning of the functional relationship between these variables. We present a novel Bayesian method to deal with such a problem, in which we learn the system function of a stationary dynamical system, for which only test data on a vector-valued observable is available, though the distribution of this observable is unknown. Thus, we are motivated to learn the state space probability density function (*pdf*), where the state space vector is wholly or partially observed. As there is no training data available for either this *pdf* or the system function, we cannot learn their respective correlation structures. Instead, we perform inference (using Metropolis-within-Gibbs), on the discretised forms of the sought functions, where the *pdf* is constructed such that the unknown system parameters are embedded within its support. The likelihood of the unknowns given the available data is defined in terms of such a *pdf*. We make an application of this methodology, to learn the density of all gravitating matter in a real galaxy.

Keywords Absence of training data · Bayesian learning · Dark Matter in galaxies · Metropolis-within-Gibbs · State space density

1 Introduction

The study of rich correlation structures of high-dimensional random objects is often invoked when learning the unknown functional relationship between an observed random variable, and some other parameters that might inform on the properties of a system. A problem in which a vector of system parameters (say $\boldsymbol{\rho} \in \mathscr{R} \subseteq \mathbb{R}^p$) is related

C. Spire (✉) · D. Chakrabarty (✉)
Department of Mathematical Sciences, Loughborough University, Loughborough, England
e-mail: c.spire@lboro.ac.uk

D. Chakrabarty
e-mail: d.chakrabarty@lboro.ac.uk

R. Argiento et al. (eds.), *Bayesian Statistics and New Generations*,
Springer Proceedings in Mathematics & Statistics 296,
https://doi.org/10.1007/978-3-030-30611-3_5

to an observed response variable (say $\boldsymbol{Y} \in \mathscr{Y} \subseteq \mathbb{R}^d$), is easily visualised by the equation: $\boldsymbol{Y} = \boldsymbol{\xi}(\boldsymbol{\rho})$, where $\boldsymbol{\xi} : \mathscr{R} \longrightarrow \mathscr{Y}$. Given training data $\mathbf{D} = \{(\boldsymbol{\rho}_i, \boldsymbol{y}_i)\}_{i=1}^{N_{data}}$, we aim to learn this unknown mapping $\boldsymbol{\xi}(\cdot)$ within the paradigm of *supervised learning*. Here, "training data" comprises pairs of chosen design points $\boldsymbol{\rho}_i$, and the output $\boldsymbol{y}_i$ that is generated at the chosen $\boldsymbol{\rho}_i$; $i = 1, \ldots, N_{data}$. Methods to perform supervised learning are extensively covered in the literature [12, 16–18]. Having learnt $\boldsymbol{\xi}(\cdot)$, one could use this model to predict the value $\boldsymbol{\rho}$ [5], at which the test datum $\boldsymbol{y}_{test}$ on $\boldsymbol{Y}$ is realised—either in the conventional framework as $\boldsymbol{\rho} = \boldsymbol{\xi}^{-1}(\boldsymbol{Y})|_{\boldsymbol{Y}=\boldsymbol{y}_{test}}$, or as the Bayesian equivalent. Such prediction is possible, only subsequent to the learning of the functional relation between $\boldsymbol{\rho}$ and $\boldsymbol{Y}$ using training data $\mathbf{D}$.

However, there exist physical systems for which only measurements on the observable $\boldsymbol{Y}$ are known, i.e. training data is not available. The disciplines affected by the absence of training data are diverse. In engineering [20], anomaly detection is entirely sample-specific. There are no training data that allow for the learning of a functional relationship between anomaly occurrence (parametrised by type and severity of anomaly), and conditions that the sample is subjected to. Yet, we need to predict those anomalies. In finance, such anomalies in stock price trends are again outside the domain of supervised learning, given that the relationship between the market conditions and prices have not been reliably captured by any "models" yet. In neuroscience [1], a series of neurons spike at different amplitudes, and for different time widths, to cause a response (to a stimulus). We can measure the response's strength and the parameters of firing neurons, but do not know the relation between these variables. Again, in petrophysics, the proportion of the different components of a rock (e.g. water, hydrocarbons), affects Nuclear Magnetic Resonance (NMR) measurements from the rock [7, 21]. However, this compositional signature cannot be reliably estimated given such data, using available estimation techniques. Quantification of petrological composition using the destructive testing of a rock is highly exclusive and expensive, to allow for a sample that is large and diverse enough to form a meaningful training data set that counters the rock-specific, latent factors (geological influences) affecting the system property (composition). Equally, dependence on such latent geological influence annuls the possibility of using numerical simulations to generate NMR data, at chosen compositional values. Thus, generation of training data is disallowed.

In this work, we capacitate learning of the (possibly high-dimensional) functional relation between an observable, and a system parameter vector, in such a challenging (absent training) data situation, given ignorance on the distribution of the observable. This could then be undertaken as an exercise in supervised learning, as long as the missing training data is generated, i.e. we are able to generate the system parameter vector $\boldsymbol{\rho}_i$ at which the measured (test) datum, $\boldsymbol{y}_i$ on $\boldsymbol{Y}$, is recorded, $\forall i \in \{1, \ldots, N_{data}\}$. Our new method invokes a system property that helps link $\boldsymbol{\rho}$ with $\boldsymbol{Y}$, and this is possible in physical systems for which we have—at least partial—observed information. To clarify, in the face of absent training data, we advance the pursuit of the probability density function of the observable $\boldsymbol{Y}$, on which data is available, and employ this to learn the system parameter vector $\boldsymbol{\rho}$. We undertake such an exercise in a Bayesian framework, in which we seek the posterior of the *pdf* of the observables, and the system parameters, given the available data.

The sought parameter vector could inform on the behaviour, or structure, of the system (e.g. it could be the vectorised version of the density function of all gravitating matter in a distant galaxy). The state space *pdf* establishes the link between this unknown vector, and measurements available on the observable (that may comprise complete or incomplete information on the state space variable). We consider dynamical systems, such that the system at hand is governed by a kinetic equation [11]; we treat the unknown system parameter vector as the stationary parameter in the model of this dynamical system. In the novel Bayesian learning method that we introduce, this parameter is embedded within the support of the state space *pdf*. We describe the general model in Sect. 2, that is subsequently applied to an astronomical application discussed in Sect. 3. Inference is discussed in Sect. 4, where inference is made on the state space *pdf* and the sought system parameters, given the data that comprises measurements of the observable, using Metropolis-within-Gibbs. Results are presented in Sect. 5, and the paper is rounded up with a conclusive section (Sect. 6).

2 General Methodology

We model the system as a dynamical one, and define the state space variable as a p-dimensional vector $\boldsymbol{S} \in \mathscr{S} \subseteq \mathbb{R}^p$. Let the observable be $\boldsymbol{Y} \in \mathscr{Y} \subseteq \mathbb{R}^d$; $d < p$, such that only some (d) of the p different components of the state space vector $\boldsymbol{S}$ can be observed. In light of this situation that is marked by incomplete information, we need to review our earlier declaration of interest in the probability density function of the full state space vector. Indeed, we aim to learn the *pdf* of the state space variable $\boldsymbol{S}$, and yet, have measured information on only $\boldsymbol{Y}$, i.e. on only d of the p components of $\boldsymbol{S}$. The data $\boldsymbol{D} = \{\boldsymbol{y}^{(k)}\}_{k=1}^{N_{data}}$ is then one set of measurements of the observable $\boldsymbol{Y}$. If the density of $\boldsymbol{S}$ is to be learnt given data on $\boldsymbol{Y}$, such incompleteness in measured information will have to be compensated for, by invoking independent information. Such independent information comprises symmetry of $\mathscr{S}$.

It follows that unobserved components of $\boldsymbol{S}$ will have to be integrated out of the state space *pdf*, in order to compare against data that comprises measurements of observables. This is equivalent to projecting the state space *pdf* onto the space $\mathscr{Y}$ of observables, and therefore, we refer to the result as the *projected state space pdf*. The likelihood of the model parameters, given the data, is simply the product of the projected state space *pdf* over all the data points (assuming iid data points). But until now, the unknown model parameters have not yet appeared in our expression of the likelihood. The next step is then to find a way for embedding the sought system parameters in the support of the projected state space *pdf*.

This can be achieved by assuming that our dynamical system is stationary, so that the rate of change of the state space *pdf* is 0. This allows us to express the *pdf* as dependent on the state space vector $\boldsymbol{S}$, but only via such functions of (some, or all, amongst) $S_1, \ldots, S_p$ that are not changing with time; in fact, the converse of this statement is also true. This is a standard result, often referred to as Jeans Theorem [3, 14]. The model parameters that we seek, can be recast as related to such identified time-independent functions of all/some state space coordinates of motion. Thus, by expressing the state space *pdf* as a function of appropriate constants of motion, we can embed system parameters into the support of the sought *pdf*.

As stated above, this *pdf* will then need to be projected into the space of observables $\mathscr{Y}$, and we will convolve such a projected *pdf* with the error density, at every choice of the model parameters. Then assuming data to be iid, the product of such a convolution over the whole dataset will finally define our likelihood. Using this likelihood, along with appropriate priors, we then define the posterior probability density of the model parameters and the state space *pdf*, given the data $\boldsymbol{D}$. Subsequently, we generate posterior samples using Metropolis-within-Gibbs.

We recall that in absence of training data on a pair of random variables, we cannot learn the correlation structure of the functional relationship between these variables. In such situations, instead of the full function, we can only learn the vectorised version of the sought function. In other words, the relevant interval of the domain of the function is discretised into bins, and the value of the function is held constant over any such bin; we can learn the functional value over this bin.

3 Astrophysics Application

Our astrophysics application is motivated to learn the contribution of dark matter, to the density function of all gravitating mass in a distant galaxy. While information on light-emitting matter is available, it is more challenging to model the effects of dark matter, since, by definition, one cannot observe such matter (as it does not emit or reflect light of any colour). However, the following physical phenomena confirm that unobservable dark matter is contributing to the overall gravitational mass density of the galaxy: distortion of the path of light by gravitational matter acting as gravitational lenses; temperature distribution of hot gas that is emanating from a galaxy; motions of stars or other galactic particles that are permitted despite the attractive gravitational pull of the surrounding galactic matter. In fact, astrophysical theories suggest that the proportion of dark matter in older galaxies (that are of interest to us here) is the major contributor to the galactic mass, over the minor fraction of luminous matter [13]. We can compute this proportion, by subtracting the density of luminous matter from the overall density. It is then necessary to learn the gravitational mass density of the whole system, in order to learn dark matter density.

We begin by considering the galaxy at hand to be a stationary dynamical system, i.e. the distribution of the state space variable does not depend on time. Let $\boldsymbol{S} = (X_1, X_2, X_3, V_1, V_2, V_3)^T \in \mathscr{S} \subseteq \mathbb{R}^6$ define the state space variable of a galactic particle, where $\boldsymbol{X} = (X_1, X_2, X_3)^T$ is defined as its 3-dimensional location vector and $\boldsymbol{V} = (V_1, V_2, V_3)^T$ as the 3-dimensional velocity vector. The data consists of measurements of the one observable velocity coordinate V_3, and two observable spatial coordinates, X_1, X_2, of N_{data} galactic particles (e.g. stars). That is, for each galactic particle, we have measurements of $\boldsymbol{Y} = (X_1, X_2, V_3)^T \in \mathscr{Y} \subseteq \mathbb{R}^3$. For N_{data} observations, our data is thus shaped as an $N_{data} \times 3$-dimensional matrix, $\boldsymbol{D} = \{\boldsymbol{y}^{(k)}\}_{k=1}^{N_{data}}$.

The system function that we are interested in learning here, is the density function $\rho(X_1, X_2, X_3)$ of the gravitational mass of all matter in the considered galaxy, where we assume that this gravitational mass density $\rho(\cdot)$ is a function of the spatial coordinates $\boldsymbol{X}$ only. This system function does indeed inform on the structure of the galactic system—for it tells us about the distribution of matter in the galaxy; it also dictates the behaviour of particles inside the galaxy, since the gravitational mass density is deterministically known as a function of the gravitational potential $\Phi(X_1, X_2, X_3)$ via the Poisson equation ($\nabla^2\Phi(X_1, X_2, X_3) = -4\pi G\rho(X_1, X_2, X_3)$, where G is the known Universal Gravitational constant, and ∇^2 is the Laplacian operator), which is one of the fundamental equations of Physics [10]. The potential of a system, along with the state space distribution, dictates system dynamics.

Here, we assume that the state space density of this dynamical system does not vary with time, i.e. $df\,[X_1(t), X_2(t), X_3(t), V_1(t), V_2(t), V_3(t)]/dt = 0$. This follows from the consideration that within a typical galaxy, collisions between galactic particles are extremely rare [3]. We thus make the assumption of a collisionless system evolving in time, according to the *Collisionless Boltzmann Equation* (CBE) [3, 6]. As motivated above, this allows us to express the state space *pdf* as dependent on those functions of $X_1, X_2, X_3, V_1, V_2, V_3$ that remain invariant with time, along any trajectory in the state space $\mathscr{S}$; such time-invariant constants of motion notably include energy, momentum, etc. It is a standard result that the state space *pdf* has to depend on the energy $E(X_1, X_2, X_3, \parallel \boldsymbol{V} \parallel)$ of a galactic particle [2, 8], where $\parallel \cdot \parallel$ represents the Euclidean norm of a vector. Here, energy is given partly by kinetic energy that is proportional to $\parallel \boldsymbol{V} \parallel^2$, and partly by potential energy, which by our assumption, is independent of velocities. Secondly, given that the state space is 6-dimensional, the number of constants of motion must be less than or equal to 5, to allow the galactic particle at least 1 degree of freedom, i.e. not be fixed in state space [8].

We ease our analysis by assuming that the state space *pdf* is a function of energy only. This can be rendered equivalent to designating the symmetry of isotropy to the state space $\mathscr{S}$, where isotropy implies invariance to rotations, i.e. the state space *pdf* is assumed to be such a function of $\boldsymbol{X}$ and $\boldsymbol{V}$, that all orthogonal transformations of $\boldsymbol{X}$ and $\boldsymbol{V}$ preserve the state space *pdf*. The simple way to achieve the equivalence between an isotropic state space *pdf* and the lone dependence on energy E of the *pdf*, is to ensure that the gravitational mass density, (and therefore the gravitational potential), at all points at a given Euclidean distance from the galactic centre, be the same, i.e. the distribution of gravitational mass abides by spherical symmetry such that $\rho(\cdot)$ (and therefore $\Phi(\cdot)$) depends on X_1, X_2, X_3 via the Euclidean norm $\parallel \boldsymbol{X} \parallel$ of the location vector $\boldsymbol{X}$, of a particle. Then energy E is given as the sum of the $\parallel \boldsymbol{V} \parallel^2$-dependent kinetic energy, and the $\parallel \boldsymbol{X} \parallel$-dependent potential energy. Spherical mass distribution is not a bad assumption in the central parts of "elliptical" galaxies that are of interest for us, as these have a global triaxial geometry. To summarise, state space *pdf* is written as $f(E)$, and we embed $\boldsymbol{\rho}(\cdot)$ into its support, by recalling that energy E is partly the gravitational potential energy $\Phi(\cdot)$ that is deterministically related to the gravitational mass density $\rho(\cdot)$, through Poisson equation.

As there is no training data available to learn the correlation structure of the sought functions $\rho(\boldsymbol{X})$ and $f(E)$, we can only learn values of these functions at specified points in their domains, i.e. learn their vectorised forms $\boldsymbol{\rho}$ and $\boldsymbol{f}$ respectively, where $\boldsymbol{\rho} := (\rho_1, ..., \rho_{N_X})^T$, with $\rho_i = \rho(\boldsymbol{x})$ for $\boldsymbol{x} \in [\boldsymbol{x}_{i-1}, \boldsymbol{x}_i]$; $i = 1, \ldots N_x$. The discretised form of $f(E)$ is similarly defined, after partitioning the relevant range of (non-positive) E-values (to indicate that the considered galactic particles are bound to the galaxy by gravitational attraction), into N_E number of E-bins. Then, in terms of these vectorised versions of the state space *pdf*, likelihood of the unknown parameters $\rho_1, \ldots \rho_{N_X}, f_1, \ldots, f_{N_E}$, given data on the observable $\boldsymbol{Y}$ is:

$$\ell\left(\boldsymbol{\rho}, \boldsymbol{f} | \{\boldsymbol{y}^{(k)}\}_{k=1}^{N_{data}}\right) = \prod_{k=1}^{N_{data}} \nu(\boldsymbol{y}^{(k)}, \boldsymbol{\rho}, \boldsymbol{f}), \tag{1}$$

where $\nu(.)$ is the projected state space *pdf*.

We also require that $\rho_1 \geq 0, \ldots \rho_{N_X} \geq 0$, $f_1 \geq 0, \ldots, f_{N_E} \geq 0$, and that $\rho_i \geq \rho_{i+1}$, $i = 1, \ldots, N_X - 1$. The latter constraint is motivated by how the mass in a gravitating system (such as a galaxy) is distributed; given that gravity is an attractive force, the stronger pull on matter closer to the centre of the galaxy, implies that gravitational mass density should not increase as we move away from the centre of the system. These constraints are imposed via the inference that we employ.

4 Inference

Inference on the unknown parameters—that are the components of $\boldsymbol{\rho}$ and $\boldsymbol{f}$—is undertaken using Metropolis-within-Gibbs. In the first block update during any iteration, the $\rho_1, \ldots, \rho_{N_X}$ parameters are updated, and subsequently, the $f_1, \ldots, f_{N_E}$ parameters are updated in the 2nd block, at the updated ρ-parameters, given the data $\boldsymbol{D}$ that comprises N_{data} measurements of the observed state space variables X_1, X_2, V_3 that are the components of the observable vector $\boldsymbol{Y}$.

Imposition of the monotonicity constraint on the ρ parameters, to ensure $\rho_i \geq \rho_{i+1}$, $i = 1, \ldots N_X - 1$, renders the inference interesting. We propose ρ_i from a truncated normal proposal density that is left truncated at ρ_{i+1}, $\forall i = 1, \ldots, N_X - 1$, and propose ρ_{N_X} from a truncated normal that is left truncated at 0. The mean of the proposal density is the current value of the parameter and the variance is experimentally chosen, as distinct for each $i \in \{1, \ldots, N_X\}$. Such a proposal density helps to maintain the non-increasing nature of the ρ_i-parameters, with increasing i. At the same time, non-negativity of these parameters is also maintained. We choose arbitrary seeds for $\rho_1, \ldots, \rho_{N_X}$, and using these as the means, a Gaussian prior is imposed on each parameter. The variance of the prior density on ρ_i, is kept quite large ($\sim 10^6$ times the chosen prior mean that is set equal to the seed), and demonstration of lack of sensitivity to the prior choices, as well as the seeds, is undertaken. A vague Gaussian prior with a large variance approximates a flat prior, thus ensuring that inference is not prior-driven. There is no correlation information on components of the vectorised state space *pdf*, unlike in the case of the components of the vectorised gravitational mass density function. We propose f_j from a truncated normal (to maintain non-negativity), where the mean of this proposal density is the current value of the parameter and the variance is chosen by hand. Vague Gaussian priors are imposed, while the same seed value is used $\forall j \in \{1, \ldots, N_E\}$.

An important consideration in our work is the choice of N_X and N_E. We could have treated these as unknowns and attempted learning these from the data; however, that would imply that the number of unknowns will vary from one iteration to another, and we desired to avoid such a complication, especially since the data strongly suggests values of N_X and N_E. We choose N_X by partitioning the range of $R_p := \sqrt{X_1^2 + X_2^2}$ values in the data $\boldsymbol{D}$, such that, each resulting R_p-bin includes at least one observed value of V_3 in it, and at the same time, the number of R_p-bins is maximised. Again, we use the available data $\boldsymbol{D}$ to compute the empirical values of energy E, where an arbitrarily scaled histogram of the observed R_p is used to mimic the vectorised gravitational mass density function, that is then employed to compute the empirical estimate of the vectorised gravitational potential function, that contributes to E values. We admit maximal E-bins over the range of the empirically computed values of E, such that each such E-bin contains at least one datum in $\boldsymbol{D}$.

5 Results

We have input data on location and velocities of 2 kinds of galactic particles (called "Globular Clusters", and "Planetary Nebulae"—respectively abbreviated as GC and PNe), available for the real galaxy NGC4494. The GC data comprises 114 measurements of $\boldsymbol{Y} = (X_1, X_2, V_3)^T$, for the GCs in NGC4494 [9]. Our second dataset (PNe data), comprises 255 measurements of PNe [15]. Results obtained using the PNe data and GC data are displayed in Fig. 1. Inconsistencies between gravitational mass density parameters learnt from data on different types of particles can suggest interesting dynamics, such as splitting of the galactic state space into multiple, non-communicating sub-spaces [4], but for this galaxy, parameters learnt from the 2 datasets, concur within the learnt 95% Highest Probability Density credible regions.

6 Conclusions

An astronomical implication of our work is that ρ_1 learnt from either dataset suggests a very high gravitational mass density in the innermost R_p-bin ($\approx$1.6 kpc), implying gravitational mass $\gtrsim 10^9$times mass of the Sun, enclosed within this innermost radial

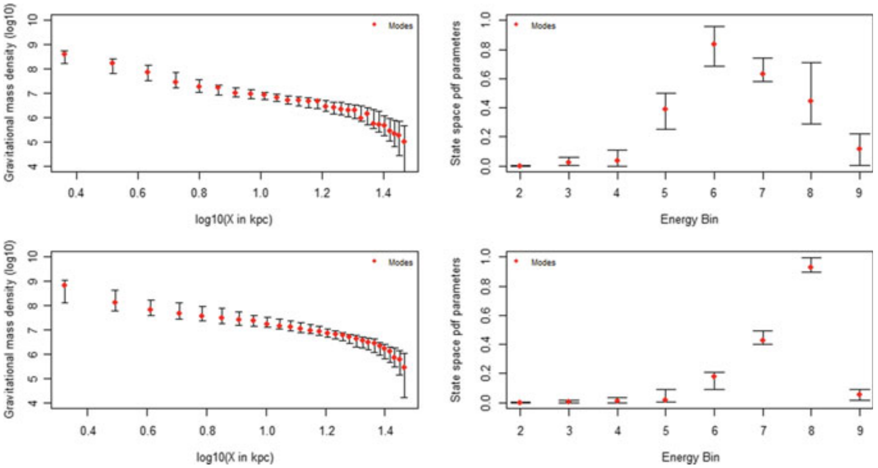


Fig. 1 Results from Metropolis-within-Gibbs showing the 95% Highest Probability Densities (HPDs) for all the parameters to learn, for both PNe (top row) and GC (bottom row) data. Mode of the marginal of each learnt parameter is shown as a red dot. Results on $\boldsymbol{\rho}$ are on the left, and on the $\boldsymbol{f}$ parameters on the right; these are depicted for the PNe data in the top row, and for the GC data in the bottom row

bin. This result alone does not contradict the suggestion that NGC4494 harbours a central supermassive blackhole (SMBH) of mass $\sim 2.69 \pm 2.04 \times 10^7$ solar masses [19]. Very interestingly, our results indicate that for both GCs and PNe, most particles lie in the intermediate range of energy values; this is also borne by the shape of the histogram of the empirically computed energy using either dataset, where this empirical E value computation is discussed in the last paragraph of Sect. 4. However, owing to its intense radially inward gravitational attraction, a central SMBH is expected to render the potential energy (and therefore the total energy E) of the particles closer to the galactic centre to be much higher negative values than those further away, while also rendering the number (density) of particles to be sharply (and monotonically) decreasing with radius away from the centre. This is expected to render the energy distribution to be monotonically decreasing as we move towards more positive E values—in contradiction to our noted non-monotonic trend. So while our results are not in contradiction to the report of a very large value of mass enclosed within the inner parts of NGC4494, interpretation of that mass as a SMBH does not follow from our learning of the state space *pdf*.

Supervised learning of the gravitational mass density function, and state space *pdf*—as well as that of the relation $\boldsymbol{\xi}(\cdot)$ between the observable state space coordinates, and the system function (or vector)—can be undertaken after generating the training dataset relevant to the functional learning problem at hand. Applications in petrophysics and finance are also planned.

References

1. Ahmad, S., Lavin, A., Purdy, S., Agha, Z.: Unsupervised real-time anomaly detection for streaming data. Neurocomputing, vol. 262 (2017). https://doi.org/10.1016/j.neucom.2017.04.070
2. Binney, J.: Dynamics of elliptical galaxies and other spheroidal components. Annu. Rev. Astron. Astrophys. **20**, 399–429 (1982)
3. Binney, J., Tremaine, S.: Galactic Dynamics. Princeton University Press, Princeton (1987)
4. Chakrabarty, D.: A new bayesian test to test for the intractability-countering hypothesis. J. Am. Stat. Assoc. **112**, 561–577 (2017)
5. Chakrabarty, D., Biswas, M., Bhattacharya, S.: Bayesian nonparametric estimation of Milky Way parameters using matrix-variate data, in a new gaussian process based method. Electron. J. Stat. **9**(1), 1378–1403 (2015)
6. Choudhuri, A.: Astrophysics for Physicists. Cambridge University Press, Cambridge (2010)
7. Coates, G.R., Xhao, L., Prammer, M.G.: NMR Logging Principles & Applications. Halliburton Energy Services Publication H02308, Houston (1999)
8. Contopoulos, G.: A classification of the integrals of motion. Astrophys. J. **138**, 1297–1305 (1963)
9. Foster, C., et al.: Global properties of ordinary early-type galaxies: photometry and spectroscopy of stars and globular clusters in NGC 4494. Mon. Not. R. Astron. Soc. **415**, 3393–3416 (2011)
10. Goldstein, H., Poole, C.P., Safko, J.: Classical Mechanics. Addison-Wesley Longman, Incorporated (2002)
11. Gressman, P.T., Strain, R.M.: Global classical solutions of the Boltzmann equation with long-range interactions. In: Proceedings of the National Academy of Sciences (2010)
12. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer, New York (2009)
13. Kalinova, V.: Mass distributions of galaxies from SAURON and CALIFA stellar kinematic maps. Doctoral thesis, Max-Planck-Institut fur Astronomie (2014)
14. Merritt, D.: Dynamics and Evolution of Galactic Nuclei. Princeton University Press, Princeton (2013)
15. Napolitano, N., et al.: The Planetary Nebula Spectrograph elliptical galaxy survey: the dark matter in NGC 4494. Mon. Not. R. Astron. Soc. **393**, 329–353 (2009)
16. Neal, R.M.: Regression and classification using gaussian process priors (with discussion). In: Bernardo, J.M., et al. (eds.) Bayesian Statistics, vol. 6, pp. 475–501. Oxford University Press, Oxford (1998)
17. Rasmussen, C.E., Williams, C.K.: Gaussian Processes for Machine Learning. The MIT Press, MIT (2006)
18. Russel, S., Norvig, P.: Artificial Intelligence: A Modern Approach, 3rd edn. Pearson, New York (2009)
19. Sadoun, R., Colin, J.: $\mathrm{M}_{BH} - \sigma$ relation between supermassive black holes and the velocity dispersion of globular cluster systems. Mon. Not. R. Astron. Soc.: Lett. **426**(1), L51–L55 (2012)
20. Sun, X., Yao, H., Ji, R., Liu, X., Xu, P.: Unsupervised fast anomaly detection in crowds. In: Proceedings of the 19th ACM International Conference on Multimedia, MM '11, pp. 1469–1472. ACM, New York, NY, USA (2011). https://doi.org/10.1145/2072298.2072042. http://doi.acm.org/10.1145/2072298.2072042
21. Wang, P., Jain, V., Venkataramanan, L.: Sparse Bayesian T1-T2 inversion from borehole NMR measurements. In: Proceedings of SPWLA 57 Annual Logging Symposium, 25–29 June 2016, Reykjavik (2016)

Multiplicative Latent Force Models

Daniel J. Tait and Bruce J. Worton

Abstract Bayesian modelling of dynamic systems must achieve a compromise between providing a complete mechanistic specification of the process while retaining the flexibility to handle those situations in which data is sparse relative to model complexity, or a full specification is hard to motivate. Latent force models achieve this dual aim by specifying a parsimonious linear evolution equation with an additive latent Gaussian process (GP) forcing term. In this work we extend the latent force framework to allow for multiplicative interactions between the GP and the latent states leading to more control over the geometry of the trajectories. Unfortunately inference is no longer straightforward and so we introduce an approximation based on the method of successive approximations and examine its performance using a simulation study.

Keywords Gaussian processes · Latent force models

1 Introduction

Modern statistical inference must often achieve a balance between an appeal to the *data driven paradigm* whereby models are flexible enough to allow inference to be chiefly driven by the observations, and on the other hand the *mechanistic approach* whereby the structure of the data generating process is well specified up to some, usually modest, set of random parameters. The conflict between these two philosophies can be particularly pronounced for complex dynamic systems for which a complete mechanistic description is often hard to motivate and instead we would like a framework that allows for the specification of a, potentially over-simplistic, representative evolution equation which would enable the modeller to embed as much

D. J. Tait (✉) · B. J. Worton
School of Mathematics, University of Edinburgh, Edinburgh, Scotland
e-mail: Tait.djk@gmail.com

B. J. Worton
e-mail: Bruce.Worton@ed.ac.uk

R. Argiento et al. (eds.), *Bayesian Statistics and New Generations*,
Springer Proceedings in Mathematics & Statistics 296,
https://doi.org/10.1007/978-3-030-30611-3_6

prior knowledge as they feel comfortable doing while at the same time ensuring the model is sufficiently flexible to allow for any unspecified dynamics to be captured during the inference process.

Such a compromise is provided by a class of hybrid models introduced in [1] which they term *latent force model* (LFM). This is a combination of a simple mechanistic model with added flexibility originating from a flexible Gaussian process (GP) forcing term. The aim is to encode minimal dynamic systems properties into the resulting state trajectories without necessarily having to provide a complete mechanistic description of how the system evolves.

One of the appealing features of the LFM is the fact that the resulting trajectories are given by Gaussian processes and therefore inference can proceed in a straightforward manner. However, for many classes of systems the Gaussian trajectories are unlikely to be realistic; examples include time series of circular, directional or tensor valued data. For all of these cases, if we have a suitably dense sample then the Gaussian trajectory assumption may be acceptable, however when data are sparse comparative to model complexity we would like to be able to consider models that move beyond this assumption and allow a priori embedding of geometric constraints.

In this paper we briefly review the LFM before introducing our extension in Sect. 3 and then discuss how our model now allows for the embedding of strong geometric constraints. Unfortunately it is no longer straightforward to solve for the trajectories as some transformation of the latent random variables and therefore in Sect. 4 we introduce an approximate solution method for this class of models based on the method of successive approximations for the solution of certain integral equations. We then demonstrate by way of a simulation study that our approximate model performs well for cases which possess a solvable ground truth.

2 Latent Force Models

The LFM was initially proposed as a model of the transcriptional regulation of gene activities in [3, 5], in subsequent developments the modelling philosophy shifted from this mechanistic perspective to the hybrid setting in [1]. For a K-dimensional state variable $\mathbf{x}(t) \in \mathbb{R}^K$ the first order LFM is described by a system of ordinary differential equations (ODE) in matrix-vector form as

$$\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} = -\mathbf{D}\mathbf{x}(t) + \mathbf{b} + \mathbf{S}\mathbf{g}(t), \tag{1}$$

where $\mathbf{D}$ is a $K \times K$ real-valued diagonal matrix, $\mathbf{b}$ is a real-valued K-vector and $\mathbf{g}(t)$ is the $\mathbb{R}^R$-valued stochastic process with smooth independent GP components $g_r(t)$, $r = 1, \ldots, R$. In this work the kernel functions of the GP terms are chosen so that the sample paths are almost surely smooth, allowing (1) to be interpreted as an ODE rather than as a stochastic differential equation. The $K \times R$ rectangular sensitivity matrix $\mathbf{S}$ acts to distribute linear combinations of the independent latent forces to each component of the evolution equation.

The model (1) gives only an implicit link between the latent random variables and the observed trajectories, but to carry out inference we would ideally like to represent this connection as an explicit transformation. It turns out that for the model (1) with constant coefficient matrix and additive inhomogeneous forcing term this is easily done and an explicit solution is given by

$$\mathbf{x}(t) = e^{-\mathbf{D}(t-t_0)}\mathbf{x}(t_0) + \int_{t_0}^{t} e^{-\mathbf{D}(t-\tau)}\mathrm{d}\tau \cdot \mathbf{b} + L[\mathbf{g}](t), \tag{2}$$

where $L[f](t)$ is the linear integral transformation acting on functions $f : \mathbb{R} \rightarrow \mathbb{R}^R$ to produce a function $L[f] : \mathbb{R} \rightarrow \mathbb{R}^K$ given by

$$L[f](t) = \int_{t_0}^{t} e^{-\mathbf{D}(t-\tau)}\mathbf{S}f(\tau)\mathrm{d}\tau. \tag{3}$$

The decomposition of the solution of the LFM (2) makes it clear that, for given values of the initial condition $\mathbf{x}(t_0)$ and the model parameters $\boldsymbol{\theta} = (\mathbf{D}, \mathbf{b}, \mathbf{S})$, the trajectory is given by a linear integral transformation of the smooth latent GPs, and it follows that the trajectory and the latent force variables will have a joint Gaussian distribution. This property enables a marginalisation over the latent GPs and so allows the LFM to be viewed as a particular instance of a GP regression model. In this interpretation the model parameters are to be regarded as kernel hyperparameters, and inference for these variables may be done using standard techniques, see [6].

3 Multiplicative Latent Force Models

While from a computational point of view the GP regression framework of the LFM is appealing we would like to move beyond the restriction of having Gaussian state trajectories. We therefore introduce an extension of the LFM which will allow us to represent non-Gaussian trajectories while at the same time keeping the same fundamental components: a linear ODE with the time dependent behaviour of the evolution equation coming from a set of independent smooth latent forces. In matrix/vector form our model is given by

$$\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} = \mathbf{A}(t)\mathbf{x}(t), \qquad \mathbf{A}(t) = \mathbf{A}_0 + \sum_{r=1}^{R} \mathbf{A}_r \cdot g_r(t). \tag{4}$$

The coefficient matrix $\mathbf{A}(t)$ will be a square matrix of dimension $K \times K$ formed by taking linear combinations of a set of *structure matrices* $\{\mathbf{A}_r\}_{r=0}^{R}$ which we multiply by scalar GPs. By linearity $\mathbf{A}(t)$ will be a Gaussian process in $\mathbb{R}^{K \times K}$ although typically the choice of the set of structure matrices will be guided by geometric considerations and in general the dimension of this space will be much less than that of the ambient K^2 dimensional space. In the specification (4) the matrix valued Gaussian process $\mathbf{A}(t)$ will interact multiplicatively with the state variable in the evolution equation, rather than as an additive forcing term in (1), and so we refer to this model as the *multiplicative latent force model* (MLFM).

Since the GP terms are smooth (4) has, almost surely, a unique pathwise solution on compact intervals $[0, T]$. While for the LFM it was possible to use (2) to perform a marginalisation over the latent forces and so learn structural parameters independently of the variables we cannot guarantee the existence of the marginal distribution to (4). The existence of L_p solutions to this problem are considered in [8] from which it may be possible to construct moment matching approximations. Because we cannot perform this marginalisation we restrict ourselves to approximations constructed around a dense realisation of the GP terms which, with increasingly fine partitions of the interval, will approach the unique sample path solution.

The multiplicative interaction in (4) and the freedom to choose the support of the coefficient matrix will allow us to embed strong geometric constraints on solutions to ODEs of this form. In particular, by choosing the elements $\{A_r\}$ from some Lie algebra $\mathfrak{g}$ corresponding to a Lie group G then the fundamental solution of (4) will itself be a member of the group G [4], allowing dynamic models with trajectories either within the group itself or formed by an action of this group on a vector space.

4 Method of Successive Approximations

In general non-autonomous linear ODEs do not possess a closed form solution and therefore it is no longer straightforward to carry out inference for the MLFM; we lack the explicit representation of the trajectories in terms of the latent random processes which was possible for the LFM using the solution (2). To proceed we first note that a pathwise solution to the model (4) on the interval $[0, T]$ is given by

$$\mathbf{x}(t) = \mathbf{x}(0) + \int_0^t \mathbf{A}(\tau)\mathbf{x}(\tau)\mathrm{d}\tau, \qquad 0 \leq t \leq T,$$

a solution to which can be obtained by starting from an initial approximation of the trajectory, $\mathbf{x}_0(t)$, and then repeatedly iterating the linear integral operator

$$\mathbf{x}_{m+1}(t) = \mathbf{x}_0(0) + \int_0^t \mathbf{A}(\tau)\mathbf{x}_m(\tau)\,\mathrm{d}\,\tau. \tag{5}$$

This process is known as the *method of successive approximations* and is a classical result in the existence and uniqueness theorems for the solutions of ODEs.

We introduce some probabilistic content into this approximation by placing a mean zero Gaussian process prior on the initial state variable $\mathbf{x}_0(t)$ independent of the latent force terms. Since (5) is a linear operator for known $\mathbf{A}(t)$ and $\mathbf{x}_m(t)$ then the marginal distribution of the $(m+1)$th successive approximation conditional on the process $\mathbf{A}(t)$ will be mean zero Gaussian with covariance given recursively by

$$\tilde{\mathbb{E}}\left[\mathbf{x}_{m+1}(t)\mathbf{x}_{m+1}(t')^{\top}\right] = \int_{t_0}^{t}\int_{t_0}^{t'} \mathbf{A}(\tau)\tilde{\mathbb{E}}[\mathbf{x}_m(\tau)\mathbf{x}_m(\tau')^{\top}]\mathbf{A}(\tau')^{\top}\,\mathrm{d}\,\tau\,\mathrm{d}\,\tau', \qquad (6)$$

where $\tilde{\mathbb{E}}$ denotes expectation conditional on the stochastic process $\mathbf{A}(t)$ on $[0, T]$.

In practice, we will not be dealing with complete trajectories, but instead with the process observed at a finite set of points $t_0 < \cdots < t_N$, and so we replace the map (5) by a numerical quadrature

$$\mathbf{x}(t_0) + \int_{t_0}^{t_i} \mathbf{A}(\tau)\mathbf{x}(\tau)\mathrm{d}\tau \approx \mathbf{x}(t_0) + \sum_{j=1}^{N_i} \mathbf{A}(\tau_{ij})\mathbf{x}(\tau_{ij})w_{ij}, \qquad i = 1, \ldots, N, \qquad (7)$$

for a set of weights $\{w_{ij}\}$ which are determined by our choice of quadrature rule and we have a set of nodes τ_{ij} labelled such that $\tau_{i1} = t_{i-1}$ and $\tau_{iN_i} = t_i$. It follows that methods with more than two nodes over a particular interval $[t_i, t_{i+1}]$ must necessarily augment the latent state vector. Increasing the number of nodes will cause the error in (7) to decrease, we defer discussion of the finer points of this approximation, but for practical purposes the important detail is that this error can be made arbitrarily small because we are free to increase the resolution of the trajectories by treating this as a missing data problem albeit with a corresponding computational cost. In terms of a linear operator acting on the whole trajectory we replace the operator (5) with a matrix operator $K[\mathbf{g}]$ acting on the discrete trajectories such that each row of $K[\mathbf{g}]$ performs the quadrature (7), that is if $\mathbf{x}$ is a dense realisation of a continuous process $\mathbf{x}(t)$ evaluated at the points $\{\tau_{ij}\}$ then

$$(\mathbf{K}[\mathbf{g}]\mathbf{x})_i = \mathbf{x}(t_0) + \sum_{j=1}^{N_i} \mathbf{A}(\tau_{ij})\mathbf{x}(\tau_{ij})w_{ij}, \qquad i = 1, \ldots, N. \qquad (8)$$

For suitably dense realisations of the trajectory we can conclude that the majority of the informational content in the linear map (5) is captured by applying the matrix operator form of the integral operator (8) and therefore there will be minimal loss of information if we replace the (Gaussian) correlated error term with an *independent* additive noise term leading to a conditional distribution of the form

$$p(\mathbf{x}_{m+1} \mid \mathbf{x}_m, \mathbf{g}, \mathbf{\Gamma}) = \mathcal{N}\left(\mathbf{x}_{m+1} \mid K[\mathbf{g}]\mathbf{x}_m, \mathbf{\Gamma}\right), \qquad (9)$$

where $\mathbf{\Gamma}$ is the covariance of the independent noise term approximating the quadrature error in (7). A similar use of quadrature is proposed in [9] applied to the integral operator (3) to allow for nonlinear transformation of the GP variables. No attempt is made to proxy for the quadrature error and it effectively gets absorbed into the

GP model. For our application the additive error may be viewed as a regularisation term to prevent singularities in the covariance matrix. Heuristically in the limit with $\Gamma = 0$ and $M \to \infty$ the covariance matrix can be represented as the outer product of the K eigenvectors of the discretised matrix operator $K[\mathbf{g}]$ with unit eigenvalues so that the resulting covariance matrix is singular.

If we specify a Gaussian initial distribution $p(\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_0 \mid \mathbf{0}, \Sigma_0)$ then carry out iterates of the map (7) up to some truncation order M we have an approximation to the distribution of a finite sample of a complete trajectory of (4) conditioned on a discrete realisation of the latent forces which is given by

$$\begin{aligned} p(\mathbf{x}_M \mid \mathbf{g}, \boldsymbol{\Gamma}) &= \int \cdots \int p(\mathbf{x}_M, \mathbf{x}_{M-1}, \dots, \mathbf{x}_0 \mid \mathbf{g}, \boldsymbol{\Gamma}) \mathrm{d}\mathbf{x}_0 \cdots \mathrm{d}\mathbf{x}_{M-1} \\ &= \int \cdots \int \prod_{m=1}^{M} p(\mathbf{x}_m \mid \mathbf{x}_{m-1}, \mathbf{g}, \boldsymbol{\Gamma}) p(\mathbf{x}_0) \mathrm{d}\mathbf{x}_0 \cdots \mathrm{d}\mathbf{x}_{M-1} \\ &= \mathcal{N}(\mathbf{x}_M \mid \mathbf{0}, \Sigma_M(\mathbf{g}, \boldsymbol{\Gamma})), \end{aligned} \tag{10}$$

where the covariance matrix $\Sigma_M(\mathbf{g}, \boldsymbol{\Gamma})$ is defined recursively by $\Sigma_0(\mathbf{g}, \boldsymbol{\Gamma}) = \Sigma_0$ and

$$\Sigma_m(\mathbf{g}, \boldsymbol{\Gamma}) = K[\mathbf{g}]\Sigma_{m-1}(\mathbf{g}, \boldsymbol{\Gamma})K[\mathbf{g}]^\top + \Gamma, \qquad m = 1, \dots, M, \tag{11}$$

and this model should then be viewed as a discretisation of the true marginal distribution with moments (6).

It is now possible to specify a complete joint distribution $p(\mathbf{x}, \mathbf{g})$ of the latent state and force variables by completing the likelihood term (10) with the prior on the latent force variable. On inspection of (8) we see that the entries of $K[\mathbf{g}]$ will be linear in the latent forces and so the entries of the covariance matrix (11) will be degree $2M$ polynomials in the latent forces and as such there is no analytical expression for the posterior conditional density for orders greater than one. Despite this it is straightforward to use sampling methods and gradient based approximations.

5 Simulation Study

Reasonably we would expect that by increasing the truncation order of the approximation introduced in the previous section we gain increasingly accurate approximations to the true conditional distribution and in this section we demonstrate that this is indeed the case by considering an exactly solvable model.

We demonstrate our method on the Kubo oscillator [7] which can be expressed by the ODE in $\mathbb{R}^2$ with a single latent force and evolution equation

$$\begin{bmatrix} \dot{x}(t) \\ \dot{y}(t) \end{bmatrix} = \begin{bmatrix} 0 & -g(t) \\ g(t) & 0 \end{bmatrix} \begin{bmatrix} x(t) \\ y(t) \end{bmatrix}, \tag{12}$$

which for $\mathbf{x}(t) = (x(t), y(t))^\top$ has solution given by

$$\mathbf{x}(t) = R\left[\int_0^t g(\tau)\mathrm{d}\tau\right]\mathbf{x}(t_0), \tag{13}$$

where $R[\theta]$ in (13) is the 2×2 matrix rotating a vector in $\mathbb{R}^2$ by θ-radians anticlockwise around the origin. It follows that given a set of data points $\mathcal{Y} = (\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_N)$ with $t_0 < t_1 < \cdots < t_N$ and zero measurement error that the values of $G_i := \int_{t_{i-1}}^{t_i} g(\tau)\mathrm{d}\tau$ are constrained to satisfy $\mathbf{x}_i = R[G_i]\mathbf{x}_{i-1}$, for $i = 1, \ldots, N$ which defines the vector $\mathbf{G} = (G_1, \ldots, G_N)^\top$ up to translation of each component by 2π, moreover since $\mathrm{Var}(G_i) = \mathcal{O}(|t_i - t_{i-1}|^2)$ we can consider only the component in $[-\pi, \pi]$ and approximate the true conditional distribution of $\mathbf{g} = (g(t_0), g(t_1), \ldots, g(t_N))^\top$ by the Gaussian distribution with density $p(\mathbf{g} \mid \mathbf{G} = \boldsymbol{\gamma})$ where $\boldsymbol{\gamma} \in [-\pi, \pi]^N$ with components satisfying $\mathbf{x}_i = R[\gamma_i]\mathbf{x}_{i-1}$ for each $i = 1, \ldots, N$.

While the distribution implied by the likelihood term (10) is not available in closed form, we can investigate the qualitative properties of the method introduced in Sect. 4 by considering the Laplace approximation. Using the Laplace approximation has the benefit of allowing us to carry out the comparison with the ground truth distribution using a proper metric on the space of distributions by considering the Wasserstein distance between two multivariate Gaussians [2].

The method of successive approximations fixes a point and is therefore local in character, as such we implement a simulation study that enables us to assess the performance of our approximation as the total interval length increases. We consider two methods of varying the interval length T; the first by fixing the sample size, N, and then varying the spacing between samples, Δt, and the second by fixing the sample frequency and varying the total number of observations. For each combination of sample size and frequency we perform 100 simulations of the Kubo oscillator (12) on the interval $[0, T]$ assuming a known *radial basis function* (RBF) kernel $k(t, t'; \psi) = \psi_0 \exp\{-(t - t')^2/2\psi_1^2\}$ with $\psi = (1, 1)^\top$ for the latent force. We consider interval lengths $T \in \{3, 6, 9\}$ and sample frequencies $\Delta t \in \{0.50, 0.75, 1.00\}$. This implies a sample size of $N = T/\Delta t + 1$ for each experiment and we use Simpson's quadrature rule so that the latent state vector is augmented to size $2N + 1$.

Our principal interest is in the impact of the truncation order, M, on the accuracy of our approximation and so for each simulated experiment we fit the model with orders $M = 3, 5, 7, 10$. The covariance of the initial approximation is formed by placing independent GP priors on the first and second components with RBF kernels $k(t, t'; \phi_k)$ and the parameters ϕ_k, $k = 1, 2$ are optimised during the fitting process. The regularisation matrix $\mathbf{\Gamma}$ is given by multiplying an appropriately sized identity matrix by a small scale parameter 0.0001 and this value is kept fixed.

The results of the experiment are displayed in Table 1. Along each row we observe that across all sampling specifications increasing the order of approximation leads to increasingly accurate approximations of the true distributions, and that this conclusion holds whether we vary the sample size or the sample frequency. Inspecting

Table 1 Comparison of the successive approximations MLFM introduced in Sect. 4 with the true distribution for the Kubo oscillator based on 100 simulations of the process on $[0, T]$ with $N = T/\Delta t + 1$ evenly spaced observations. Reported are the sample averages and standard errors of the Wasserstein distance between the Laplace approximation and the true conditional distribution

T	Δt	order = 3	order = 5	order = 7	order = 10
9	1.00	0.965 (0.477)	0.863 (0.573)	0.711 (0.672)	0.527 (0.632)
	0.75	0.983 (0.315)	0.874 (0.407)	0.762 (0.448)	0.584 (0.415)
	0.50	1.517 (0.556)	1.068 (0.450)	0.701 (0.227)	0.517 (0.225)
6	1.00	0.865 (0.606)	0.619 (0.503)	0.433 (0.475)	0.319 (0.412)
	0.75	0.738 (0.392)	0.629 (0.463)	0.513 (0.426)	0.328 (0.325)
	0.50	0.846 (0.256)	0.591 (0.194)	0.532 (0.234)	0.399 (0.192)
3	1.00	0.374 (0.311)	0.294 (0.384)	0.202 (0.256)	0.185 (0.211)
	0.75	0.421 (0.440)	0.272 (0.440)	0.136 (0.217)	0.076 (0.064)
	0.50	0.421 (0.190)	0.395 (0.289)	0.235 (0.132)	0.191 (0.051)

the columns we observe that for each order a decrease in the sampling interval T leads to a general increase in accuracy of the approximation with some variations with the sample size and frequency. The fact that within most blocks of fixed T and M that the distances are of a similar magnitude strongly suggests it is the size of the window T that is a larger determinant of the accuracy of the introduced approximation than the number of sample points or their frequency. In fact, we see that dense samples can lead to a slower convergence of the approximation and this is particularly pronounced for the row $T = 9$ and $\Delta t = 0.50$ which does a very poor job of approximating the true distribution at lower orders compared to the sparser samples, but eventually outperforms these methods as the approximation order increases.

6 Discussion

In this paper we have introduced the MLFM, a hybrid model which enables the embedding of prior geometric knowledge into statistical models of dynamic systems. By using the method of successive approximations we were able to motivate a family of truncated approximations to the joint distribution, and while the distribution is not available in closed form it is still amenable to sampling and gradient based methods. In future work we discuss variational methods formed by retaining the successive approximations rather than performing the marginalisation (10) and exploiting the interpretation of (9) as a linear Gaussian dynamical system in the truncation order.

The simulation study in Sect. 5 showed the method performs well over moderate sample windows with only a few orders of approximation, but that as the length of window over which a solution is sought increases the order required to achieve good performance increases. It may therefore be of interest to replace a single, high order, approximation with a collection of local methods of lower order. Combining these local models in a principled manner is the subject of ongoing work, nevertheless the results of Sect. 5 show that the method introduced in this paper can perform well, as well as being an important precursor to more involved methods.

Acknowledgements The authors would like to thank Dr. Durante and two anonymous referees for their valuable comments and suggestions on an earlier version of this paper, which substantially improved the content and clarity of the article. Daniel J. Tait is supported by an EPSRC studentship.

References

1. Alvarez, M., Luengo, D., Lawrence, N.D.: Latent force models. In: van Dyk, D., Welling, M. (eds.) Proc. Mach. Learn. Res. **5**, 9–16 (2009)
2. Dowson, D.C., Landau, B.V.: The Fréchet distance between multivariate normal distributions. J. Multivar. Anal. **12**, 450–455 (1982)
3. Gao, P., Honkela, A., Rattray, M., Lawrence, N.D.: Gaussian process modelling of latent chemical species: applications to inferring transcription factor activities. Bioinform. **24**, i70–i75 (2008)
4. Iserles, A., Norsett, S.P.: On the solution of linear differential equations in Lie groups. Philos. Trans. Roy. Soc. A, Math., Phys. and Eng. Sci. **357**, 983–1019 (1999)
5. Lawrence, N.D., Sanguinetti, G., Rattray, M.: Modelling transcriptional regulation using Gaussian processes. In: Schölkopf, B., Platt, J.C., Hofmann, T. (eds.) Adv. Neural Inf. Process. Syst. **19**, 785–792 (2007)
6. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. MIT Press, Cambridge (2006)
7. Risken, H.: The Fokker-Planck Equation: Methods of Solution and Applications. Springer, Berlin (1989)
8. Strand, J.L.: Random ordinary differential equations. J. Differ. Equ. **7**, 538–553 (1970)
9. Titsias M.K., Lawrence, N.D., Rattray, M.: Efficient sampling for Gaussian process inference using control variables. In: Koller, D., Schuurmans, D., Bengio, Y., Bottou, L. (eds.) Adv. Neural Inf. Process. Syst. **21**, 1681–1688 (2009)

Computational Statistics

particleMDI: A Julia Package for the Integrative Cluster Analysis of Multiple Datasets

Nathan Cunningham, Jim E. Griffin, David L. Wild and Anthony Lee

Abstract We present particleMDI, a Julia package for performing integrative cluster analysis on multiple heterogeneous data sets, built within the framework of multiple data integration (MDI). particleMDI updates cluster allocations using a particle Gibbs approach which offers better mixing of the MCMC chain—but at greater computational cost—than the original MDI algorithm. We outline approaches for improving computational performance, finding the potential for greater than an order-of-magnitude improvement. We demonstrate the capability of particleMDI to uncovering the ground truth in simulated and real datasets. All files are available at https://github.com/nathancunn/particleMDI.jl.

Keywords Bayesian inference · Cluster analysis · Computational statistics · Data integration · Particle Monte Carlo methods

1 Introduction

Cluster analysis is the task of inferring a latent group structure in data, such that observations within groups are, in some sense, 'closer' to one another than to observations in other groups. Standard methods, such as k-means, are not equipped

Nathan Cunningham's research is funded by EPSRC grant 1654596 and supported by The Alan Turing Institute under grant TU/D/000013. David Wild acknowledges support from EPSRC grant EP/R014337/1.

N. Cunningham (✉)
Department of Statistics, University of Warwick, Coventry CV4 7AL, England
e-mail: n.cunningham@warwick.ac.uk

J. E. Griffin
University College London, London, England

D. L. Wild
University of Warwick, Coventry, England

A. Lee
University of Bristol, Bristol, England

R. Argiento et al. (eds.), *Bayesian Statistics and New Generations*,
Springer Proceedings in Mathematics & Statistics 296,
https://doi.org/10.1007/978-3-030-30611-3_7

for cases where the units of observation have data arising from multiple sources. Integrating multiple data sources into a composite analysis is a key challenge in the analysis of genomic data where multiple heterogeneous datasets can give different—but complementary—views of the same underlying process. In this context, one may perform cluster analysis to infer risk cohorts among groups of patients for whom we have multiple biological data sets recorded. We introduce particleMDI, a package developed in the statistical programming language Julia [2] for performing integrative cluster analysis in this context. While many such approaches exist (see e.g. [7, 12, 13, 17]), particleMDI is built within the framework of multiple data integration (MDI) [11].

MDI facilitates integrative cluster analysis by allowing for the borrowing of information between datasets of potentially different types. Observations arise from a Dirichlet-multinomial allocation mixture model [9]—a finite approximation to the Dirichlet process mixture model. To infer dependence between the cluster structure of different datasets, Kirk et al. introduce a parameter, Φ, measuring the similarity between pairs of datasets at the level of the cluster allocations. The inferred value of, e.g., $\phi_{k,l}$ is used to inflate the probability of assigning observations in dataset k to the clusters they are assigned to in dataset l.

Inference in MDI is performed via a Gibbs sampler, alternating between updating cluster allocations and hyperparameters; full details are available in [11]. As conjugate priors are used in MDI, the cluster parameters can be analytically integrated over and individual cluster allocations are updated while holding all other allocations fixed [14]. A result of this one-at-a-time approach is that once MDI infers an allocation which is 'good enough', in some sense, it can be difficult to consider alternatives unless they are similar to this current allocation.

2 particleMDI

particleMDI extends the original MDI algorithm, replacing the one-at-a-time approach to clustering with a conditional particle filter, which has demonstrated good mixing properties even when the number of particles is relatively low [8]. This approach to cluster analysis (see [4, 5, 8]) infers a latent cluster allocation, $c_{i,k}$, for an observation, $x_{i,k}$, given observations $x_{1:i,k}$ and allocations $c_{1:(i-1),k}$, using a weighted cloud of approximations, termed particles. The particle approximation of the Gibbs sampler [1] uses a conditional sequential Monte Carlo (SMC) update, which uses a single particle, sampled appropriately from the particle filter, to update the hyperparameters. The trajectory of this 'reference particle' is held fixed through a subsequent pass of the conditional SMC update, thus guiding other particles towards relevant regions of the state space.

We use the parameter Φ to share information across datasets by inflating the weights of particles in which allocations agree across datasets, as detailed in Algorithm 1.

2.1 Improving Computational Performance

particleMDI is much more computationally costly than the original MDI algorithm. Fearnhead [5] discusses the inherent inefficiencies in particle Monte Carlo algorithms as applied to clustering algorithms: resampling and the discrete nature of the state-space mean it is likely some particles will be duplicates of others. Calculation of mutation weights—the weights for assigning an observation to each cluster—involves evaluating the posterior predictive of assigning an observation to each cluster. This step is wholly deterministic meaning identical particles will have identical mutation weights and, thus, there is no value in evaluating them more than once. To tackle this, we identify duplicated particles via the following ID assignments

$$ID_{i+1}^{(m)} = ID_i^{(m)} \times (M \times N) + c_{i+1}^{(m)}$$

where particle m assigns observation $i + 1$ to cluster $c_{i+1}^{(m)}$, M is the total number of particles, and N the maximum number of clusters.

It is also likely that where particles differ, they may share commonality, be it a shared subset of clusters, or even a shared partition up to a permutation of cluster labels. Again, this will involve redundant calculations evaluating posterior predictives multiple times for the same clusters. We adapt our algorithm so that each particle indexes into a global environment of clusters containing only a single copy of each unique cluster. We now need only evaluate posterior predictives once for each unique cluster and then combine these at the level of the particle to form the mutation weights.

A separate layer of inefficiency arises in the sequential nature of SMC methods. Evaluation of the posterior predictive of observation i conditional on the cluster allocations of observations $1 : (i - 1)$ is uninformative for very small values of i. To address this, we augment the particle Gibbs sampler such that we only update a predetermined number of cluster labels, holding $\lfloor n\rho \rfloor$ labels fixed for $0 < \rho < 1$. As the observations are exchangeable, we permute the observation indices according to a uniform permutation function σ and hold the first $\lfloor n\rho \rfloor$ cluster labels fixed from a previous pass of the conditional particle filter. The idea of updating blocks of sequential observations in the particle Gibbs sampler has previously been discussed [1] and a similar idea has been explored in the context of cluster analysis [3]. The permutation function, $\sigma(\cdot)$, is updated at every Gibbs iteration, ordering observation such that $\sigma(c)_{i,k}$ is the allocation for observation $\sigma(x)_{i,k}$ and $\sigma(x)_{i,l}$ corresponds to the same observational unit in a different data set. Therefore, where the standard particle Gibbs algorithm samples alternately from $p(\theta | x_{1:n}, c_{1:n})$ and $p_\theta(c_{1:n} | x_{1:n})$, our approach samples from $p(\theta | x_{1:n}, c_{1:n})$ and $p_\theta(\sigma(c)_{\lceil n\rho \rceil:n} | \sigma(x)_{1:n}, \sigma(c)_{1:\lfloor n\rho \rfloor})$. ($\theta$ here refers to the hyperparameters of the model, not the cluster parameters indicated in Fig. 1—as we use conjugate priors, the cluster parameters can be integrated out.) As well as giving the algorithm a 'warm start', this also avoids introducing a dependency between the inferred allocations and the order data are observed. Other approaches, such as that in [8], resolve this issue by instead updating all previous allocations during the resampling step. In a worst-case scenario—where resampling

Algorithm 1 particleMDI

```
Inputs:
  $\boldsymbol{\pi}$, $\boldsymbol{\Phi}$, cluster allocations $c^*_{1:n,1:K}$, a random permutation over observation
  indices, $\sigma(\cdot)$, and thresholds $\alpha$ and $\rho$ to control resampling and the portion of
  data conditioned-on, respectively
Initialize:
  Set particle weights $\xi^{(1)}, \dots, \xi^{(M)} = 1$
  Set $\sigma(c)^{(1)}_{1:n,1:K} = \sigma(c)^*_{1:n,1:k}, \sigma(c)^{(2:M)}_{1:\lfloor n\rho \rfloor,1:K} = \sigma(c)^*_{1:\lfloor n\rho \rfloor,1:K}$
for $i = \lceil n\rho \rceil, \dots, n$ do                          ▷ (iterate over remaining observations)
  for $m = 1, \dots, M$ do                                          ▷ (iterate over particles)
    for $k = 1, \dots, K$ do                                        ▷ (iterate over datasets)
      if $m \neq 1$ then                                            ▷ (particle 1 is the reference)
        Sample $\sigma(c)^{(m)}_{i,k}$                              ▷ (assign observation $\sigma(x)_{i,k}$ to a cluster)
        $q(\sigma(c)^{(m)}_{i,k} = a) \propto f(\sigma(x)_{i,k} | \sigma(c)^{(m)}_{i,k} = a) \times \pi_{a,k}$
      end if
    end for
    $\xi^{(m)} = \xi^{(m)} \times \prod_{k=1}^{K} \sum_{a=1}^{N} \pi_{a,k} f(\sigma(x)_{i,k} | \sigma(c)^{(m)}_{i,k} = a) \times \prod_{k=1}^{K-1} \prod_{l=k+1}^{K} (1 + \phi_{k,l} \mathbb{1}(\sigma(c)^{(m)}_{i,k} =$
$\sigma(c)^{(m)}_{i,l}))$                                           ▷ (Update particle weights accounting for agreement across datasets)
  end for
  Calculate effective sample size $(ESS) = \frac{(\sum_{m=1}^{M} \xi^{(m)})^2}{\sum_{m=1}^{M} \xi^{(m)2}}$.
  if ESS $< \alpha M$ then
    resample particles according to $\frac{\xi^{(m)}}{\sum_{m=1}^{M} \xi^{(m)}}$ and reset particle weights $\xi^{(1)}, \dots, \xi^{(M)} = 1$
  end if
end for
Select a final cluster allocation according to $\frac{\xi^{(m)}}{\sum_{m=1}^{M} \xi^{(m)}}$ and use to update $\boldsymbol{\pi}$, $\boldsymbol{\Phi}$ and use as $c^*_{1:n,1:K}$
and return to start.
```

is performed at every step—this would increase the complexity of the algorithm from $\mathcal{O}(n)$ to $\mathcal{O}(n^2)$, assuming the mutation weights can be computed in constant time. The choice of ρ warrants careful consideration as it imposes a trade-off between computation time and the mixing of the algorithm. However, where computation time is not a concern, lower values of ρ are not strictly to be preferred; setting ρ too low can result in too few conditioned-on observations to overcome the dependency in the observation order. We explore the impact of this in Sect. 4.

3 Using the `particleMDI` Package

The `pmdi()` function provides the primary functionality of the particleMDI algorithm. It takes the following inputs:

- `dataFiles` a vector of matrices containing the data to be clustered
- `dataTypes` a vector of types. For convenience, Gaussian and categorical data types are included and can be specified as `particleMDI.GaussianCluster`

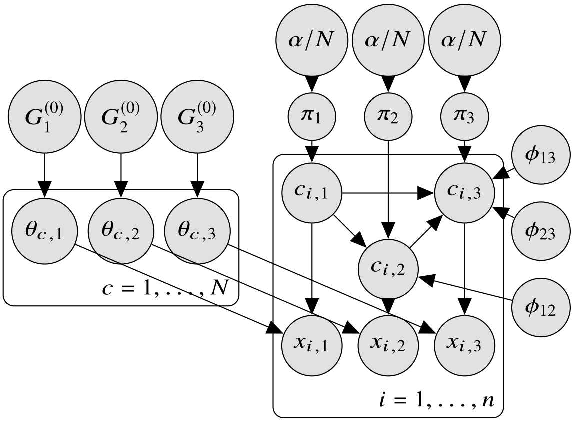


Fig. 1 A graphical model representation of MDI and particleMDI in a three dataset case ($K = 3$). $x_{i,k}$ denotes observation i in dataset k arising from cluster $c_{i,k}$ with parameters $\theta_{c_i,k}$, which are given a prior $G_k^{(0)}$. Cluster a in dataset k has prior allocation weight $\pi_{a,k}$ which is given a $Dirichlet(\alpha/N)$ prior. The $\phi_{i,j}$ values allow the allocations in data set i to inform those in data set j. (Figure recreated from [11])

and `particleMDI.CategoricalCluster` respectively. However, this can easily be extended to any other data type for which a posterior predictive can be specified, as detailed in Sect. 3.1

- `N` the maximum number of clusters to be inferred in each dataset
- `particles` an integer indicating the number of particles to use in the analysis
- ρ a value in (0, 1) indicating the proportion of the data whose allocations are held fixed at each iteration of the Gibbs sampler, as outlined in Sect. 2.1
- `iter` an integer specifying the number of iterations of the Gibbs sampler
- `outputFile` a string specifying the path of a .csv file in which to store the output

`pmdi()` outputs a .csv file, where each row contains the mass parameters, the phi values, and the allocations $c_{1:n,1:K}$. A user can assess this output file via some plotting functions built in the Julia library Gadfly [10]. In order to visualise the cluster allocations from multiple iterations of the Gibbs sampler, as well as across multiple datasets, `generate_psm()` and `consensus_map()`, can be used to visualise the posterior similarity matrices [13, 16] as heatmaps. That is, for each of K datasets, an $n \times n$ heatmap is constructed where element (i, j) reflects the frequency that observations i and j are assigned to the same cluster, as seen in Fig. 2. `plot_phimatrix()`, `plot_phichain()`, and `plot_nclust()` can each be useful tools for monitoring convergence of the Gibbs sampler, returning a heatmap of mean Φ values, a line graph of inferred Φ values, and the number of clusters inferred respectively.

3.1 Extending particleMDI for User-Defined Data Types

One of the strengths of the original MDI method is its ability to cluster a variety of different data types within a single analysis. While we provide the functionality for Gaussian and categorical data types, we take advantage of Julia's multiple dispatch capabilities to allow users to extend `particleMDI` to perform cluster analysis on other data types. As Julia code is just-in-time compiled, these user-specified data types do not suffer any penalty in terms of computation time. We illustrate this capability with a trivial example of assigning observations to clusters based on their sign.

We first create a cluster struct, a structure containing a single cluster and sufficient statistics for calculating the posterior predictive. In this case, we just need indicators of whether any observations belong to the cluster, as well as their sign.

```
mutable struct SignCluster
  n::Int64    # No. of observations in cluster
  isneg::Bool # Positive/negative flag
  SignCluster(dataFile) = new(0, false)
end
```

We then define `calc_logprob`, a function which returns the log posterior predictive of an observation, `obs`, given the observations assigned to cluster `cl`. It is important to specify `cl` as being of type `SignCluster`.

```
function particleMDI.calc_logprob(cl::SignCluster, obs)
    if cl.n == 0
        return log(0.5)
    else
        return ((obs[1] <= 0) == cl.isneg) ? 0 : - 10
    end
end
```

Finally, the function `cluster_add!` updates a cluster, `cl`, when an observation, `obs`, is added to it.

```
function particleMDI.cluster_add!(cl::SignCluster, obs)
    cl.n += 1
    cl.isneg = (obs[1] < 0)
end
```

We can now cluster univariate data into positive and negative clusters by passing `SignCluster` as a data type in `pmdi()`.

4 Application

We demonstrate particleMDI on three simulated Gaussian datasets, with cluster means $\mu_{\cdot,1} = [-0.5, 0, 0.75]$, $\mu_{\cdot,2} = [0, 0.75, -0.5]$, $\mu_{\cdot,3} = [0.75, -0.5, 0]$, where $\mu_{i,j}$ indicates the mean of observations belonging to cluster i in dataset j. We choose balanced clusters for clarity of illustrating results; analysis on other data suggests this does not unduly impact the results. All observations are drawn independently with standard deviation $\sigma = 1$ meaning we expect significant overlap across all clusters. Each data set has 150 observations with 16 dimensions, with 25% being noise. The analyses are run for 1000 iterations, with $M = 32$, and $\rho = 0.25$. The results, in Fig. 2, show that, by considering all three datasets simultaneously, particleMDI is able to recover the true underlying structure of the data.

Figure 3 shows the empirical effect on computation time as a function of observations, dimensions, and the number of particles used. Where the relevant parameters are not altered, particleMDI is run for $n = 1000$, $M = 32$, clustering two Gaussian and one categorical dataset with $n = 150$ observations with 16 dimensions. In all cases 25% of dimensions are drawn as random noise. All analyses were performed in Julia 0.6.4 on a Windows laptop with a 2.80 GHz Intel Core i7-7700HQ CPU and 32.0 GB RAM. We contrast the computation times between two implementations of the algorithm: one which benefits from the performance improvements obtained by exploiting the redundancy of the particle filter as outlined in Sect. 2.1; and one without these improvements. As we are only avoiding performing redundant calculations, these improvements do not come at the cost of any decrease in accuracy. Figure 3 shows we can improve computation time by more than an order of magnitude.

In order to assess the impact of ρ, we examine cluster accuracy from analysis on Fisher's iris dataset [6] for varying levels of ρ. We assess cluster accuracy by means of the adjusted Rand index [15]—a measure of agreement between two partitions adjusted for agreement by chance, a value of 1 indicating perfect agreement, and 0 indicating agreement no better than chance. As expected, very large ρ values lead to slow mixing of the Gibbs sampler, leading to many iterations before the

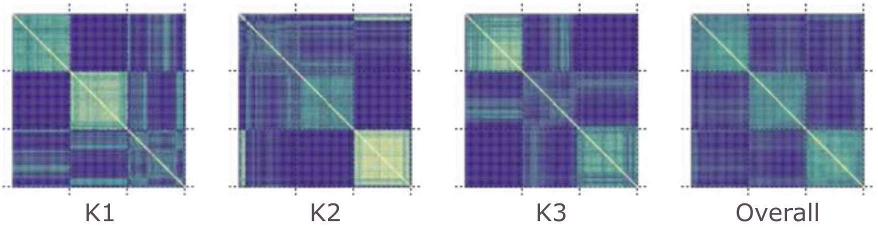


Fig. 2 Heatmap representation of the posterior similarity matrices as output from generate_psm() and consensus_map for three Gaussian datasets (K1, K2, K3) with different degrees of overlap in clusters across data sets. The brightness of point (i, j) in each reflects the empirical probability that observations i and j are clustered together in each dataset, with these values averaged across datasets to give the value in 'overall'

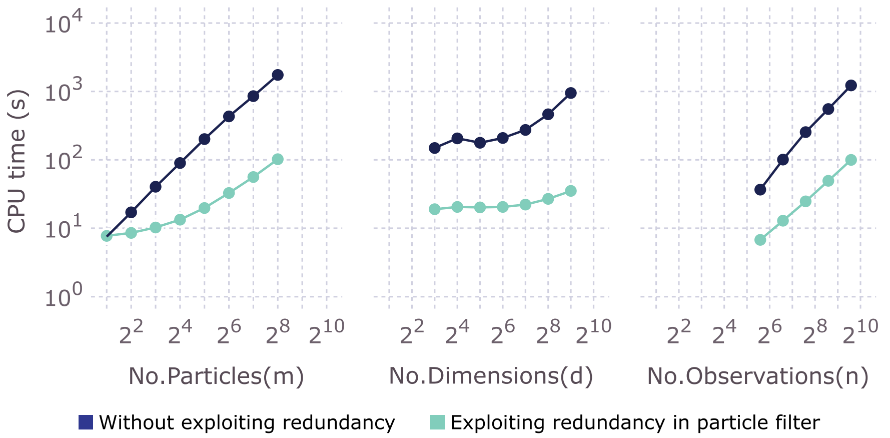
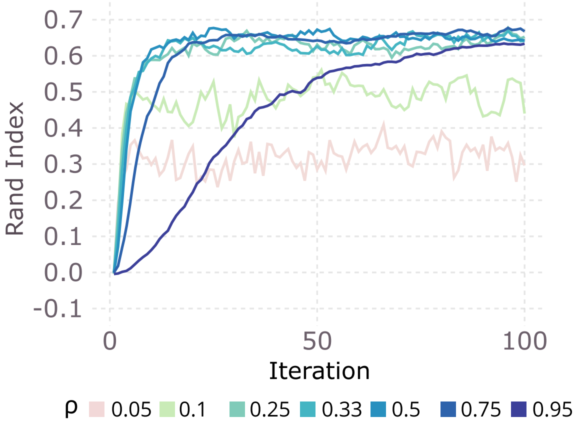


Fig. 3 Computation time as a function of the number of particles, the number of dimensions, and the number of observations respectively. The results show that reducing inefficient calculations can contribute to greater than an order of magnitude improvement in computation time

Fig. 4 The effect on cluster accuracy as a function of ρ from analyses on Fisher's iris dataset. Analyses were performed 10× and adjusted Rand index values per iteration were averaged across runs. The results suggest extreme values can negatively influence the output while there is little observable difference between thresholds in the range 0.25–0.5



algorithm converges. Interestingly though, very small values of ρ appear to be more problematic, with values of $\rho = 0.05$ and $\rho = 0.1$ struggling to get close to the ground truth. As discussed in Sect. 2.1 when ρ is very small, the conditional particle filter has little information on which to base allocations for observations it encounters at the beginning, inducing a strong dependence on the order of the observations (Fig. 4).

5 Discussion

In this paper, we have presented particleMDI, a Julia package implementing a particle Monte Carlo approach to the integrative cluster analysis of multiple data sets. We have demonstrated the capability of the package to uncover the ground truth cluster structure in a group of synthetic datasets of different data types. In Sect. 3.1 we showed how this package can perform cluster analysis on any data type for which a posterior predictive distribution can be specified. We outlined methods for improving computational performance of our algorithm in Sect. 2.1 and demonstrated that these approaches can achieve performance improvements of an order of magnitude or more in terms of computation time. While the context of our work is in integrative cluster analyses, these approaches are also applicable to the single-data context of cluster analysis using particle Monte Carlo methods.

All files relevant to this package are available on Github. (https://github.com/nathancunn/particleMDI.jl)

References

1. Andrieu, C., Doucet, A., Holenstein, R.: Particle Markov chain Monte Carlo methods. J. R. Stat. Soc.: Ser. B (Stat. Methodol.) **72**(3), 269–342 (2010)
2. Bezanson, J., Edelman, A., Karpinski, S., Shah, V.: Julia: a fresh approach to numerical computing. SIAM Rev. **59**(1), 65–98 (2017). https://doi.org/10.1137/141000671
3. Bouchard-Côté, A., Doucet, A., Roth, A.: Particle Gibbs split-merge sampling for Bayesian inference in mixture models. J. Mach. Learn. Res. **18**(28), 1–39 (2017)
4. Chopin, N.: A sequential particle filter method for static models. Biometrika **89**(3), 539–552 (2002)
5. Fearnhead, P.: Particle filters for mixture models with an unknown number of components. Stat. Comput. **14**(1), 11–21 (2004)
6. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Ann. Eugen. **7**(2), 179–188 (1936)
7. Gabasova, E., Reid, J., Wernisch, L.: Clusternomics: integrative context-dependent clustering for heterogeneous datasets. PLoS Comput. Biol. **13**(10), e1005781 (2017)
8. Griffin, J.: Sequential Monte Carlo methods for mixtures with normalized random measures with independent increments priors. Stat. Comput. **27**(1), 131–145 (2017)
9. Ishwaran, H., Zarepour, M.: Exact and approximate sum representations for the Dirichlet process. Can. J. Stat. **30**(2), 269–283 (2002)
10. Jones, D.C., Arthur, B., Nagy, T., Gowda, S., Godisemo, Holy, T., Noack, A., Sengupta, A., Darakananda, D., Mattriks, Leblanc, S., Dunning, I., Fischer, K., Chudzicki, D., Yu, Y., Breloff, T., Kleinschmidt, D., Mellnik, A., john verzani, inkyu, Innes, M.J., Huchette, J., Bauman, M., Buzby, K., Hyatt, K., Forsyth, J., Borje, G., Saba, E., Coalson, C., Pelenitsyn, A.: Giovineitalia/gadfly.jl: v0.7.0 (2018). https://doi.org/10.5281/zenodo.1284282
11. Kirk, P., Griffin, J.E., Savage, R.S., Ghahramani, Z., Wild, D.L.: Bayesian correlated clustering to integrate multiple datasets. Bioinformatics **28**(24), 3290–3297 (2012)
12. McParland, D., Gormley, I.C., McCormick, T.H., Clark, S.J., Kabudula, C.W., Collinson, M.A.: Clustering South African households based on their asset status using latent variable models. Ann. Appl. Stat. **8**(2), 747–776 (2014)

13. Monti, S., Tamayo, P., Mesirov, J., Golub, T.: Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Mach. Learn. **52**(1–2), 91–118 (2003)
14. Neal, R.M.: Markov chain sampling methods for Dirichlet process mixture models. J. Comput. Graph. Stat. **9**(2), 249–265 (2000)
15. Rand, W.M.: Objective criteria for the evaluation of clustering methods. J. Am. Stat. Assoc. **66**(336), 846–850 (1971). https://doi.org/10.1080/01621459.1971.10482356; https://www.tandfonline.com/doi/abs/10.1080/01621459.1971.10482356
16. Rasmussen, C., de la Cruz, B., Ghahramani, Z., Wild, D.: Modeling and visualizing uncertainty in gene expression clusters using Dirichlet process mixtures. IEEE/ACM Trans. Comput. Biol. Bioinform. **6**(4), 615–628 (2009)
17. Shen, R., Olshen, A.B., Ladanyi, M.: Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics **25**(22), 2906–2912 (2009)

Approaches Toward the Bayesian Estimation of the Stochastic Volatility Model with Leverage

Darjus Hosszejni and Gregor Kastner

Abstract The sampling efficiency of MCMC methods in Bayesian inference for stochastic volatility (SV) models is known to highly depend on the actual parameter values, and the effectiveness of samplers based on different parameterizations varies significantly. We derive novel algorithms for the centered and the non-centered parameterizations of the practically highly relevant SV model with leverage, where the return process and innovations of the volatility process are allowed to correlate. Moreover, based on the idea of ancillarity-sufficiency interweaving (ASIS), we combine the resulting samplers in order to guarantee stable sampling efficiency irrespective of the baseline parameterization. We carry out an extensive comparison to already existing sampling methods for this model using simulated as well as real world data.

Keywords Ancillarity-sufficiency interweaving strategy (ASIS) · Auxiliary mixture sampling · Bayesian inference · Markov chain Monte Carlo (MCMC) · State-space model

1 Introduction and Model Specification

Stochastic volatility (SV) models [19] are an increasingly popular choice for modeling financial return data. The basic SV model assumes an autoregressive structure for the log-volatility, and it is able to match the empirically observable low serial autocorrelation in the return series but high serial autocorrelation in the squared return series. The SV model with leverage (SVL) [6] extends the SV model by allowing the return series and the increment of the log-volatility series to correlate. This

D. Hosszejni (✉) · G. Kastner
Department of Finance, Accounting and Statistics, Institute for Statistics and Mathematics, WU Vienna University of Economics and Business, Welthandelsplatz 1/D4/level 4, Vienna, Austria
e-mail: darjus.hosszejni@wu.ac.at

G. Kastner
e-mail: gregor.kastner@wu.ac.at

R. Argiento et al. (eds.), *Bayesian Statistics and New Generations*,
Springer Proceedings in Mathematics & Statistics 296,
https://doi.org/10.1007/978-3-030-30611-3_8

correlation models a real world phenomenon, the asymmetric relationship between returns and their volatility.

SVL, in its centered parameterization (C), is typically formulated as

$$\begin{aligned} y_t &= \exp(h_t/2)\varepsilon_t, \\ h_{t+1} &= \mu + \varphi(h_t - \mu) + \sigma\eta_t, \\ \operatorname{cor}(\varepsilon_t, \eta_t) &= \rho, \end{aligned} \tag{1}$$

for $t = 1, \ldots, T$, where $\varepsilon_t, \eta_t \sim$ i.i.d. $\mathcal{N}(0, 1)$. The only observed variable is $\mathbf{y} = (y_1, \ldots, y_T)'$, usually some de-meaned financial return series. An AR(1) structure is assumed for the latent log variance $\mathbf{h} = (h_1, \ldots, h_T)'$, with mean μ, persistence φ, and increment volatility σ. The leverage effect is captured by ρ, which is zero in the basic SV model by assumption.

An equivalent specification, called the non-centered parameterization (NC), can be obtained by substituting $\tilde{h}_t = (h_t - \mu)/\sigma$ into (1), thereby moving μ and σ from the state equation to the observation equation. The resulting formulation is given by

$$\begin{aligned} y_t &= \exp((\mu + \sigma\tilde{h}_t)/2)\varepsilon_t, \\ \tilde{h}_{t+1} &= \varphi\tilde{h}_t + \eta_t. \end{aligned} \tag{2}$$

Common priors are chosen from the literature: $(\varphi + 1)/2 \sim \text{Beta}(a_\varphi, b_\varphi)$, $(\rho + 1)/2 \sim \text{Beta}(a_\rho, b_\rho), \sigma^2 \sim \text{Gamma}(\alpha_\sigma, \beta_\sigma), \mu \sim \mathcal{N}(\mu_\mu, \sigma_\mu^2), h_1 \sim \mathcal{N}(\mu, \sigma^2/(1 - \varphi^2))$ [5, 9, 15].

While the SV model is accessible through the `R` [18] package `stochvol` [8], it does not cater for the leverage effect, and, to the best of our knowledge, there is no implementation of SVL that works out-of-the-box in a free, open source environment.[1] Our goal is to extend the package with an easy-to-use MCMC sampler that performs reasonably well on a diverse variety of data sets. To this end, we compare various sampling algorithms through a large simulation study from a practical viewpoint. In doing so, it is important to note that `stochvol` is often used as a subsampler for hierarchical models such as (vector auto) regressions or multivariate (factor) SV models. Consequently, in order to use the extended package in a similar manner, adaptive algorithms are not preferred, as their adaptation state can be cumbersome to implement within a larger MCMC scheme.

2 Estimation Strategies

The state-of-the-art solution [15] for estimating $\mathbf{h}$ is based on linearizing the observation equation in (2), and employing a ten-component bivariate Gaussian mixture approximation to the joint law of $(\log\varepsilon_t^2, \eta_t)$ separately for each time point, thus

[1] Editorial note: as of February 1, 2019, `stochvol` handles leverage.

introducing a new array of latent variables $s_t \in \{1, \ldots, 10\}$, $t = 1, \ldots, T$, encoding the mixture components. The resulting conditionally Gaussian state space can be written as

$$\begin{aligned} y_t^* &= \mu + \sigma \tilde{h}_t + m_{s_t}^{(1)} + v_{s_t}^{(1)} w_t, \\ \tilde{h}_{t+1} &= \varphi \tilde{h}_t + \sqrt{1-\rho^2} z_t + d_t \rho \left(m_{s_t}^{(2)} + v_{s_t}^{(2)} w_t \right), \end{aligned} \tag{3}$$

where $y_t^* = \log(y_t^2)$, $d_t = \operatorname{sgn}(y_t)$, w_t, $z_t \sim$ i.i.d. $\mathcal{N}(0, 1)$ for $t = 1, \ldots, T$, and $m_j^{(i)}$, $v_j^{(i)}$ are model-independent constants for $i = 1, 2$, and $j = 1, \ldots, 10$, defined in [15].

Let $\boldsymbol{\vartheta} = (\varphi, \rho, \sigma, \mu)'$, and $\mathbf{s} = (s_1, \ldots, s_T)'$. The sampling algorithm of the auxiliary model (AUX) consists of repeating the steps below.

- Step 1: Draw $\mathbf{s} \mid \mathbf{y}, \mathbf{h}, \boldsymbol{\vartheta}$ using inverse transform sampling with the posterior probabilities calculated as in Sect. 2.3.2 of [15].
- Step 2: Draw $\varphi, \rho, \sigma \mid \mathbf{y}, \mathbf{s}$ via an independent Metropolis-Hastings (MH) step utilizing the Laplace approximation of the collapsed distribution of $\varphi, \rho, \sigma \mid \mathbf{y}, \mathbf{s}$ as the proposal. The calculation of the acceptance ratio includes Kalman filter evaluations, numerical optimization, and numerical differentiation.
- Step 3: Draw $\mu \mid \mathbf{y}, \mathbf{s}, \varphi, \rho, \sigma$, and then $\mathbf{h} \mid \mathbf{y}, \mathbf{s}, \boldsymbol{\vartheta}$, using Gaussian simulation smoothing [2, 4].

At least three issues arise with this method. First, due to the involvement of Kalman filter evaluations and the numerical optimization part in Step 2, the execution time of the sampler is significantly worse than the runtime of methods with more naïve proposals, e.g. MH algorithms based on sampling from the full conditional distribution. According to our measurements, Step 2 requires around 80% of the total runtime. Second, for extreme data sets, the sampler might get stuck in a state and be unable to accept a new state for many iterations. Third, and finally, the numerical optimization step is sensitive to its configuration, possibly returning a negative semi-definite Hessian matrix at the found mode.

Hence, for parameter sampling, we replace Step 2 by a random-walk MH (RWMH) method which estimates (1) or (2) without resorting to the auxiliary mixture approximation. For the latent vector, we again use the highly efficient Step 3 of AUX as a proposal, followed by an MH acceptance-rejection step to correct for the difference between models (1) and (3).

As already shown for SV [9, 13], samplers based on different parameterizations can have substantially different sampling efficiencies on the same data set due to the altered dependence structure. To exploit this phenomenon, the ancillarity-sufficiency interweaving strategy (ASIS) [21] can utilize samplers of both C and NC, and thus ASIS may be able to deliver a markedly higher effective sample size than samplers based on a single parameterization only.

The RWMH sampling algorithm estimates SVL by repeating the steps below.

- Step 1: Draw $\mathbf{h} \mid \mathbf{y}, \boldsymbol{\vartheta}$. A candidate $\mathbf{h}^*$ is proposed using the AUX sampler by drawing $\mathbf{s} \mid \mathbf{y}, \mathbf{h}, \boldsymbol{\vartheta}$ and then drawing $\mathbf{h} \mid \mathbf{y}, \mathbf{s}, \boldsymbol{\vartheta}$ as explained in Steps 1 and 3 of algorithm AUX. Subsequently, $\mathbf{h}^*$ is accepted with probability

$$\min\left\{1, \frac{p_{\mathrm{C}}\left(\mathbf{h}^* \mid \mathbf{y}, \boldsymbol{\vartheta}\right)}{p_{\mathrm{C}}\left(\mathbf{h} \mid \mathbf{y}, \boldsymbol{\vartheta}\right)} \frac{p_{\mathrm{A}}\left(\mathbf{h} \mid \mathbf{y}, \boldsymbol{\vartheta}\right)}{p_{\mathrm{A}}\left(\mathbf{h}^* \mid \mathbf{y}, \boldsymbol{\vartheta}\right)}\right\},$$

where p_C and p_A denote the corresponding posteriors resulting from specifications (1) and (3), respectively.

- Step 2: Draw $\boldsymbol{\vartheta} \mid \mathbf{y}, \mathbf{h}$. In order to avoid a possibly problematic truncation of the proposal distribution, the parameter vector $\boldsymbol{\vartheta}$ is transformed from $(-1, 1) \times (-1, 1) \times (0, \infty) \times \mathbb{R}$ to $\mathbb{R}^4$ by applying the transformation $x \mapsto 0.5 \log((1 + x)/(1 - x))$ to φ and ρ, and by taking the logarithm of σ^2. Then, in the resulting unbounded space, a simple four-dimensional Gaussian random walk is proposed. Its innovation covariance matrix elements are fixed at 0.1 on the diagonal and zero elsewhere.
- If ASIS is applied, then, after Step 2, $\tilde{\mathbf{h}} = (\tilde{h}_1, \dots, \tilde{h}_T)'$ is calculated using the new values of σ and μ, followed by a new draw from $\boldsymbol{\vartheta} \mid \mathbf{y}, \tilde{\mathbf{h}}$. Finally, in order to move back to the original parameterization, $\mathbf{h}$ is recalculated from $\tilde{\mathbf{h}}$ and the new values of σ and μ.

ASIS is a natural extension to the RWMH samplers for the centered and the non-centered parameterizations. However, in the case of AUX, resampling in a different parameterization is detrimental to sampling efficiency for two reasons. First, in Step 2, the parameters φ, ρ and σ are drawn from a collapsed distribution that is independent of $\mathbf{h}$. Consequently, ASIS provides only negligible gains in sampling efficiency. Second, if ASIS were applied to AUX, the computationally most expensive parts of Step 2 would be repeated, thereby increasing the execution time by around 80%.

3 Simulation Study

In order to assess the efficiency of our estimation algorithms for the parameter vector $\boldsymbol{\vartheta}$, we simulate data using SVL from an extensive grid of data generating processes (DGPs). The parameters $\varphi_{\text{true}}, \rho_{\text{true}}, \sigma_{\text{true}}$ vary on a $\{0, 0.5, 0.9, 0.95, 0.99\} \times \{-0.6, -0.3, 0, 0.3, 0.6\} \times \{0.1, 0.3, 0.5\}$ grid. For the sake of readability, μ_{true} is set to -9 in all cases, resulting in 75 distinct parameter settings. This choice covers previously investigated ranges [7, 9]. After the burn-in, respectively, adaptation phase, 50 000 MCMC draws are obtained from the posterior distribution. We repeat this exercise for ten data sets of length 300, and ten data sets of length 3000, for eight sampling algorithms: AUX, Stan-C, Stan-N, JAGS-C, JAGS-N, RWMH-C, RWMH-N, and RWMH-ASISx5, where C and N stand for the centered and, respectively, non-centered parameterization, while ASISx5 denotes the algorithm repeating the two steps of ASIS five times after each draw of $\mathbf{h}$, which in general we found to be superior to executing the two ASIS steps only once. Note, that, although they do not fit our needs due to their adaptation phase, we include Stan [1] and JAGS [16] as benchmarks, and all reported results are based on the chain after adaptation has stopped. We fix the priors throughout the simulation study to $a_\varphi = 20$, $b_\varphi = 1.5$, $a_\rho = 3$, $b_\rho = 6$, $\alpha_\sigma = 0.5$, $\beta_\sigma = 0.5$, $\mu_\mu = -10$, and $\sigma^2_\mu = 100$. The prior

Table 1 Typical execution times (in min) for 50 000 draws after the burn-in when $T = 3000$. The displayed values correspond to the first and the ninth deciles of all wall clock times. The choice of the parameterization affects the execution time when JAGS or Stan is used and thus in these cases runtimes are shown separately for C and NC

Stan-C	Stan-N	JAGS-C	JAGS-N	RWMH	RWMH-ASISx5	AUX
90–642	59–441	22–31	50–106	6–21	14–29	44–86

hyperparameters of φ, σ^2, and μ are chosen from previous studies [9, 14], and the slightly informative prior on ρ is chosen to improve the estimation process of Stan-C and of AUX in the extremes of the parameter grid. However, results not reported here due to space constraints indicate that when $T = 300$, the posterior distribution of ρ is only mildly affected by this choice compared to a uniform prior, whereas when $T = 3000$ the differences are barely noticeable.

The resulting 12 000 MCMC chains were computed on a cluster of computers consisting of 400 Intel E5 2.3 GHz cores running `R` version 3.4.3. The Stan and the JAGS models were estimated using `rstan` [1] version 2.17.3 and `rjags` [16] version 4-6. The RWMH samplers and AUX were based on our computationally efficient `Rcpp` [3] implementation. The typical runtime of the samplers is summarized in Table 1. Inefficiency factors and effective sample sizes were calculated using the `coda` [17] package, data analysis and visualization was done with the help of the `tidyverse` [20] package.

We assess the statistical efficiency of the different competitors through their inefficiency factor (IF), an estimator for the integrated autocorrelation time τ, given by $\tau = 1 + 2\sum_{t=1}^{\infty} \rho_{\text{auto}}(t)$, where $\rho_{\text{auto}}(t)$ denotes the autocorrelation function at lag t. For an MCMC sample S, the IFs reported here are calculated as $\text{IF}_S = n_S/\text{ESS}_S$ [9], where n_S is the size of S, and ESS_S stands for the effective sample size of S, the size of a serially uncorrelated sample having the same Monte Carlo standard error as S. A good sampler has low serial correlation, thus the aim is to provide samples with low IF, or, in other words, high ESS. In practice, computational speed is comparably important to computational efficiency. Hence, the final assessment is based on the effective sampling rate (ESR), defined as the ESS divided by the execution time. We note that incorporating runtime in the assessment of algorithms may be problematic due to inconsistent implementations [11]; however, as one of our objectives is a software package, we consider the computational speed an essential part of our study.

3.1 Collapsed Versus Full Conditional Sampling

AUX employs a well-known technique for improving the statistical efficiency of MCMC simulations by using a collapsed distribution for sampling φ, ρ, and σ. This means that some variables are marginalized out in order to decrease the serial

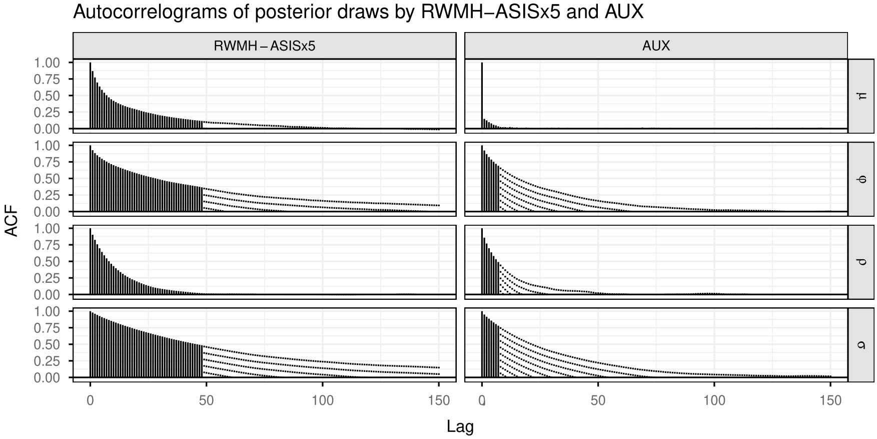


Fig. 1 Autocorrelation functions of the posterior draws for μ, φ, ρ, and σ, using RWMH-ASISx5 and AUX, for an illustrative example where $\varphi_{\text{true}} = 0.95$, $\rho_{\text{true}} = -0.3$, $\sigma_{\text{true}} = 0.3$, and $T = 300$. The line type indicates the speed of the Monte Carlo simulation: the number of solid lines equals the average number of samples drawn in 0.1 s

dependence in the chain [12]. ASIS, on the other hand, takes advantage of being able to reorganize the dependence. Which technique is superior in practice largely depends on computational aspects. Figure 1 exemplifies the problem by displaying the autocorrelograms of the outputs of RWMH-ASISx5 and AUX, for the parameters μ, φ, ρ, and σ, based on a selected data set. The figure illustrates the execution time as well: in both columns, the number of solid lines indicates the average number of samples drawn in 0.1 s. Thus, in each facet, the height of the rightmost solid line visualizes the ESR for the given parameter and sampler. Although the autocorrelation functions of AUX decay faster than the ones of RWMH-ASISx5, the latter counterbalances its disadvantages by its speed. Note, however, that different DGPs tend to produce qualitatively different pictures, making the choice between AUX and RWMH-ASISx5 non-trivial.

3.2 Efficiency Overview

The minimal ESR is the minimum taken over the ESRs of φ, ρ, σ, and μ, and, thus, it measures the speed of discovering the joint posterior $p(\boldsymbol{\vartheta} \mid \mathbf{y})$. In order to provide an overview, Fig. 2 displays the minimal ESRs for each sampler and strategy, and for all DGPs with $\rho_{\text{true}} = -0.3$ and $T = 3000$. Taking into account that Stan and JAGS are general-purpose probabilistic modeling frameworks, they perform surprisingly well compared to AUX and our RWMH implementations which have been developed specifically for the model at hand. However, the absence of a best performing method is eye-catching. In particular, the choice between AUX and RWMH is noticeably difficult.

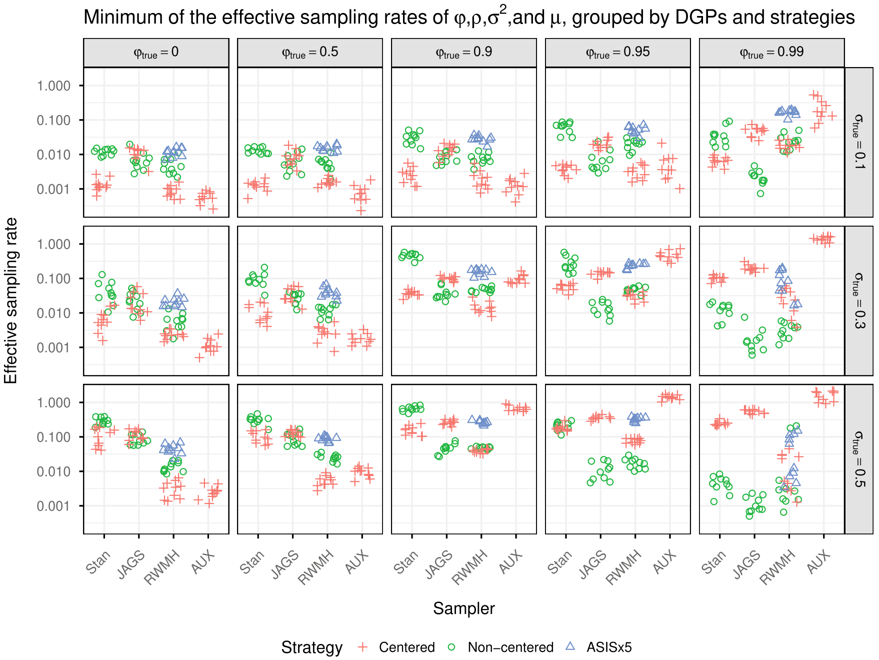


Fig. 2 Minimal effective sampling rates of all the examined samplers and strategies, for the whole range of φ_{true} and σ_{true} values, while, for the sake of readability, ρ_{true} is set to -0.3, and T to 3000. In each facet, there are 10 data points plotted for each sampler and strategy, corresponding to the 10 repetitions of the simulation exercise. A small horizontal noise has been applied to the position of the points to increase their visibility

In terms of variability, note that the ESRs of AUX range from below 0.001 to above 1, while the ESRs of RWMH-ASISx5 fall between 0.01 and 0.1. This renders the latter more stable by around two orders of magnitude.

4 Application to Financial Data Sets

We apply the eight estimation methods to seven univariate time series of daily financial log-returns covering four asset types and two economic periods. The first time interval is a booming, pre-crisis period starting from 2005-01-01 and ending on 2007-12-31, including a total of 872 business days. The second interval is a more recent, more volatile period between 2015-01-01 and 2018-06-30, including 1014 business days. The series under consideration are the Bitcoin price in USD (ticker: BTCUSD=X), hereafter BTC, the German DAX index (ticker: ˆGDAXI), the EUR/USD exchange rate (ticker: EURUSD=X), hereafter EUR, and a large German company, the Merck KG's equity (ticker: MRK.DE), hereafter MRK. The data is provided by Yahoo! Finance.

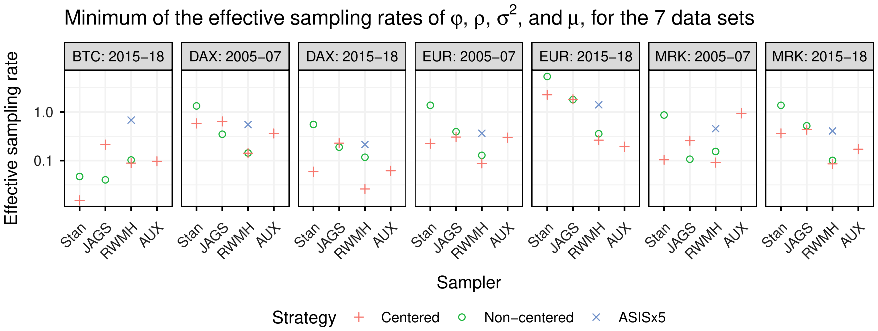


Fig. 3 Minimal effective sampling rate for seven data sets. Facets correspond to time series, where "2005–07" and "2015–18" denote the first and the second time period, respectively. In each facet, each point corresponds to a certain sampler and strategy

Figure 3 summarizes the results of the exercise, carried out under the same prior specification as in Sect. 3. It is interesting to note that Stan generally shows high ESRs with the only exception of BTC where RWMH-ASISx5 excels. Focusing on the comparison of RWMH and AUX it stands out that without interweaving, AUX is generally to be preferred, whereas RWMH-ASISx5 tends to outperform AUX in all scenarios but one. The overall picture is similar to Fig. 2, as there is no single algorithm that dominates on all data sets.

5 Conclusion

The paper at hand contributes to the literature on MCMC sampling algorithms by investigating the efficiency of several competing methods for the stochastic volatility model with leverage. We derived an RWMH sampler and improved it through ASIS and an efficient latent state sampler. Moreover, we carried out a computational experiment to compare our novel method to the state-of-the-art approach, an auxiliary mixture sampler, and to Stan and JAGS implementations as benchmarks. Based on our results, we conclude that employing our boosted naïve estimator for the latent space stabilizes the effective sampling rate of the algorithm by avoiding numerical optimization and differentiation.

Current research is directed towards further financial applications including factor models [10], and further extending the R package stochvol.

References

1. Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A.: Stan: a probabilistic programming language. J. Stat. Softw. **76**(1), 1–32 (2017). https://doi.org/10.18637/jss.v076.i01

2. Carter, C.K., Kohn, R.: On Gibbs sampling for state space models. Biometrika **81**(3), 541–553 (1994). https://doi.org/10.1093/biomet/81.3.541
3. Eddelbuettel, D., François, R.: Rcpp: Seamless R and C++ integration. J. Stat. Softw. **40**(8), 1–18 (2011). https://doi.org/10.18637/jss.v040.i08
4. Frühwirth-Schnatter, S.: Data augmentation and dynamic linear models. J. Time Ser. Anal. **15**(2), 183–202 (1994). https://doi.org/10.1111/j.1467-9892.1994.tb00184.x
5. Frühwirth-Schnatter, S., Wagner, H.: Stochastic model specification search for Gaussian and partial non-Gaussian state space models. J. Econ. **154**(1), 85–100 (2010). https://doi.org/10.1016/j.jeconom.2009.07.003
6. Harvey, A.C., Shephard, N.: Estimation of an asymmetric stochastic volatility model for asset returns. J. Bus. Econ. Stat. **14**(4), 429–434 (1996). https://doi.org/10.1080/07350015.1996.10524672
7. Jacquier, E., Polson, N.G., Rossi, P.E.: Bayesian analysis of stochastic volatility models with fat-tails and correlated errors. J. Econ. **122**(1), 185–212 (2004). https://doi.org/10.1016/j.jeconom.2003.09.001
8. Kastner, G.: Dealing with stochastic volatility in time series using the R package stochvol. J. Stat. Softw. **69**(5), 1–30 (2016). https://doi.org/10.18637/jss.v069.i05
9. Kastner, G., Frühwirth-Schnatter, S.: Ancillarity-sufficiency interweaving strategy (ASIS) for boosting MCMC estimation of stochastic volatility models. Comput. Stat. Data Anal. **76**, 408–423 (2014). https://doi.org/10.1016/j.csda.2013.01.002
10. Kastner, G., Frühwirth-Schnatter, S., Lopes, H.F.: Efficient Bayesian inference for multivariate factor stochastic volatility models. J. Comput. Graph. Stat. **26**(4), 905–917 (2017). https://doi.org/10.1080/10618600.2017.1322091
11. Kriegel, H.P., Schubert, E., Zimek, A.: The (black) art of runtime evaluation: are we comparing algorithms or implementations? Knowl. Inf. Syst. **52**, 341–378 (2016). https://doi.org/10.1007/s10115-016-1004-2
12. Liu, J.S.: The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. J. Am. Stat. Assoc. **89**(427), 958–966 (1994). https://doi.org/10.1080/01621459.1994.10476829
13. McCausland, W.J., Miller, S., Pelletier, D.: Simulation smoothing for state-space models: a computational efficiency analysis. Comput. Stat. Data Anal. **55**(1), 199–212 (2011). https://doi.org/10.1016/j.csda.2010.07.009
14. Nakajima, J., Omori, Y.: Leverage, heavy-tails and correlated jumps in stochastic volatility models. Comput. Stat. Data Anal. **53**(6), 2335–2353 (2009). https://doi.org/10.1016/j.csda.2008.03.015
15. Omori, Y., Chib, S., Shephard, N., Nakajima, J.: Stochastic volatility with leverage: fast and efficient likelihood inference. J. Econ. **140**(2), 425–449 (2007). https://doi.org/10.1016/j.jeconom.2006.07.008
16. Plummer, M.: JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In: Hornik, K., Leisch, F., Zeileis, A. (eds.) Proceedings of the 3rd International Workshop on Distributed Statistical Computing (2003)
17. Plummer, M., Best, N., Cowles, K., Vines, K.: CODA: convergence diagnosis and output analysis for MCMC. R News **6**(1), 7–11 (2006)
18. R Core Team: R: a language and environment for statistical computing. R Foundation for Statistical Computing (2018). https://www.R-project.org/
19. Taylor, S.J.: Financial returns modeled by the product of two stochastic processes: a study of daily sugar prices 1961–75. Time Series Analysis, Theory and Practice, pp. 203–226. North-Holland, Amsterdam (1982)
20. Wickham, H.: tidyverse: Easily Install and Load 'tidyverse' Packages (2017). https://CRAN.R-project.org/package=tidyverse. R package version 1.1.1
21. Yu, Y., Meng, X.L.: To center or not to center: that is not the question–an ancillarity-sufficiency interweaving strategy (ASIS) for boosting MCMC efficiency. J. Comput. Graph. Stat. **20**(3), 531–570 (2011). https://doi.org/10.1198/jcgs.2011.203main

Efficient Metropolis-Hastings Sampling for Nonlinear Mixed Effects Models

Belhal Karimi and Marc Lavielle

Abstract The ability to generate samples of the random effects from their conditional distributions is fundamental for inference in mixed effects models. Random walk Metropolis is widely used to conduct such sampling, but such a method can converge slowly for medium dimension problems, or when the joint structure of the distributions to sample is complex. We propose a Metropolis–Hastings (MH) algorithm based on a multidimensional Gaussian proposal that takes into account the joint conditional distribution of the random effects and does not require any tuning, in contrast with more sophisticated samplers such as the Metropolis Adjusted Langevin Algorithm or the No-U-Turn Sampler that involve costly tuning runs or intensive computation. Indeed, this distribution is automatically obtained thanks to a Laplace approximation of the original model. We show that such approximation is equivalent to linearizing the model in the case of continuous data. Numerical experiments based on real data highlight the very good performances of the proposed method for continuous data model.

Keywords Nonlinear · MCMC · Metropolis · Mixed effects · Sampling

1 Introduction

Mixed effects models are reference models when the inter-individual variability that can exist within the same population is considered (see [9] and the references therein). Given a population of individuals, the probability distribution of the series of observations for each individual depends on a vector of individual parameters. For complex priors on these individual parameters or models, Monte Carlo methods must be used to approximate the conditional distribution of the individual parameters

B. Karimi (✉) · M. Lavielle
Inria, Paris, France
e-mail: belhal.karimi@inria.fr

M. Lavielle
e-mail: marc.lavielle@inria.fr

R. Argiento et al. (eds.), *Bayesian Statistics and New Generations*,
Springer Proceedings in Mathematics & Statistics 296,
https://doi.org/10.1007/978-3-030-30611-3_9

given the observations. Most often, direct sampling from this conditional distribution is impossible and it is necessary to have resort to a Markov chain Monte Carlo (MCMC) procedure.

Designing a fast mixing sampler is of utmost importance for several tasks in the complex process of model building. The most common MCMC method for nonlinear mixed effects models is the *random walk Metropolis* algorithm [9, 14, 15]. Despite its simplicity, it has been successfully used in many classical examples of pharmacometry, when the number of random effects is not too large. Nevertheless, maintaining an optimal acceptance rate (advocated in [15]) most often implies very small moves and therefore a very large number of iterations in medium and high dimensions since no information of the geometry of the target distribution is used.

To make better use of this geometry and in order to explore the space faster, the Metropolis-adjusted Langevin algorithm (MALA) uses evaluations of the gradient of the target density for proposing new states which are accepted or rejected using the Metropolis-Hastings algorithm [16, 18]. The No-U-Turn Sampler (NUTS) is an extension of the Hamiltonian Monte Carlo [11] that allows an automatic and optimal selection of some of the settings required by the algorithm, [3]. Nevertheless, these methods may be difficult to use in practice, and are computationally involved, in particular when the structural model is a complex ODE based model.

The algorithm we propose is a Metropolis-Hastings algorithm, but for which the proposal is a good approximation of the target distribution. For general data model (i.e. categorical, count or time-to-event data models or continuous data models), the Laplace approximation of the incomplete pdf $\mathtt{p}(y_i)$ leads to a Gaussian approximation of the conditional distribution $\mathtt{p}(\psi_i|y_i)$.

In the special case of continuous data, linearisation of the model leads, by definition, to a Gaussian linear model for which the conditional distribution of the individual parameter ψ_i given the data y_i is a multidimensional normal distribution that can be computed and we fall back on the results of [8].

2 Mixed Effect Models

2.1 Population Approach and Hierarchical Models

We will adopt a population approach in the sequel, where we consider N individuals and n_i observations for individual i. The set of observed data is $y = (y_i, 1 \leq i \leq N)$ where $y_i = (y_{ij}, 1 \leq j \leq n_i)$ are the observations for individual i. For the sake of clarity, we assume that each observation y_{ij} takes its values in some subset of $\mathbb{R}$. The distribution of the n_i-vector of observations y_i depends on a vector of individual parameters ψ_i that takes its values in a subset of $\mathbb{R}^p$. We assume that the pairs (y_i, ψ_i) are mutually independent and consider a parametric framework: the joint distribution of (y_i, ψ_i) is denoted by $\mathtt{p}(y_i, \psi_i; \theta)$, where θ is the vector of fixed parameters of the model. A natural decomposition of this joint distribution writes

$\mathtt{p}(y_i, \psi_i; \theta) = \mathtt{p}(y_i|\psi_i; \theta)\mathtt{p}(\psi_i; \theta)$, where $\mathtt{p}(y_i|\psi_i; \theta)$ is the conditional distribution of the observations, given the individual parameters, and where $\mathtt{p}(\psi_i; \theta)$ is the so-called population distribution used to describe the distribution of the individual parameters within the population. A particular case of this general framework consists in describing each individual parameters ψ_i as a typical value ψ_{pop}, and a vector of individual random effects η_i: $\psi_i = \psi_{\text{pop}} + \eta_i$. In the sequel, we will assume a multivariate Gaussian distribution for the random effects: $\eta_i \sim_{\text{i.i.d.}} \mathcal{N}(0, \Omega)$. Several extensions of this model are straightforward, considering for instance transformation of the normal distribution, or adding individual covariates in the model.

2.2 Continuous Data Models

A regression model is used to express the link between continuous observations and individual parameters:

$$y_{ij} = f(t_{ij}, \psi_i) + \varepsilon_{ij} \ , \tag{1}$$

where y_{ij} is the j-th observation for individual i measured at time t_{ij}, ε_{ij} is the residual error, f is the structural model assumed to be a twice differentiable function of ψ_i. We start by assuming that the residual errors are independent and normally distributed with zero-mean and a constant variance σ^2. Let $t_i = (t_{ij}, 1 \leq n_i)$ be the vector of observation times for individual i. Then, the model for the observations rewrites $y_i|\psi_i \sim \mathcal{N}(f_i(\psi_i), \sigma^2 \mathtt{Id}_{n_i \times n_i})$, where $f_i(\psi_i) = (f(t_{i,1}, \psi_i), \ldots, f(t_{i,n_i}, \psi_i))$. If we assume that $\psi_i \sim_{\text{i.i.d.}} \mathcal{N}(\psi_{\text{pop}}, \Omega)$, then the parameters of the model are $\theta = (\psi_{\text{pop}}, \Omega, \sigma^2)$.

3 Sampling from Conditional Distributions

The conditional distribution $\mathtt{p}(\psi_i|y_i; \theta)$ plays a crucial role in most methods used for inference in nonlinear mixed effects models.

One of the main task to perform is to compute the maximum likelihood (ML) estimate of θ, $\hat{\theta}_{\text{ML}} = \arg\max_{\theta \in \Theta} \mathcal{L}(\theta, y)$, where $\mathcal{L}(\theta, y) \triangleq \log \mathtt{p}(y; \theta)$. The stochastic approximation version of EM [7] is an iterative procedure for ML estimation that requires to generate one or several realisations of this conditional distribution at each iteration of the algorithm.

Metropolis-Hasting algorithm is a powerful MCMC procedure widely used for sampling from a complex distribution [4]. To simplify the notations, we remove the dependency on θ. For a given individual i, the MH algorithm, to sample from the conditional distribution $\mathtt{p}(\psi_i|y_i)$, is described in Algorithm 1.

Algorithm 1 Metropolis-Hastings algorithm

Initialization: Initialize the chain sampling $\psi_i^{(0)}$ from some initial distribution ξ_i .
Iteration k: given the current state of the chain $\psi_i^{(k-1)}$:

1. Sample a candidate ψ_i^c from a proposal distribution $q_i(\,\cdot\,|\psi_i^{(k-1)})$.
2. Compute the MH ratio:

$$\alpha(\psi_i^{(k-1)}, \psi_i^c) = \frac{\mathrm{p}(\psi_i^c|y_i)}{\mathrm{p}(\psi_i^{(k-1)}|y_i)} \frac{q_i(\psi_i^{(k-1)}|\psi_i^c)}{q_i(\psi_i^c|\psi_i^{(k-1)})} \ . \tag{2}$$

3. Set $\psi_i^{(k)} = \psi_i^c$ with probability $\min(1, \alpha(\psi_i^c, \psi_i^{(k-1)})$ (otherwise, keep $\psi_i^{(k)} = \psi_i^{(k-1)}$).

Current implementations of the MCMC algorithm, to which we will compare our new method, in Monolix [5], saemix (R package) [6], nlmefitsa (Matlab) and NONMEM [2] mainly use the same combination of proposals. The first proposal is an independent Metropolis-Hasting algorithm which consists in sampling the candidate state directly from the marginal distribution of the individual parameter ψ_i. The other proposals are component-wise and block-wise random walk procedures [10] that update different components of ψ_i using univariate and multivariate Gaussian proposal distributions. Nevertheless, those proposals fail to take into account the nonlinear dependence structure of the individual parameters. A way to alleviate these problems is to use a proposal distribution derived from a discretised Langevin diffusion whose drift term is the gradient of the logarithm of the target density leading to the Metropolis Adjusted Langevin Algorithm (MALA) [16, 18]. The MALA proposal is given by:

$$\psi_i^c \sim \mathcal{N}(\psi_i^{(k)} - \gamma \nabla_{\psi_i} \log \mathrm{p}(\psi_i^{(k)}|y_i), 2\gamma) \ , \tag{3}$$

where γ is a positive stepsize. These methods still do not take into consideration the multidimensional structure of the individual parameters. Recent works include efforts in that direction, such as the Anisotropic MALA for which the covariance matrix of the proposal depends on the gradient of the target measure [1]. The MALA algorithm is a special instance of the Hybrid Monte Carlo (HMC), introduced in [11]; see [4] and the references therein, and consists in augmenting the state space with an auxiliary variable p, known as the velocity in Hamiltonian dynamics.

All those methods aim at finding the proposal q that accelerates the convergence of the chain. Unfortunately they are computationally involved and can be difficult to implement (stepsizes and numerical derivatives need to be tuned and implemented).

We see in the next section how to define a multivariate Gaussian proposal for both continuous and noncontinuous data models, that is easy to implement and that takes into account the multidimensional structure of the individual parameters in order to accelerate the MCMC procedure.

4 A Multivariate Gaussian Proposal

For a given parameter value θ, the MAP estimate, for individual i, of ψ_i is the one that maximises the conditional distribution $\mathrm{p}(\psi_i|y_i,\theta)$:

$$\hat{\psi}_i = \arg\max_{\psi_i} \mathrm{p}(\psi_i|y_i,\theta) = \arg\max_{\psi_i} \mathrm{p}(y_i|\psi_i,\theta)\mathrm{p}(\psi_i,\theta)$$

4.1 General Data Models

For both continuous and noncontinuous data models, the goal is to find a simple proposal, a multivariate Gaussian distribution in our case, that approximates the target distribution $\mathrm{p}(\psi_i|y_i)$. In our context, we can write the marginal pdf $\mathrm{p}(y_i)$ that we aim to approximate as $\mathrm{p}(y_i) = \int e^{\log \mathrm{p}(y_i,\psi_i)} \mathrm{d}\psi_i$. Then, the Taylor expansion of $\log(\mathrm{p}(y_i,\psi_i)$ around the MAP $\hat{\psi}_i$ (that verifies by definition $\nabla \log \mathrm{p}(y_i,\hat{\psi}_i) = 0$) yields the Laplace approximation of $-2\log(\mathrm{p}(y_i))$ as follows:

$$-2\log \mathrm{p}(y_i) \approx -p\log 2\pi - 2\log \mathrm{p}(y_i,\hat{\psi}_i) + \log\left(\left|-\nabla^2 \log \mathrm{p}(y_i,\hat{\psi}_i)\right|\right) .$$

We thus obtain the following approximation of $\log \mathrm{p}(\hat{\psi}_i|y_i)$:

$$\log \mathrm{p}(\hat{\psi}_i|y_i) \approx -\frac{p}{2}\log 2\pi - \frac{1}{2}\log\left(\left|-\nabla^2 \log \mathrm{p}(y_i,\hat{\psi}_i)\right|\right) ,$$

which is precisely the log-pdf of a multivariate Gaussian distribution with mean $\hat{\psi}_i$ and variance-covariance $-\nabla^2 \log \mathrm{p}(y_i,\hat{\psi}_i)^{-1}$, evaluated at $\hat{\psi}_i$.

Proposition 1 *The Laplace approximation of the conditional distribution $\psi_i|y_i$ is a multivariate Gaussian distribution with mean $\hat{\psi}_i$ and variance-covariance*

$$\Gamma_i = -\nabla^2 \log p(y_i,\hat{\psi}_i)^{-1} = \left(-\nabla^2 \log p(y_i|\hat{\psi}_i) + \Omega^{-1}\right)^{-1} .$$

We shall now see another method to derive a Gaussian proposal distribution in the specific case of continuous data models.

4.2 Nonlinear Continuous Data Models

When the model is described by (1), the approximation of the target distribution can be done twofold: either by using the Laplace approximation, as explained above, or by linearizing the structural model f_i for any individual i of the population. Once the

MAP estimate $\hat{\psi}_i$ has been computed, using an optimisation procedure, the method is based on the linearisation of the structural model f around $\hat{\psi}_i$:

$$f_i(\psi_i) \approx f_i(\hat{\psi}_i) + \mathrm{J}_{f_i(\hat{\psi}_i)}(\psi_i - \hat{\psi}_i) \ , \tag{4}$$

where $\mathrm{J}_{f_i(\hat{\psi}_i)}$ is the Jacobian matrix of the vector $f_i(\hat{\psi}_i)$. Defining $z_i \triangleq y_i - f_i(\hat{\psi}_i) + \mathrm{J}_{f_i(\hat{\psi}_i)}\hat{\psi}_i$ yields a linear model $z_i = \mathrm{J}_{f_i(\hat{\psi}_i)}\psi_i + \epsilon_i$ which tractable conditional distribution can be used for approximating $\mathrm{p}(\psi_i|y_i, \theta)$:

Proposition 2 *Under this linear model, the conditional distribution $\psi_i|y_i$ is a Gaussian distribution with mean μ_i and variance-covariance Γ_i where*

$$\mu_i = \hat{\psi}_i \quad \textit{and} \quad \Gamma_i = \left(\frac{J'_{f_i(\hat{\psi}_i)} J_{f_i(\hat{\psi}_i)}}{\sigma^2} + \Omega^{-1} \right)^{-1} \ . \tag{5}$$

We can note that linearizing the structural model is equivalent to using the Laplace approximation with the expected information matrix. Indeed:

$$\mathbb{E}_{y_i|\hat{\psi}_i}\left(-\nabla^2 \log \mathrm{p}(y_i|\hat{\psi}_i)\right) = \frac{\mathrm{J}'_{f_i(\hat{\psi}_i)} \mathrm{J}_{f_i(\hat{\psi}_i)}}{\sigma^2} \ . \tag{6}$$

We then use this normal distribution as a proposal in Algorithm 1 for model (1).

5 A Pharmacokinetic Example

5.1 Data and Model

32 healthy volunteers received a 1.5 mg/kg single oral dose of warfarin, an anticoagulant normally used in the prevention of thrombosis [12], for who we measure warfarin plasmatic concentration at different times. We will consider a one-compartment pharmacokinetics (PK) model for oral administration, assuming first-order absorption and linear elimination processes:

$$f(t, ka, V, k) = \frac{D\, ka}{V(ka - k)}(e^{-ka\, t} - e^{-k\, t}) \ , \tag{7}$$

where ka is the absorption rate constant, V the volume of distribution, k the elimination rate constant, and D the dose administered. Here, ka, V and k are PK parameters that can change from one individual to another. Then, let $\psi_i = (ka_i, V_i, k_i)$ be the vector of individual PK parameters for individual i lognormally distributed. We will assume in this example that the residual errors are independent and normally distributed with mean 0 and variance σ^2. We can use the proposal given by Proposition 2 and based on a linearisation of the structural model f proposed in (7). For the method to be easily extended to any structural model, the gradient is calculated by automatic differentiation using the R package ‘Madness’ [13].

5.2 MCMC Convergence Diagnostic

We will consider one of the 32 individuals for this study and fix θ to some arbitrary value, close to the Maximum Likelihood (ML) estimate obtained with SAEM (saemix R package [6]): $ka_{\text{pop}} = 1$, $V_{\text{pop}} = 8$, $k_{\text{pop}} = 0.01$, $\omega_{ka} = 0.5$, $\omega_V = 0.2$, $\omega_k = 0.3$ and $\sigma^2 = 0.5$. First, we compare our our nlme-IMH, which is a MH sampler using our new proposal, with the RWM, the MALA, which proposal, at iteration k, is defined by $\psi_i^c \sim \mathcal{N}(\psi_i^{(k)} - \gamma_k \nabla \log \pi(\psi_i^{(k)}), 2\gamma_k)$. The stepsize ($\gamma = 10^{-2}$) is constant and is tuned such that the optimal acceptance rate of 0.57 is reached [15]. 20 000 iterations are run for each algorithm. Figure 1 highlights quantiles stabilisation using the MALA similar to our method for all orders and dimensions. The NUTS, implemented in rstan (R Package [17]), is fast and steady and presents similar, or even better convergence behaviors for some quantiles and dimension, than our method (see Fig. 1).

Then, we produce 100 independent runs of the RWM, the IMH using our proposal distribution (called the nlme-IMH algorithm), the MALA and the NUTS for 500 iterations. The boxplots of the samples drawn at a given iteration threshold are presented Fig. 2 against the ground truth (calculated running the NUTS for 100 000 iterations) for the parameter ka.

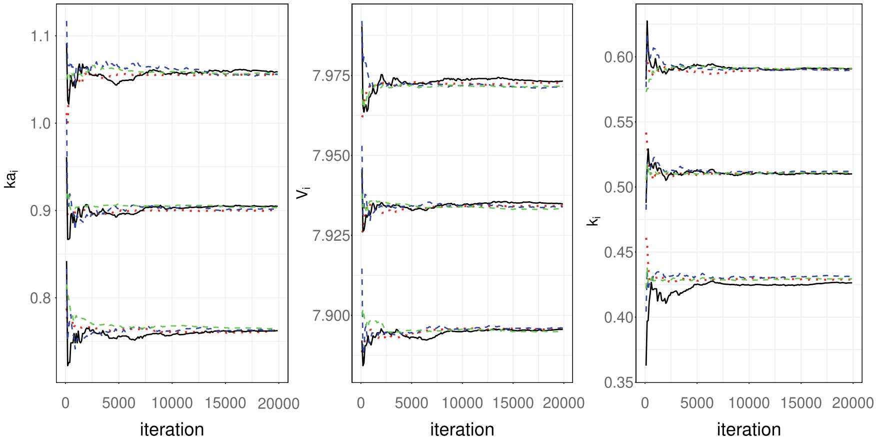


Fig. 1 Modelling of the warfarin PK data: convergence of the empirical quantiles of order 0.1, 0.5 and 0.9 of $\mathrm{p}(\psi_i | y_i; \theta)$ for a single individual. Our new MH algorithm is in red and dotted, the RWM is in black and solid, the MALA is in blue and dashed and the NUTS is in green and dashed

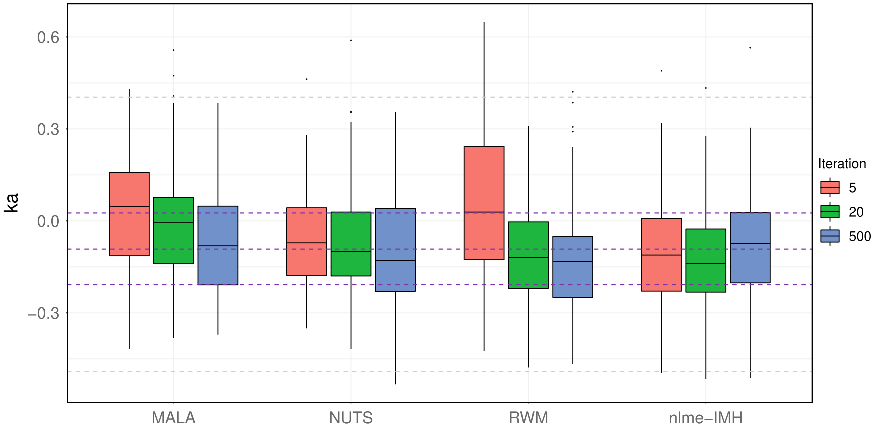


Fig. 2 Modelling of the warfarin PK data: Boxplots, over 100 parallel runs, for the RWM, the nlme-IMH, the MALA and the NUTS algorithm. The ground truth median, 0.25 and 0.75 percentiles are plotted as a dashed purple line and its maximum and minimum as a dashed grey line

For the three numbers of iteration considered in Fig. 2, the median of the nlme-IMH and NUTS samples are closer to the ground truth. Figure 2 also highlights that all those methods succeed in sampling from the whole distribution after 500 iterations. Similar comments can be made for the other parameters.

We decided to conduct a comparison between those sampling methods just in terms of number of iterations (one iteration is one transition of the Markov Chain). We acknowledge that the transition cost is not the same for each of those algorithms, though, our nmle-IMH algorithm, except the initialisation step that requires a MAP and a Jacobian computation, has the same iteration cost as RWM. The call to the structural model f being very costly in real applications (when the model is the solution of a complex ODE for instance), the MALA and the NUTS, computing its first order derivatives at each transition, are thus far computationally involved.

Since computational costs per transition are hard to accurately define for each sampling algorithm and since runtime depends on the actual implementation of those methods, comparisons are based on the number of iterations of the chain here.

6 Conclusion and Discussion

We presented in this article an independent Metropolis-Hastings procedure for sampling random effects from their conditional distributions in nonlinear mixed effects models. The numerical experiments that we have conducted show that the proposed sampler converges to the target distribution as fast as state-of-the-art samplers. This good practical behaviour is partly explained by the fact that the conditional mode of the random effects in the linearised model coincides with the conditional mode of the random effects in the original model. Initial experiments embedding this fast and easy-to-implement IMH algorithm within the SAEM algorithm [7], for Maximum Likelihood Estimation, indicate a faster convergence behavior.

References

1. Allassonniere, S., Kuhn, E.: Convergent stochastic expectation maximization algorithm with efficient sampling in high dimension. Application to Deformable Template Model Estimation. Comput. Stat. Data Anal. **91**, 4–19 (2015)
2. Beal, S., Sheiner, L.: The NONMEM system. Am. Stat. **34**(2), 118–119 (1980)
3. Betancourt, M.: A Conceptual Introduction to Hamiltonian Monte Carlo (2017). arXiv:1701.02434
4. Brooks, S., Gelman, A., Jones, G., Meng, X.-L.: Handbook of Markov Chain Monte Carlo. CRC Press (2011)
5. Chan, P.L.S., Jacqmin, P., Lavielle, M., McFadyen, L., Weatherley, B.: The use of the SAEM algorithm in MONOLIX software for estimation of population pharmacokinetic-pharmacodynamic-viral dynamics parameters of maraviroc in asymptomatic HIV subjects. J. Pharmacokinet. Pharmacodyn. **38**(1), 41–61 (2011)
6. Comets, E., Lavenu, A., Lavielle, M.: Parameter estimation in nonlinear mixed effect models using saemix, an R implementation of the SAEM algorithm. J. Stat. Softw. **80**(3), 1–42 (2017)
7. Delyon, B., Lavielle, M., Moulines, E.: Convergence of a stochastic approximation version of the EM algorithm. Ann. Stat. **27**(1), 94–128 (1999)
8. Karimi, B., Lavielle, M., Moulines, E.: Non linear mixed effects models: bridging the gap between independent Metropolis-Hastings and variational inference. ICML 2017 Implicit Models Workshop (2017)
9. Lavielle, M.: Mixed Effects Models for The Population Approach: Models, Tasks. CRC Press, Methods and Tools (2014)
10. Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E.: Equation of state calculations by fast computing machines. J. Chem. Phys. **21**(6), 1087–1092 (1953)
11. Neal, R.M.: MCMC using Hamiltonian dynamics. Handbook of Markov Chain Monte Carlo, vol. 2(11) (2011)
12. O'Reilly, R.A., Aggeler, P.M.: Studies on Coumarin anticoagulant drugs initiation of Warfarin therapy without a lading dose. Circulation **38**(1), 169–177 (1968)
13. Pav, S.E.: Madness: A Package for Multivariate Automatic Differentiation (2016)
14. Robert, C.P., Casella, G.: Monte Carlo Statistical Methods. Springer Texts in Statistics (2004)
15. Roberts, G.O., Gelman, A., Gilks, W.R.: Weak convergence and optimal scaling of random walk Metropolis algorithms. Ann. Appl. Probab. **7**(1), 110–120 (1997)
16. Roberts, G.O., Tweedie, R.L.: Exponential convergence of Langevin distributions and their discrete approximations. Bernoulli **2**(4), 341–363 (1996)
17. Stan Development Team: RStan: the R interface to Stan. R Package Version 2.17.3 (2018)
18. Stramer, O., Tweedie, R.L.: Langevin-type models I: Diffusions with given stationary distributions and their discretizations. Methodol. Comput. Appl. Probab. **1**(3), 283–306 (1999)

Comparison Between Suitable Priors for Additive Bayesian Networks

Gilles Kratzer, Reinhard Furrer and Marta Pittavino

Abstract Additive Bayesian networks (ABN) are types of graphical models that extend the usual Bayesian-generalised linear model to multiple dependent variables through the factorisation of the joint probability distribution of the underlying variables. When fitting an ABN model, the choice of the prior for the parameters is of crucial importance. If an inadequate prior—like a not sufficiently informative one—is used, data separation and data sparsity may lead to issues in the model selection process. In this work we present a simulation study to compare two weakly informative priors with a strongly informative one. For the weakly informative prior, we use a zero mean Gaussian prior with a large variance, currently implemented in the R package *abn*. The candidate prior belongs to the Student's t-distribution. It is specifically designed for logistic regressions. Finally, the strongly informative prior is Gaussian with a mean equal to the true parameter value and a small variance. We compare the impact of these priors on the accuracy of the learned additive Bayesian network as function of different parameters. We create a simulation study to illustrate Lindley's paradox based on the prior choice. We then conclude by highlighting the good performance of the informative Student's t-prior and the limited impact of Lindley's paradox. Finally, suggestions for further developments are provided.

Keywords Graph theory · Structural search · Binomial regression

G. Kratzer (✉)
Department of Mathematics, University of Zurich, Zurich, Switzerland
e-mail: gilles.kratzer@math.uzh.ch

R. Furrer
Department of Mathematics and Department of Computational Science,
University of Zurich, Zurich, Switzerland
e-mail: reinhard.furrer@math.uzh.ch

M. Pittavino
Geneva School of Economics and Management, Research Center for Statistics,
University of Geneva, Geneva, Switzerland
e-mail: marta.pittavino@unige.ch

R. Argiento et al. (eds.), *Bayesian Statistics and New Generations*,
Springer Proceedings in Mathematics & Statistics 296,
https://doi.org/10.1007/978-3-030-30611-3_10

1 Introduction to ABN

Additive Bayesian network (ABN) models are types of graphical models that extend the usual generalised regression framework to multiple dependent variables through the representation of the joint probability distribution. ABNs are a special type of Bayesian network (BN) model in that each node in the graph comprises a generalised linear model (GLM). This property is used to compute additively the global score of the structure. All types of BN models consist of two reciprocally dependent parts: a qualitative one (the structure) and a quantitative one (the model parameters). BN models are statistical models that derive a directed acyclic graph (DAG) from empirical data, describing the dependency structure of the random variables. The DAG is the graphical representation of the joint probability distribution of all random variables represented by the data. The model parameters stem from the local probability distribution of all the variables in the network.

In the last few decades, BN modelling has been widely used in biomedical science and in systems biology to analyse multi-dimensional data [3, 4, 12, 13, 25]. Recently, ABN modelling approaches have been successfully applied to the field of veterinary epidemiology [10, 24, 28]. A very appealing feature of ABN is its ability to generalise standard regression methodology. A general introduction to BN and ABN modelling in veterinary epidemiology is provided by [18]. Further applications of BN to veterinary studies were described by [29]. Graphical modelling techniques based on ABN used to analyse epidemiological data were used by [17, 19, 20], resulting in several publications.

The paper is structured as follows. Relevant technical details of ABN models are presented in Sect. 2. Section 3 explains the issue of data separation and Lindley's paradox and highlights the importance of appropriate prior choice. Section 4 reports the results of a simulation study underpinning the necessity of careful prior selection with respect to data separation and Lindley's paradox. We conclude the article in Sect. 5 with future research directions.

The main original contribution of this paper is to compare the impact of different priors on the fitting performance of ABN models using synthetic datasets in a realistic epidemiological setting. The underlying idea is to render the ABN approach more robust in practice when using an adequate prior. This is of high importance, as a typical systems epidemiology dataset has a relatively small sample size. Our findings show that none of the currently proposed alternative priors are fully satisfactory and further studies need to be conducted.

2 Additive Bayesian Networks in a Nutshell

A BN model $\mathcal{B}$ for a set of random variables $\boldsymbol{X} = \{X_1, \dots, X_n\}$ consists of:

- A DAG structure $\mathcal{S} = (V, E)$, where V is a finite set of nodes and E is a finite set of directed edges between the nodes. A DAG is *acyclic*; hence, the edges in

E do not form directed cycles. A random variable X_j corresponds to each node $j \in V = \{1, \ldots, n\}$ in the graph. We do not distinguish between a variable X_j and the corresponding node j.

- A node k is said to be a *parent* of a node j if the edge set E contains an edge from k to j. A set of parents for a node j is denoted by $\mathbf{Pa}_j$. P_j indicates the total number of parents for a node j, i.e., $\dim(\mathbf{Pa}_j) = P_j \geq 0$ and $P_j = 0$ for orphan nodes.
- A set of local probability distributions for all variables in the network is encoded by $\boldsymbol{\theta}_{\mathcal{B}}$. Each node j, with parent set $\mathbf{Pa}_j$, is parametrised by a local probability distribution: $P(X_j \mid \mathbf{Pa}_j)$.

We denote a BN model $\mathcal{B}$ by the pair $\mathcal{B} = (\mathcal{S}, \boldsymbol{\theta}_{\mathcal{B}})$, representing the DAG *structure* $\mathcal{S}$ and the model *parametrization* $\boldsymbol{\theta}_{\mathcal{B}}$, respectively.

Edges represent both *marginal* and *conditional dependencies*. The main role of the network structure is to express the conditional independence relationships among the variables in the model through graphical separation, thus specifying the factorisation of the global probability distribution:

$$P(\boldsymbol{X}) = \prod_{j=1}^{n} P(X_j \mid \mathbf{Pa}_j). \tag{1}$$

The left panel of Fig. 1 shows a simple BN example for four nodes.

A general BN has an arbitrary distribution for each of the factors in (1), whereas an ABN defines each of these factors through a GLM. For example, assuming a discrete distribution for node 3 of Fig. 1, using classical notation for the exponential family parametrisation [22] its probability mass function writes

$$P(X_3 = x \mid X_1 = x_1, X_2 = x_2) = \exp\big(\eta(\theta_3)T(x) - A(\theta_3)\big)H(x), \tag{2}$$

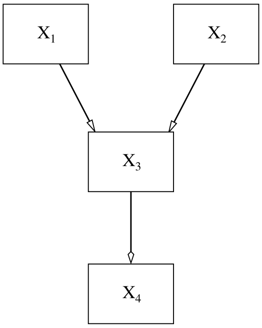
where the functions η, T, A, H may be node-dependent (we have omitted the indices to simplify the notation) and where θ_3 incorporates the configuration of the parents' node

$$\eta(\theta_3) = \beta_{3,0} + \beta_{3,1}x_1 + \beta_{3,2}x_2. \tag{3}$$

For nodes with continuous distributions or with different number of parents, (2) and (3) can be adapted accordingly.

Using only binary variables, i.e., $X_j \in \{0, 1\}$, simplifies the previous equations. More specifically, each node equation (2) collapses to θ_j and represents 2^{P_j} nonzero probabilities. In (2) we use the logit link function, resulting in classical logistic regression models for all nodes.

In the Bayesian framework, the priors are placed upon the parameters $\beta_{j,k}$, $j = 1, \ldots, n$ and $k = 0, \ldots, P_j$.



X_1 is (conditionally) independent of X_2:

$$P(X_1 = 1) = \theta_1, \qquad \log\Big(\frac{\theta_1}{1-\theta_1}\Big) = \beta_{1,0}$$

X_2 is (conditionally) independent of X_1:

$$P(X_2 = 1) = \theta_2, \qquad \log\Big(\frac{\theta_2}{1-\theta_2}\Big) = \beta_{2,0}$$

X_3 is conditionally dependent jointly upon X_1 and X_2:

$$P(X_3 = 1 \mid x_1, x_2) = \theta_3,$$
$$\log\Big(\frac{\theta_3}{1-\theta_3}\Big) = \beta_{3,0} + \beta_{3,1}x_1 + \beta_{3,2}x_2$$

X_4 is conditionally dependent upon X_3:

$$P(X_4 = 1 \mid x_3) = \theta_4, \qquad \log\Big(\frac{\theta_4}{1-\theta_4}\Big) = \beta_{4,0} + \beta_{4,1}x_3$$

Fig. 1 Left: A four node Bayesian network with structure $\mathcal{S} = (V, E)$, where $V = \{1, \dots, 4\}$, $E = \{(1,3), (2,3), (3,4)\}$, e.g., $\mathbf{Pa}_1 = \mathbf{Pa}_2 = \varnothing$, $\mathbf{Pa}_3 = \{1,2\}$ and $\mathbf{Pa}_4 = \{3\}$. Right: ABN parametrisation for binary nodes based on the logit link function. Note that the probabilities θ_3 and θ_4 depend on the configuration of the parents

3 Potential Limitation When Fitting ABN Models

From an applied perspective when fitting ABN models, we need to be aware of two possible issues: data separation and Lindley's paradox. The former is linked to the data and the latter is inherently linked to the fitting procedure. An adequate prior could potentially control both of their impacts on the fitted structure.

3.1 Data Separation

The data separation arises when a linear combination of predictors perfectly predicts the outcome. This is surprisingly common in applied logistic regression. Data separation induces estimation problems for the entire model, not only for the parameters directly involved.

Due to the large number of models necessary to evaluate (for each node a GLM for each parent configuration), data separation is a serious concern when modelling discrete data with an ABN model. The separation occurs when the dataset is too small to observe events with low probabilities. Therefore, the smaller the sample size, the higher the probability of not observing given instances which have a low probability. The issue is intensified with increasing complexity of the model. A popular solution is to remove predictors until the design matrix becomes fully ranked. However, this often leads to the deletion of the strongest predictors, which is not desirable, especially in the context of ABN [30]. Alternatively, the natural "Bayesian" solution is to use a prior that will drive the posterior whenever data separation arises. Multiple prior distributions have been proposed to tackle this issue. A notable one is the Jeffreys prior [5] which is, however, hard to interpret as prior information. Indeed, the Jeffreys prior is not parametrised on the scale of the parameter. Moreover, when applied to sparse data the prior may lead to poor numerical results. As a result, dedicated priors have been developed which are weakly informative enough to be used in a general context and which can still drive the posterior if separation arises [8]. They have been designed to produce stable and regularised estimates. These priors are based on the Student's t-distribution. This paper compares the effect of the currently implemented prior in the *abn* R package, which is non-informative with the weakly informative prior promoted by [8].

3.2 Lindley's Paradox

A common approach to fitting an ABN model is using a score-based method. A popular choice is to take a score that is decomposable, i.e., each node of the network has a score contribution which is additive. The score of the total network is the sum of each individual atomic network. Typically a cache of scores is pre-computed and an optimisation algorithm is applied to select the structure that contains all nodes and maximises the score, i.e., choosing a global structure. The cache-building essentially implies fitting all possible combinations of the random variables. Even for very limited numbers of random variables, approximations are required to render the computation tractable.

Under the viewpoint of building a global structure based on the nodes' individual models, ABN can be viewed as a model selection technique. It is known that when a weakly informative prior is used, Bayesian model selection will asymptotically always prefer the simpler model, regardless of the data. This is called Lindley's paradox [21]. Using a weakly informative prior for the parameters leads to reasonable parameter estimates compared to a pure maximum likelihood estimation for a given network. But the main objective of ABN analysis is performing structural inference, which is precisely negatively affected by weakly informative priors. So, Lindley's paradox is potentially a massive threat for an ABN analysis. Indeed, the quality of the inferred structure relies on the paradigm that the score should be representative (as a proxy) of the ability of a given structure to fit the data. If a systematic asymptotic error is made regarding the scoring of the atomic structures, then the final model selected will tend to be too simple compared to an ABN selection using frequentist computed scores. Controlling the final complexity of the model through the adequate choice of the parameter prior is a major challenge when using Bayesian model selection approach applied to ABN models.

4 Implementation and Simulation Study

From a practical perspective, computational speed is the major concern in an ABN context. Indeed the number of models to be evaluated is a function of the number of nodes. This function is super-exponential with the number of random variables [27].

The estimation of Bayesian regression coefficients using Gibbs or Metropolis algorithms is usually not fast enough, especially because the model selection approach is based on a point estimate of the posterior rather than on the full network information. An appealingly fast and reliable procedure to fit the model and return an approximate posterior mode is described in [8] and a detailed explanation is given in the Sect. 14.8 of [7]. The procedure is an alteration of the classical iterative reweighted least squares algorithm that uses an approximate expectation-maximisation algorithm to update the regression coefficients at each step. The prior information is taken into account through augmented data. This procedure is used to estimate the posterior mode for every possible combination of all the variables. The output of this procedure is a comprehensive list of scores. Further details for this first step are given in [16]. In a second step, an exact search is performed to select the network with the highest possible global score [14]. The simulation study has been carried out using the package *abn* [15] in the R software environment [26].

4.1 Data Separation

In order to illustrate the influence of the prior on an ABN analysis, we randomly simulate BNs consisting of 10 binary random variables with 80% of the possible edges expressed. Each edge represents the same regression coefficient (β coefficients in the right panel of Fig. 1) set to 0.99 on the logit scale, i.e., =expit(5). For sample sizes $N = 100$, 500, 1000 and 10,000 we randomly generate 50 distributions of the selected network. The two priors used are a weakly informative prior (WI) which is a normal distribution with mean zero and variance 1000 and a Student's t-prior (ST) with one degree of freedom (i.e., Cauchy) and scale parameter 2.5. The simulation results are not very sensitive to the scale parameter within the range of 1.5–3. Then the true positive rate (TPR) and the false positive rate (FPR) are used to measure the accuracy of the selected networks. Every selected network is transformed to an essential graph, as two networks of the same Markov class of equivalence could differ substantially in terms of structure but have the same score because they represent the same assertions of conditional independence [6]. Indeed, the implemented scoring approach can differentiate networks up to the Markov class of equivalence only.

Figure 2 shows the TPR and FPR as a function of the sample size for two different priors and illustrates that both priors exhibit a proper "asymptotic" behaviour when sample size increases: TPR and FPR tend to 100% and 0%, respectively. The chosen coefficients (0.99) of the edges in each BN leads almost surely to data separation for most of the possible variables' combinations. Not surprisingly, the Student's t-prior has a better accuracy for network scoring for selecting both positive and negative edges.

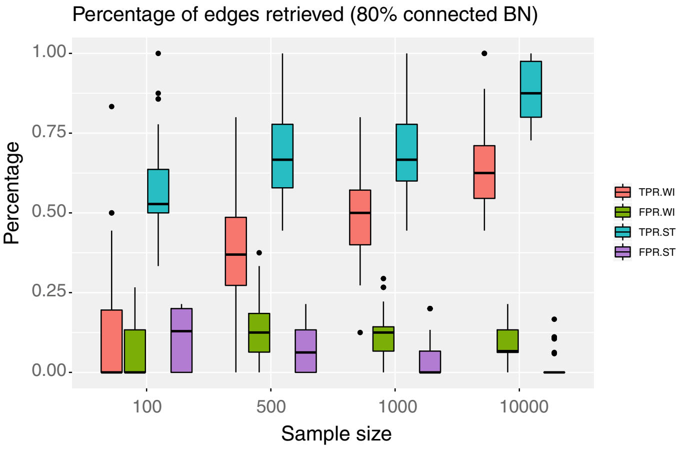


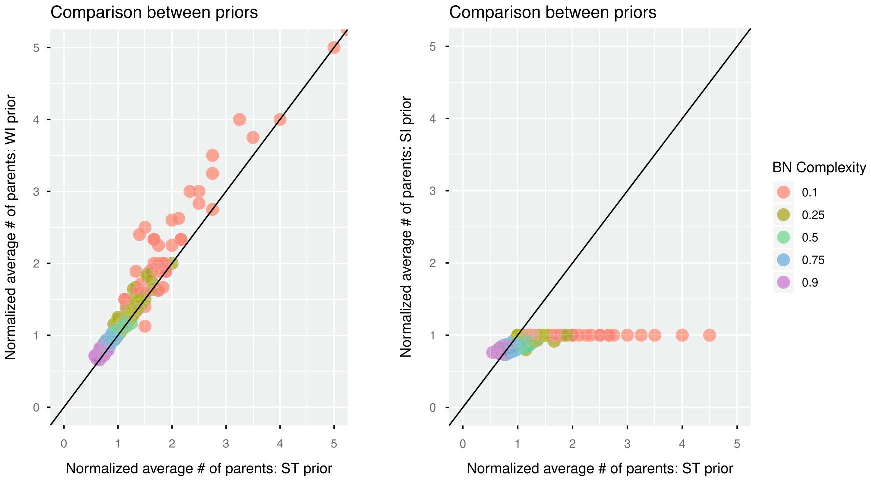
Fig. 2 Accuracy measures for retrieved edges for 80%-connected ten nodes ($n = 10$) simulated Bayesian networks as a function of data sample size ($N = 100$, 500, 1000 and 10,000). The boxplots (each based on 50 simulations) show the true positive rate (TPR) and the false positive rate (FPR) outcome of the weakly informative prior (WI) and Student's t-prior (ST)

4.2 Lindley's Paradox

An ABN modelling approach is based on multiple approximations. The network score is a proxy for selecting the best network. This score is conveniently chosen as decomposable, i.e., the marginal likelihood in this case. The procedure to compute the marginal likelihood is subject to numerical approximations. So even if Lindley's paradox is a known theoretical concern, it could potentially have a limited impact in practice.

In order to illustrate Lindley's paradox in a plausible situation, we randomly simulate BNs of $n = 10$ nodes with a range of different edge densities. Each edge has a known regression coefficient (β coefficients in the right panel of Fig. 1). Then, we simulate 50 synthetic datasets of 1000 observations per network density. For this simulation study three priors have been used: the two priors described above and a strongly informative prior (SI), which is a normal distribution with its mean set to the true regression coefficient for each edge and variance 0.1. Of course, this last prior is not realistic in practice but it is added here to illustrate the "asymptotic" behaviour. The average normalised number of parents is used to illustrate Lindley's paradox. For this illustration, we divide the average number of a simulated network by the true number of parents of the original network. Then, BNs are fitted using binomial regression with different priors, and the essential graphs are extracted.

Figure 3 summarises the simulation result and compares the (normalised) average number of parents of the fitted BN under different priors. If the selected DAGs are subjected to Lindley's paradox due to the weakness of prior information, one should see a scatter plot deviating from the diagonal. Indeed, the Lindley's paradox implies



(a) Student's t-prior (ST) and the weakly informative prior (WI).

(b) Student's t-prior (ST) and the strongly informative prior (SI).

Fig. 3 Comparison of different priors for different network complexities (edge densities varying between 0.1 and 0.9)

that the selected model tends to be too simple. Then the normalised average number of parents tend to be higher than one and should depend on the prior used. As seen in Fig. 3, sparse networks, i.e. low network complexity, are more impacted by than highly connected ones. The marginal posterior likelihood seems to overfit the sparse network structure and to underfit dense networks. In Fig. 3a one can see that a weakly informative prior performs comparably as Student's t-prior, whereas the effect of highly informative prior is clearly visible in Fig. 3b. The selected networks almost never exceed the true networks in term of complexity. Surprisingly, even such a prior does not allow the scoring procedure to optimally select dense networks. Indeed, one could expect that such amounts of information would lead to perfect scoring and thus quasi no sampling error. Again, with 1000 observations, the simulation results are not very sensitive to the scale parameter.

5 Future Developments

In Sect. 4 we showed that parameter priors play a major role in ABN modelling by (i) comparing the effect of different priors on data separation when dealing with discrete data and (ii) selecting networks depending on the prior information to address the Lindley's paradox conditional to the structure complexity. The simulation study highlights the need to further study suitable priors for ABN modelling. Indeed, the presented priors are not fully satisfactory. They partially improve the situation in the discrete case, but as mentioned in the introduction, an ABN modelling usually involves a mixture of distributions.

A conjugate prior that contains enough information regarding the data would lead to evident benefits. For example, a closed-form distribution for the posterior might be available. This result would lead to huge advantages in terms of marginal likelihood computation by reducing the time for the structure selection process. Similarly, the parameters estimates will also benefit from this choice. Another common issue of the BN literature is the so-called score equivalence problem [11, 23] that could be solved using an appropriate prior.

In order to achieve this goal, we consider the link between ABN models and GLMs and exploit features of the exponential family. A good candidate for this purpose is the conjugate prior distribution that belongs to a flexible family of priors called the Diaconis–Ylvisaker conjugate priors [2]. This prior distribution was introduced by [1]. A change of variables and the resulting properties need to be checked (as in [9]) in order to apply this distribution to our specific case. Further work will be conducted in this direction in order to formally verify all the desirable assumptions. Additionally, the R package *abn* [15] should be equipped with further priors for practical usage and availability for the statistical community.

References

1. Chen, M., Ibrahim, J.G.: Conjugate priors for generalized linear models. Statistica Sinica **13**, 461–476 (2003)
2. Diaconis, P., Ylvisaker, D.: Conjugate priors for exponential families. Ann. Stat. **7**(2), 269–281 (1979)
3. Djebbari, A., Quackenbush, J.: Seeded Bayesian networks: constructing genetic networks from microarray data. BMC Syst. Biol. **2**(1), 57 (2008)
4. Dojer, N., Gambin, A., Mizera, A., Wilczyński, B., Tiuryn, J.: Applying dynamic Bayesian networks to perturbed gene expression data. BMC Bioinform. **7**(1), 249 (2006)
5. Firth, D.: Bias reduction of maximum likelihood estimates. Biometrika **80**(1), 27–38 (1993)
6. Flesch, I., Lucas, P.J.: Markov equivalence in Bayesian networks. In: Lucas, P., Gámez, J.A., Salmerón, A. (eds.) Advances in Probabilistic Graphical Models, pp. 3–38. Springer, Berlin, Heidelberg (2007)
7. Gelman, A., Stern, H.S., Carlin, J.B., Dunson, D.B., Vehtari, A., Rubin, D.B.: Bayesian Data Analysis. Chapman and Hall/CRC (2013)
8. Gelman, A., Jakulin, A., Pittau, M.G., Su, Y.S.: A weakly informative default prior distribution for logistic and other regression models. Ann. Appl. Stat. **2**(4), 1360–1383 (2008)
9. Gutiérrez-Peña, E., Smith, A.F.M.: Conjugate parameterizations for natural exponential families. J. Am. Stat. Assoc. **90**(432), 1347–1356 (1995)
10. Hartnack, S., Springer, S., Pittavino, M., Grimm, H.: Attitudes of Austrian veterinarians towards euthanasia in small animal practice: impacts of age and gender on views on euthanasia. BMC Vet. Res. **12**(1), 26 (2016)
11. Heckerman, D., Geiger, D., Chickering, D.M.: Learning Bayesian networks: the combination of knowledge and statistical data. Mach. Learn. **20**(3), 197–243 (1995)
12. Hodges, A.P., Dai, D., Xiang, Z., Woolf, P., Xi, C., He, Y.: Bayesian network expansion identifies new ROS and biofilm regulators. PLOS One **5**(3), e9513 (2010)

13. Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S., Emili, A., Snyder, M., Greenblatt, J.F., Gerstein, M.: A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science **302**(5644), 449–453 (2003)
14. Koivisto, M., Sood, K.: Exact Bayesian structure discovery in Bayesian networks. J. Mach. Learn. Res. **5**(May), 549–573 (2004)
15. Kratzer, G., Pittavino, M., Ian, L.F., Lewis, I.F.: abn: an R package for modelling multivariate data using additive Bayesian networks, R package version 1.3 (2018). https://CRAN.R-project.org/package=abn
16. Kratzer G, Furrer R (2018) Information-Theoretic Scoring Rules to Learn Additive Bayesian Network Applied to Epidemiology. arXiv:1808.01126
17. Lewis, F.I.: Bayesian networks as a tool for epidemiological systems analysis. In: AIP Conference Proceedings vol. 1493, pp. 610–617 (2012)
18. Lewis, F.I., Brülisauer, F., Gunn, G.J.: Structure discovery in Bayesian networks: an analytical tool for analysing complex animal health data. Prev. Vet. Med. **100**(2), 109–115 (2011)
19. Lewis, F.I., McCormick, B.J.: Revealing the complexity of health determinants in resource-poor settings. Am. J. Epidemiol. **176**(11), 1051–1059 (2012)
20. Lewis, F.I., Ward, M.P.: Improving epidemiologic data analyses through multivariate regression modelling. Emerg. Themes Epidemiol. **10**(1), 4 (2013)
21. Lindley, D.V.: A statistical paradox. Biometrika **44**(1/2), 187–192 (1957)
22. Pitman, E.J.G.: Sufficient statistics and intrinsic accuracy. Math. Proc. Camb. Philos. Soc. **32**(4), 567–579 (1936)
23. Pittavino, M.: Additive Bayesian networks for multivariate data: parameter learning, model fitting and applications in veterinary epidemiology. Ph.D. thesis, University of Zurich (2016)
24. Pittavino, M., Dreyfus, A., Heuer, C., Benschop, J., Wilson, P., Collins-Emerson, J., Torgerson, P.R., Furrer, R.: Comparison between generalized linear modelling and additive Bayesian network; identification of factors associated with the incidence of antibodies against Leptospira interrogans sv Pomona in meat workers in New Zealand. Acta Trop. **173**, 191–199 (2017)
25. Poon, A.F.Y., Lewis, F.I., Pond, S.L.K., Frost, S.D.W.: Evolutionary interactions between N-linked glycosylation sites in the HIV-1 envelope. PLOS Comput. Biol. **3**(1), e11 (2007)
26. R Core Team: R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2017)
27. Robinson, R.W.: Counting unlabeled acyclic digraphs. In: Little, C.H.C. (ed.) Combinatorial Mathematics V, pp. 28–43. Springer, Berlin, Heidelberg (1977)
28. Sanchez-Vazquez, M.J., Nielen, M., Edwards, S.A., Gunn, G.J., Lewis, F.I.: Identifying associations between pig pathologies using a multi-dimensional machine learning methodology. BMC Vet. Res. **8**(1), 151 (2012)
29. Ward, M.P., Lewis, F.I.: Bayesian graphical modelling: applications in veterinary epidemiology. Prev. Vet. Med. **110**(1), 1–3 (2013)
30. Zorn, C.: A solution to separation in binary response models. Polit. Anal. **13**(2), 157–170 (2005)

A Bayesian Nonparametric Model for Integrative Clustering of Omics Data

Iliana Peneva and Richard S. Savage

Abstract Cancer is a complex disease, driven by a range of genetic and environmental factors. Many integrative clustering methods aim to provide insight into the mechanisms underlying cancer but few of them are computationally efficient and able to estimate the number of subtypes. We have developed a Bayesian nonparametric model for combined data integration and clustering called BayesCluster, which aims to identify cancer subtypes and addresses many of the issues faced by the existing integrative methods. The proposed method can integrate and use the information from multiple different datasets, and offers better cluster interpretability by using nonlocal priors. We incorporate feature learning because of the large number of predictors, and use a Dirichlet process mixture model approach to produce the patient subgroups. We ensure tractable inference with simulated annealing. We apply the model to datasets from the Cancer Genome Atlas project of glioblastoma multiforme, which contains clinical and biological data about cancer patients with extremely poor prognosis of survival. By combining all available information we are able to be better identify clinically meaningful subtypes of glioblastoma.

Keywords Bayesian nonparametrics · Data integration · Glioblastoma · Mixture models · Non-local priors

1 Introduction

Every year approximately 14.1 million people are diagnosed with a type of cancer [44] and the survival prognosis for many of them is poor due to the lack of understanding of the causes of some cancers. Modern large-scale projects such as the Cancer

I. Peneva (✉)
University of Warwick, Warwick, UK
e-mail: i.peneva@warwick.ac.uk

R. S. Savage
Department of Statistics, University of Warwick, Warwick, UK
e-mail: r.s.savage@warwick.ac.uk

R. Argiento et al. (eds.), *Bayesian Statistics and New Generations*,
Springer Proceedings in Mathematics & Statistics 296,
https://doi.org/10.1007/978-3-030-30611-3_11

Genome Atlas (TCGA) [42], METABRIC [5] and the International Cancer Genome Consortium (ICGC) [12] offer a great opportunity to study the mechanisms underlying different types of cancer. However, they also bring the challenges of selecting informative features, estimating the number of cancer subtypes, and providing interpretative results.

Many recent studies propose an integrative clustering approach to solving these problems, based on the idea that none of the individual datasets can fully capture the complexity of cancer, but collectively, they can offer a better understanding of the true oncogenic mechanisms. These studies can broadly be divided into two categories: those which cluster each dataset individually and then define the final partition by a post hoc integration of the separate clusterings, and those which combine all data sources to determine a single partition. Using the first approach [6, 41] fails to identify the common structure shared between the different data types, whereas the second approach [19, 38] fails to identify patterns that are unique to the individual data types. Both approaches often have slow parameter inference and are limited in the type of the datasets they can model.

In this paper, we present an efficient clustering algorithm called BayesCluster, which addresses these drawbacks. It makes use of mixture models, which have become increasingly popular in integrative clustering [1, 16, 21, 35]. In this way, we can easily model different types of data and perform faster, more efficient inference. We illustrate the model with the TCGA study of glioblastoma, which is the most aggressive brain cancer and has some of the worst survival prognosis, with a median survival of about 14 months [37]. The results from this work have the potential for improving this situation by providing more specific patient subtypes and possible biomarkers.

2 Data

We downloaded gene expression (GE), copy number variation (CNV), microRNA (miRNA), and methylation (ME) data from the glioblastoma multiforme project, available on TCGA data portal (http://cancergenome.nih.gov/). After matching samples across all data types, we were left with 211 samples for which we have complete genomic data. There were a few duplicate samples for the same patient in the datasets, for which we made a blind selection of the first sample, based on barcode ordering.

We selected for subsequent analyses the most highly variable genes in each of the genomic datasets. We used the publicly available level 3 gene expression data on the UNC AgilentG4502A_07 platform, level 2 copy number data, level 3 miRNA data, generated by UNC on the H-miRNA 8×15 K platform, and the publicly available level 3 methylation data on HumanMethylation450 platform. We set all missing values to 0 because we assume that the data is zero-centred and normalised. This resulted in the selection of 122, 112 and 125 genes from the gene expression, copy number variation and miRNA data, respectively. In the case of the methylation data, the measurements were in the form of beta values, which are equal to the ratio of methylation signal to methylation + background signal. We binarised the data ($\beta > 0.95$) and removed any features with fewer than 10 hits. That left us with 106 features.

3 Methodology

3.1 Statistical Models

We use two different statistical models to model the real-valued and discrete types of data in the glioblastoma study.

We model each D_t-dimensional continuous observation $\mathbf{x}_{it}$ in dataset t with N observations by a Gaussian likelihood with unknown mean and variance:

$$p(\mathbf{x}_{it}|\mathbf{z}_i, \mathbf{W}_t) = \mathcal{N}(\mathbf{x}_{it}|\mathbf{W}_t\mathbf{z}_i, \sigma_t^2\mathbf{I}), \tag{1}$$

where the latent variables $\mathbf{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_N\} \in \mathbb{R}^{P\times N}$ represent the true molecular subtypes to be discovered, $\mathbf{W}_t \in \mathbb{R}^{D_t\times P}$ is the loadings matrix associated with dataset t and that maps the data to a lower dimensional space, and $\sigma_t^2 \in \mathbb{R}^+$ is the residual variance. Following [15], we place $\mathcal{N}(\mathbf{0}, \mathbf{I})$ prior on each row in $\mathbf{W}_t$.

We assume that the methylation observations are modelled as realisations of a multinomial distribution whose parameters are achieved through a softmax transformation of the linear projection of the latent factor vector:

$$p(\mathbf{x}_{i,me}|\mathbf{z}_i, \mathbf{W}_{1:D_{me}}, \mathbf{w}_{01:0D_{me}}) = \prod_{d=1}^{D_{me}} Cat(\mathbf{x}_{id,me}|S(\mathbf{W}_d^{\mathsf{T}}\mathbf{z}_i + \mathbf{w}_{0d})), \tag{2}$$

where $\mathbf{W}_d \in \mathbb{R}^{P\times M}$ is the loadings matrix for the dth response variable, $\mathbf{w}_{0d} \in \mathbb{R}^M$ is the offset term for the dth response variable and $M \in \mathbb{N}$ is the number of categories (in this case 2). We place $\mathcal{N}(\mathbf{0}, \mathbf{I})$ prior on every row in each loadings matrix $\mathbf{W}_d$ and $\mathcal{N}(\mathbf{0}, \mathbf{I})$ prior on the offsets $\mathbf{w}_{0d}$, similarly to [15].

3.2 Integrative Framework

We assume that the datasets we model share the same set of latent variables $\mathbf{Z}$, which represent the underlying subtypes or shared structure between the datasets. We can jointly estimate $\mathbf{Z}$ from the available datasets. We represent each of the glioblastoma datasets with a matrix, with the columns corresponding to the patients and the rows to the genes. For example, $\mathbf{X}_{ge}$, with dimensions $D_{ge} \times N$, denotes the gene expression data. By using (1) and (2), we can derive the mathematical form of the model which integrates the glioblastoma datasets as follows:

$$\begin{aligned}
\mathbf{X}_{ge} &= \mathbf{W}_{ge}\mathbf{Z} + \boldsymbol{\varepsilon}_{ge} \\
\mathbf{X}_{cnv} &= \mathbf{W}_{cnv}\mathbf{Z} + \boldsymbol{\varepsilon}_{cnv} \\
p(\mathbf{X}_{me}|\mathbf{W}_{1:D_{me}}, \mathbf{w}_{01:0D_{me}}) &= \prod_{i=1}^{N}\prod_{d=1}^{D_{me}} Cat(x_{id,me}|S(\mathbf{W}_d^{\mathsf{T}}\mathbf{z}_i + \mathbf{w}_{0d})) \qquad (3) \\
\mathbf{X}_{miRNA} &= \mathbf{W}_{miRNA}\mathbf{Z} + \boldsymbol{\varepsilon}_{miRNA},
\end{aligned}$$

where $\mathbf{W}_{ge}, \mathbf{W}_{cnv}, \mathbf{W}_1, \ldots, \mathbf{W}_{D_{me}}, \mathbf{W}_{miRNA}$ are the loading matrices, $\mathbf{w}_{01}, \ldots, \mathbf{w}_{0D_{me}}$ are the offset terms, and $\boldsymbol{\varepsilon}_{ge} \in \mathbb{R}^{D_{ge}\times N}$, $\boldsymbol{\varepsilon}_{cnv} \in \mathbb{R}^{D_{cnv}\times N}$, $\boldsymbol{\varepsilon}_{miRNA} \in \mathbb{R}^{D_{miRNA}\times N}$ are the remaining variances unique to each data type after accounting for correlation between data types. We place $\mathcal{N}(\mathbf{0}, \sigma_{ge}^2\mathbf{I})$, $\mathcal{N}(\mathbf{0}, \sigma_{cnv}^2\mathbf{I})$, $\mathcal{N}(\mathbf{0}, \sigma_{miRNA}^2\mathbf{I})$ priors on the error terms $\boldsymbol{\varepsilon}_{ge}, \boldsymbol{\varepsilon}_{cnv}$ and $\boldsymbol{\varepsilon}_{miRNA}$, respectively.

The key idea of the integrative framework is to reduce the high-dimensional datasets to a low-dimensional subspace which still captures the major data variations. We then model the lower dimensional representation using a finite approximation to a Dirichlet process mixture model [13], known as a Dirichlet-multinomial allocation mixture model [10]:

$$p(\mathbf{Z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{Z}|\boldsymbol{\mu}_k, \mathbf{I}), \qquad (4)$$

where π_k's are the mixing proportions, which are $\text{Dir}(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K})$ distributed and K is the number of clusters which we learn from the data. In (4), the latent variables $\mathbf{Z}$ are conditionally independent and identically distributed with a mixture density. We use cluster indicators c_i, with $c_i|\boldsymbol{\pi} \sim \text{Mult}(\boldsymbol{\pi})$, to indicate the cluster membership of the ith latent variable. We assume that the cluster means of latent variables $\boldsymbol{\mu}_k$ have a moment prior, introduced in [33]:

$$p(\boldsymbol{\mu}_k) = \frac{1}{C_k} \prod_{1\le i<j\le k} \frac{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^{\mathsf{T}}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}{g} \mathcal{N}(\boldsymbol{\mu}_k|\mathbf{0}, g\mathbf{I}), \qquad (5)$$

where C_k is the normalising constant and g is the dispersion parameter, which drives the separation between clusters. This prior leads to the identification of more clinically meaningful subtypes, stronger model parsimony and robustness even under model misspecification. The moment prior serves also as a base distribution for the Dirichlet Process prior in (4).

3.3 *Inference*

We perform inference with simulated annealing [17], which is a stochastic optimiser and can avoid getting stuck in local maxima. In addition, using simulated annealing will provide us with a near optimal solution much faster than the MCMC methods.

The main drawback of simulated annealing is that we get only MAP estimates of the model parameters rather than the full posterior distributions but since we are mainly interested in the final cluster partition, this is a worthy trade-off.

We have marginalised the mixing proportions $\boldsymbol{\pi}$ which reduces the number of parameters we need to infer. At the beginning of each iteration, we resample the cluster indicators $C = \{c_1, \ldots, c_N\}$ using Algorithm 8 from [27] as we work with non-conjugate priors. We then consider merging all points from small clusters (with fewer than 10 points) with the large cluster which they are most likely to join, and accept the move with a Metropolis Hastings ratio. This move is an example of split-merge MCMC [10, 14]. We investigated the idea of split-merge sampling because we often observed a tail of small clusters in our experiments which might be in part due to simulated annealing getting stuck in local modes.

We use simulated annealing after the split-merge move to infer the model parameters. Since it is closely related to the MCMC methods, we can easily adapt the steps involved in random walk Metropolis Hastings to perform simulated annealing instead. We proceed in the following manner: we propose a move from the current state S to S^*, where the latent variables $\mathbf{Z}$, the noise variables $\boldsymbol{\varepsilon}_t$, the loadings matrices for the discrete observations $\mathbf{W}_1 \ldots, \mathbf{W}_{D_{me}}$ and the offset terms $\mathbf{w}_{01}, \ldots, \mathbf{w}_{0D_{me}}$ have been updated using the same proposal distributions as we would use in random walk Metropolis Hastings, and the update of the loadings matrices $\mathbf{W}_t$ has been obtained using the approximation $\mathbf{X}_t \approx \mathbf{W}_t\mathbf{Z}$. After the proposal of the new state, we compute

$$r = \exp\left(\frac{f(S^*) - f(S)}{T_k}\right), \tag{6}$$

where f is the model log posterior and T_k is the current temperature of the cooling schedule. We move to the new state S^* and update the model parameters with probability $\min(1, r)$, otherwise we stay in the current state S and do not update model parameters. We use an exponential cooling schedule $T_k = T_0 s^k$ with a starting temperature of $T_0 = 100$ and $s = 0.95$. We fix the concentration parameter α to 3 as it leads to robust results [8]. The algorithm continues until convergence or a fixed number of iterations is reached.

3.4 Model Selection and Choice of Final Partition

Since the output of the algorithm depends on the dataset we choose to initialise the latent variables with, we run BayesCluster for all possible initialisation scenarios for a range of number of latent variables ($P = 2, \ldots, 10$) and we use Bayesian information criterion to select P. The final partition is the maximum a posteriori partition from the simulated annealing algorithm for the selected P.

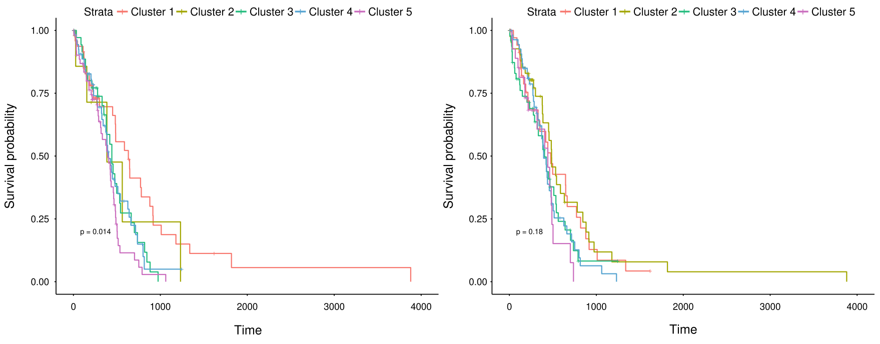
4 Results

We consider 5 ways of identifying cancer subtypes with BayesCluster—with each individual data type and with all 4 data types combined. We used clinical data, containing information about days to last follow-up, the patient treatments, age, gender, to further specify the subtypes. We compare all subtypes in terms of the right-censored event being death.

For each case, we plot Kaplan-Meier curves for the patient groups, identified by BayesCluster, after the removal of any patients with no follow-up. We test the null hypothesis that there is no difference between the subtypes in their survival prognosis, and the resulting unadjusted log-rank p-values are 0.4, 0.02, 0.66, 0.39, 0.014 for the subtypes identified using only gene expression, only copy number variation, only miRNA, only methylation and all data types, respectively. We have identified 5 glioblastoma subtypes using the information from all 4 datasets (see Fig. 1a). The heatmaps on Fig. 2 highlight the genomic differences between the 5 subtypes.

4.1 Comparison with Other Methods

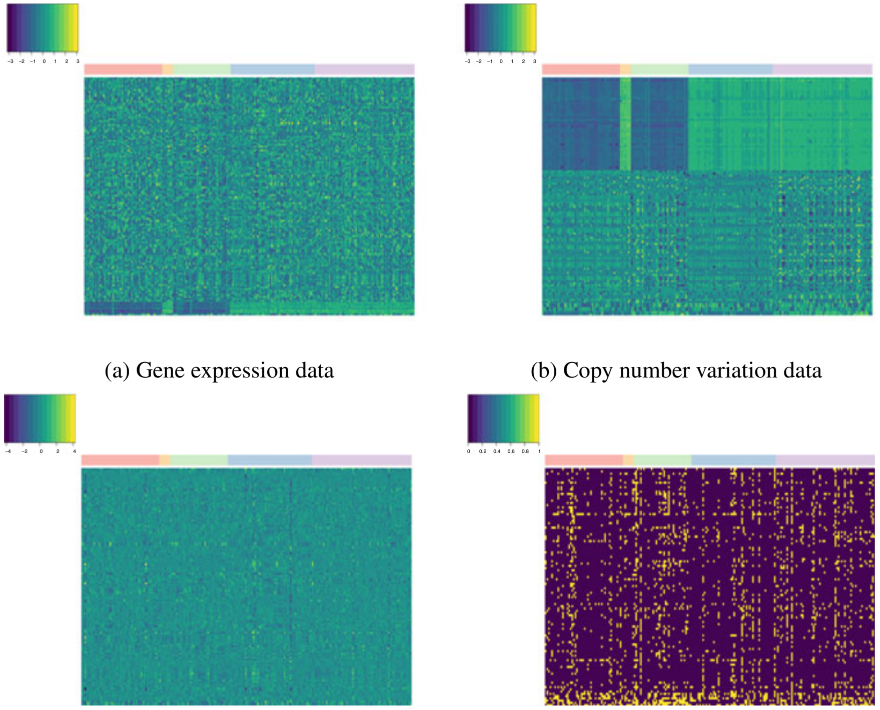
We compare the performance of BayesCluster with another integrative clustering method, called iClusterPlus [25]. iClusterPlus is a Gaussian latent variable model for genomic data integration and is an extension of iCluster [38]. It both reduces the dimensionality of the datasets and models the relations between the different types of data. Similarly to BayesCluster, it assumes that there is a common set of latent cluster membership variables across all datasets. Unlike BayesCluster, iClusterPlus uses a modified Monte Carlo Newton-Raphson algorithm [7, 23] and random walk Metropolis Hastings [32, 39] to learn the model parameters. The method uses the



(a) BayesCluster

(b) iClusterPlus

Fig. 1 Glioblastoma subtypes, identified using the integration of gene expression, methylation, copy number variation and miRNA data, with the application of BayesCluster and iClusterPlus



(a) Gene expression data

(b) Copy number variation data

(c) MiRNA data

(d) Methylation data

Fig. 2 Heatmaps of the glioblastoma data types, with the patients being on the x-axis, sorted by the integrative partition, and the selected features for each data type on the y-axis, sorted by hierarchical clustering with average linkage

deviance ratio metric, which is the percentage of the total variation explained by the model, for model selection. The optimal number of clusters is determined by the point of transition after which there is not a significant change in the deviance ratio.

The models we use to model the different data types are similar to the ones used in the Multiple Dataset Integration (MDI) model [34] as well. Both BayesCluster and MDI use split-merge to improve the mixing. In this paper we demonstrate that BayesCluster provides competitive results in comparison with iClusterPlus and simple clustering approaches. A more comprehensive comparison with other integrative methods including MDI can be found in [29].

iClusterPlus identifies 5 subtypes of glioblastoma patients using the information from all 4 datasets (Fig. 1b). The five groups do not have statistically different survival outcome, and there is no patient subtype that has a noticeably better survival prognosis than the rest.

We compare the performance of BayesCluster on each of the four individual data sources with k-means [11, 22] and Gaussian mixture model (GMM) [2, 24] in the

Table 1 Bonferroni-corrected p-values for Kaplan-Meier survival curves. As GMM resulted in one large cluster and singletons in the case of copy number variation and miRNA, we excluded these results. The p-value for k-modes was not adjusted as the model was applied only to methylation data. We have highlighted the unadjusted p-values with an asterisk

	GE	CNV	miRNA	ME	all
k-means/k-modes	0.6	0.06	1.00	0.59*	–
GMM	0.16*	–	–	–	–
iClusterPlus	0.95	0.225	1.00	0.036	0.9
BayesCluster	1.00	0.10	1.00	1.00	0.07

case of real-valued data, and in the case of discrete data, k-modes [4]. For each of the datasets, k-means/k-modes and GMM are not able to capture the difference between the patient subtypes. This suggests that using the information from multiple glioblastoma data sources could identify more clinically meaningful subtypes.

We summarise the comparison between the models in Table 1. We have applied Bonferroni correction in the cases where we test multiple times the null hypothesis of no difference between the groups in their survival.

5 Conclusion

We presented BayesCluster, a novel Bayesian method for unsupervised modelling of individual datasets and for integrative modelling of multiple datasets, which has the advantages of inferring the number of clusters from the data and of easily implementable inference. Its framework is based on the Dirichlet Process mixture model which enables the modelling of heterogeneous data types and the subtype allocation of arriving new patients, using the predictive distribution from [9].

We applied BayesCluster to glioblastoma data and managed to identify specific subtypes that are prognostic of survival outcome and could be used for more targeted treatments. We plan to validate these subtypes with an independent dataset, part of the ICGC database. Since there were cases where BayesCluster could not find meaningful subtypes, we will consider improving the feature selection and interpretability by incorporating estimation of the posterior probability of each omics feature. This can be used as a criterion for feature selection as suggested by [26].

References

1. Barash, Y., Friedman, N.: Context-specfic Bayesian clustering for gene expression data. J. Comput. Bio. **9**, 169–191 (2002)
2. Bishop, C.: Pattern Recognition and Machine Learning. Springer (2006)

3. Blei, D.M., Kucukelbir, A., McAuliffe, J.D.: Variational inference: a review for statisticians. J. Am. Stat. Assoc. **112**, 859–877 (2017)
4. Chaturvedi, A., Green, P., Caroll, J.D.: K-modes clustering. J. Classif. **18**, 35–55 (2001)
5. Curtis, C., Shah, S.P., Chin, S., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., Yuan, Y.: The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature **486**, 343 (2012)
6. Filkov, V., Skiena, S.: Heterogeneous data integration with the consensus clustering formalism. In: International Workshop on Data Integration in the Life Sciences, pp. 110–123. Springer (2004)
7. Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. **33**, 1 (2010)
8. Fúquene, J., Steel, M., Rossell, D.: On Choosing Mixture Components via Non-local Priors. J. R. Stat. Society. Ser B **81**, 809–837 (2019)
9. Görür, D., Rasmussen, C.E.: Dirichlet process Gaussian mixture models: choice of the base distribution. J. Comput. Sci. Technol. **25**, 653–664 (2010)
10. Green, P.J., Richardson, S.: Modelling heterogeneity with and without the Dirichlet process. Scand. J. Stat. **28**, 355–375 (2001)
11. Hartigan, J.A., Wong, M.A.: Algorithm AS 136: A k-means clustering algorithm. J. R. Stat. Soc. Ser. C (App. Stat) **28**, 100–108 (1979)
12. International Cancer Genome Consortium: International network of cancer genome projects. Nature **464**, 993 (2010)
13. Ishwaran, H., Zarepour, M.: Exact and approximate sum representations for the Dirichlet process. Can. J. Stat. **30**, 269–283 (2002)
14. Jain, S., Neal, R.M.: A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. J. Comput. Graph. Stat. **13**, 158–182 (2004)
15. Khan, M.E., Bouchard, G., Murphy, K.P., Marlin, B.M.: Variational bounds for mixed-data factor analysis. In: Advances in Neural Information Processing Systems, pp. 1108–1116 (2010)
16. Kirk, P., Griffin, J.E., Savage, R.S., Ghahramani, Z., Wild, D.L.: Bayesian correlated clustering to integrate multiple datasets. Bioinformatics **28**, 3290–3297 (2012)
17. Kirkpatrick, S., Gelatt, D.C., Vecchi, M.P.: Optimization by simulated annealing. Science **220**, 671–680 (1983)
18. Klami, A., Jitta, A.: Probabilistic size-constrained microclustering. In: UAI Proceedings (2016)
19. Kormaksson, M., Booth, J.G., Figueroa, M.E., Melnick, A.: Integrative model-based clustering of microarray methylation and expression data. Ann. App. Stat. **6**, 1327–1347 (2012)
20. Lee, Y., Lee, J., Ahn, S.H., Lee, J., Nam, D.: WNT signaling in glioblastoma and therapeutic opportunities. Nature **96**, 137 (2016)
21. Liu, X., Sivaganesan, S., Yeung, K.Y., Guo, J., Bumgarner, R.E., Medvedovic, M.: Context-specific infinite mixtures for clustering gene expression profiles across diverse microarray dataset. Bionformatics **22**, 1737–1744 (2006)
22. MacQueen, J.: Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability **1**, 281–297 (1967)
23. McCulloch, C.E.: Maximum likelihood algorithms for generalized linear mixed models. J. Am. Stat. Assoc. **92**, 162–170 (1997)
24. McLachlan, G., Peel, D.: Finite Mixture Models. Wiley, New York (2004)
25. Mo, Q., Wang, S., Seshan, V.E., Olshen, A.B., Schultz, N., Sander, C., Powers, S.R., Ladanyi, M., Shen, R.: Pattern discovery and cancer gene identification in integrated cancer genomic data. P. Nath. A. Sci. **110**, 4245–4250 (2013)
26. Mo, Q., Shen, R., Guo, C., Vannucci, M., Chan, K., Hilsenbeck, S.: A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. Biostatistics **19**, 71–86 (2017)
27. Neal, R.M.: Markov chain sampling methods for Dirichlet process mixture models. J. Comput. Graph. Stat. **9**, 249–265 (2000)

28. Onogi, A., Nurimoto, M., Morita, M.: Characterization of a Bayesian genetic clustering algorithm based on a Dirichlet process prior and comparison among Bayesian clustering methods. BMC Bioinform. **12**, 263 (2011)
29. Peneva, I., Savage, R.S.: Identifying cancer subtypes using Bayesian data integration. In preparation
30. Quiroz, M., Kohn, R., Villani, M., Tran, M.: Speeding up MCMC by efficient data subsampling. J. Am. Stat. Assoc. **114**, 831–843 (2019)
31. Rasmussen, C.E.: The infinite Gaussian mixture model. In: Advances in Neural Information Processing Systems, pp. 554–560 (2000)
32. Robert, C.P., Casella, G.: The Metropolis-Hastings algorithm. In: Monte Carlo Statistical Methods, pp. 231–283 Springer (1999)
33. Rossell, D., Telesca, D.: Nonlocal priors for high-dimensional estimation. J. Am. Stat. Assoc. **112**, 254–265 (2017)
34. Savage, R.S., Ghahramani, Z., Griffin, J.E., Kirk, P., Wild, D.L.: Identifying cancer subtypes in glioblastoma by combining genomic, transcriptomic and epigenomic data. In: International Conference on Machine Learning (ICML) 2012: Workshop on Machine Learning in Genetics and Genomics (2013)
35. Savage, R.S., Ghahramani, Z., Griffin, J.E., De La Cruz, B.J., Wild, D.L.: Discovering transcriptional modules by Bayesian data integration. Bioinformatics **26**, i158–i167 (2010)
36. Schwarz, G., et al.: Estimating the dimension of a model. Ann. Stat. **6**, 461–464 (1978)
37. Shaweis, H., Han, C., Sivasubramiam, V., Brazil, L., Beaney, R., Sadler, G., Al-Sarraj, S., Hampton, T., Logan, J., Hurwitz, V.: Has the survival of patients with glioblastoma changed over the years? Brit. J. Can. **114**, 146 (2016)
38. Shen, R., Olshen, A.B., Ladanyi, M.: Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioninformatics **25**, 2906–2912 (2009)
39. Sherlock, C., Fearnhead, P., Roberts, G.O.: The random walk Metropolis: linking theory and practice through a case study. Stat. Sci. **25**, 172–190 (2010)
40. Suchard, M.A., Wang, Q., Chan, C., Frelinger, J., Cron, A., West, M.: Understanding GPU programming for statistical computation: studies in massively parallel massive mixtures. J. Comput. Graph. Stat. **19**, 419–438 (2010)
41. TCGA: Comprehensive molecular portraits of human breast tumours. Nature **490**, 61–70 (2012)
42. Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M.: The Cancer Genome Atlas pan-cancer analysis project. Nat. Gen. **45**, 1113 (2013)
43. West, M., Escobar, M.D.: Hierarchical Priors And Mixture Models, With Application In Regression And Density Estimation. Institute of Statistics and Decision Sciences, Duke University (1993)
44. World Health Organisation: Cancer key facts (2018). http://www.who.int/en/news-room/fact-sheets/detail/cancer (Cited: 15 Jan 2019)

Bayesian Inference of Interaction Effects in Item-Level Hierarchical Twin Data

Inga Schwabe

Abstract Behavior geneticists are interested in the relative importance of genetic and environmental influences in the origins of individual differences in a trait (phenotype). Considerable effort has been devoted to analyses including interactions between these different sources, such as gene-environment interactions. To measure the phenotype, usually, a questionnaire is presented to monozygotic (MZ) and dizygotic (DZ) twins and the resulting sum-scores are used as proxy measures for the phenotype in the genetic model. However, earlier research has shown that using sum-scores can lead to the spurious finding of interactions and, instead, an approach based on raw item data should be adopted. This can be done by simultaneously estimating the genetic twin model and an item response theory (IRT) measurement model. Due to the hierarchical nature of twin data, this is difficult to implement in the frequentist framework. As an alternative, we can adopt the Bayesian framework and use off-the-shelf MCMC methods. This chapter contains an overview of this methodology, including different parametrizations of interaction terms. To illustrate the methodology, the depression scores of 364 MZ twin pairs and 585 DZ twin pairs are analyzed to investigate if depression is etiologically different in older ($>$60 years) twins.

Keywords Behavior genetics · Hierarchical twin data · Item response theory · MCMC · Psychometrics

1 Introduction

Inference in the field of behavior genetics focuses on determining the relative contribution of nature and nurture to individual differences in a trait (e.g., the phenotype). One of the most commonly used methods in this field is the classical twin design, which makes use of the fact that twin pairs are either identical (monozygotic, MZ) and share the same genomic sequence or non-identical (dizygotic, DZ) and share, on average, only half of the segregating genes. When MZ twin pairs are more similar in

I. Schwabe (✉)
Tilburg University, Warandelaan 2, 5037 AB Tilburg, The Netherlands
e-mail: I.Schwabe@uvt.nl

R. Argiento et al. (eds.), *Bayesian Statistics and New Generations*,
Springer Proceedings in Mathematics & Statistics 296,
https://doi.org/10.1007/978-3-030-30611-3_12

a phenotype (e.g., depression or educational achievement) than DZ twin pairs, this implies that genetic influences are important.

Traditionally, in the classical twin design, the ACE model is used which decomposes total phenotypic variance (e.g., scores on a mathematics test or a depression scale), σ_P^2, into variance due to additive genetic (A) influences (σ_A^2), common-environmental (C) influences (σ_C^2) and unique-environmental (E) influences (σ_E^2, residual variance). Additive genetic (A) influences are parametrized as latent variables that are perfectly correlated within MZ twin pairs and correlate, on average, $\frac{1}{2}$ within DZ twin pairs. Common-environmental influences are parametrized as being perfectly correlated within both MZ and DZ twin pairs and interpreted as shared influences that lead to resemblance between twins that cannot be attributed to their genetic resemblance. Contrarily to many other research fields, residual variance (denoted by σ_E^2 in the twin design) is not used for model-fit purposes, but interpreted as reflecting not only measurement error, but everything that is not related to genetic differences that makes a twin par different (e.g., unique-environmental influences, E).

In the last decade, the assessment of an interaction between these different sources of variance has received increasing attention. An interaction can be estimated between any combination of the latent variables (e.g., A, C, and E), but we can also estimate a moderation of variance components by adding a *measured* environmental moderator variable to the model. While the former parametrization usually concentrates on estimating an interaction between latent *genetic* influences and latent unique-environmental influences, the latter design is often used to estimate all different kinds of interactions (e.g., including interactions between the moderator variable(s) and latent environmental influences).

1.1 Potential Bias

When the twin design is applied, traditionally, the twins' answers to every item of a questionnaire or test are aggregated into a single score (e.g., a sum-score) and then used as a proxy measure for the phenotype. However, earlier research has shown that the use of a sum-score can lead to the spurious finding of interaction effects due to artefacts of the given measurement scale, such as heterogeneous measurement error. Instead, an approach based on the analysis of raw item data should be adapted which can be done by simultaneously estimating the genetic model and an item response theory (IRT) measurement model [5, 9, 11]. The IRT approach is model-based measurement in which a twin's latent trait is estimated using not only trait levels (e.g., performance on a mathematics test) but also item properties such as the difficulty of each test item.

1.2 Bayesian Approach

To take full advantage of the IRT approach, both the genetic twin model and the IRT measurement model have to be estimated simultaneously [13], which is not possible when using the currently available software [1, 6, 8]. The simultaneous estimation of the genetic twin model and IRT model is challenging since it requires the evaluation of multiple integrals. This problem can be avoided by adopting a Bayesian approach and using Markov chain Monte Carlo (MCMC) algorithms to approximate the posterior distribution. In the following, an overview of a Bayesian approach to the estimation of gene-environment interactions is given that enables item-level analysis of the twin data by integrating an IRT model into the genetic model [5, 9, 11].

2 Methods

2.1 ACE Model

We assume that part of the variance in the latent phenotypic trait θ_{ij} of twin j from pair i is shared with the co-twin and part of it is unique to the individual (e.g., explained by unique-environmental influences (E), captured by residual variance σ_E^2) [4]. Furthermore, we assume that the total shared variance is partly genetic (e.g., explained by additive genetic influences, represented by A) and partly environmental (e.g., explained by common-environmental factors, represented by C). We then have for MZ twins:

$$\begin{aligned} A_i &\sim N(0, \sigma_A^2) \\ C_i &\sim N(0, \sigma_C^2) \\ \theta_{ij} &\sim N(\mu + A_i + C_i + \beta_{1s}\boldsymbol{M}_i, \sigma_E^2) \end{aligned}$$

where β_{1s} denotes the main effect of a moderator variable that has the same value for every twin family i (see the following for more detail).

As it is generally observed that a hierarchical parametrization leads to more efficient estimation, we use the following (equivalent) parametrization, where the familial random effect F (e.g., the sum of C and A) of every twin family i is a deviation from the random effect C of the same family i:

$$\begin{aligned} C_i &\sim N(\mu + \beta_{1s}\boldsymbol{M}_i, \sigma_C^2) \\ F_i &\sim N(C_i, \sigma_A^2) \\ \theta_{ij} &\sim N(F_i, \sigma_E^2) \end{aligned}$$

where μ refers tot the phenotypic population mean, σ_C^2 represents variance explained by common-environmental influences, σ_A^2 denotes variance that can be explained by additive genetic influences and σ_E^2 is the residual variance for each individual twin. For every family i, the genetic part of the phenotype, A_i, can be calculated by subtracting the part of the familial random effect that is due to common-environmental influences (e.g., $A_i = F_i - C_i$).

The parametrization is the same for DZ twins, except that they are not genetically identical, but share on average 50% of the segregating genes. This genetic correlation of $\frac{1}{2}$ is modeled by assuming normally distributed familial random effects for every twin family i first (e.g., $F1_i \sim N(C_i, \frac{1}{2}\sigma^2_A)$) and then using these as expected value when modeling the familial effect of every individual twin j of family i (e.g., $F2_{ij} \sim N(F1_i, \frac{1}{2}\sigma^2_A)$).

2.2 Modeling Interaction Effects

We can take two different approaches to integrate interaction terms into the above described twin design.

First, we can model any interaction(s) among the *latent* variables (e.g., A, C, and E). As this is common practice, here, we concentrate on estimating an interaction between additive genetic influences and unique-environmental influences (e.g., A×E), but note that it is also possible to include other interaction terms (e.g., A×C or C×E). A×E can be modeled by varying the amount of variance due to unique-environmental variance systematically with genetic value A, resulting in a variance term that is unique for every individual and is portioned into a part that denotes an intercept (unique-environmental variance when $A_{ij} = 0$) and a slope parameter that represents the interaction effect (e.g. $\sigma^2_{Eji} = \exp(\beta_0 + \beta_1 A2_{ij})$ for DZ twins where β_0 denotes the intercept and β_1 the interaction term). Note that A×E is then modelled as (log)linear effect, meaning that unique-environmental variance is larger at either higher or lower levels of the genetic value where the sign of the slope determines the direction of the interaction effect. The exponential function is used to avoid negative variances.

While modeling environmental influences as latent variables provides a nice and strong omnibus test, often, one or more moderator variable(s) are used to investigate interactions with *specific* environmental influences. This makes results potentially very informative. Often, this alternative parametrization is not only used to estimate interaction effects with additive genetic influences (henceforth referred to as A×M), but also with common-environmental influences (henceforth referred to as C×M) or unique-environmental influences (henceforth referred to as E×M) - that is, moderation of variance components (henceforth referred to as ACE×M). For readability, the ACE×M parametrization presented here discusses the case where we the moderator variable takes the same value for both members of a twin pair [11]. To model ACE×M, variance components are divided into an intercept (representing variance components when $\boldsymbol{M}_i = 0$) and a linear interaction term (denoting A×M, C×M and E×M respectively). For MZ twins, we then have:

$$\begin{aligned}
\sigma^2_{Ai} &= \exp(\beta_{0a} + \beta_{1a}\boldsymbol{M}_i) \\
\sigma^2_{Ci} &= \exp(\beta_{0c} + \beta_{1c}\boldsymbol{M}_i) \\
\sigma^2_{Ei} &= \exp(\beta_{0e} + \beta_{1e}\boldsymbol{M}_i)
\end{aligned}$$

where β_{0a}, β_{0c} and β_{0e} represent additive genetic, common-environmental and unique-environmental variance respectively when $\boldsymbol{M}_i = 0$. β_{1a}, β_{1c} and β_{1e} are the interaction terms that represent A×M, C×M and E×M respectively.

Similar to the model specification for MZ twins, C×M and E×M are introduced by allowing common-environmental and unique-environmental influences to vary between DZ families (e.g., $\sigma^2_{Ci} = \exp(\beta_{0c} + \beta_{1c}\boldsymbol{M}_i)$ and $\sigma^2_{Ei} = \exp(\beta_{0e} + \beta_{1e}\boldsymbol{M}_i)$). Then, to introduce A×M, the genetic value of every individual twin, $A2_{ij}$, is scaled by multiplying it with the standard deviation σ_{Ai}, where $\sigma^2_{Ai} = \exp(\beta_{0a} + \beta_{1a}\boldsymbol{M}_i)$. This yields a genetic value $A3_{ij}$ that is unique for every individual twin j from DZ family i:

$$A3_{ij} = A2_{ij}\sqrt{\exp(\beta_{0a} + \beta_{1a}\boldsymbol{M}_i)}$$

where, similar to the model specification of DZ twins, β_{0a} represents additive genetic variance when $\boldsymbol{M}_i = 0$ and $\beta_{1a}$is the interaction term that represents A×M.

2.3 IRT Modeling

For readability, so far, we have ignored the psychometric part in the genetic models discussed above. Simultaneously to every genetic model, the latent phenotype, θ_{ij} appears in an IRT measurement model. The most simple IRT model is the so-called Rasch model, which models the probability of a correct answer to item k (e.g., of a mathematics test) by twin j from family i as a logistic function of the difference between the twin's latent trait score (e.g., mathematical ability) and the difficulty of the item:

$$\ln(P_{ijk}/(1 - P_{ijk})) = \theta_{ij} - b_k$$
$$Y_{ijk} \sim Bernoulli(P_{ijk})$$

where Y_{ijk} represents the answer of twin j of family i to item k which are assumed to have a Bernoulli distribution. It is assumed that all items discriminate equally well between varying traits, but the model can be extended to also include discrimination parameters (e.g., comparable to factor loadings in structural equation modeling) that differ between items [3]. The Rasch model and its extension are suitable for dichotomous data as is often encountered in cognitive tests (e.g., scored as correct = 1 and false = 0) and sometimes in personality questionnaires where a respondent is asked if he or she agrees with a particular statement (e.g., scored as agree = 1, 0 = disagree). For ordinal item data, as encountered in for example Likert scale data, we can use among others the generalized partial credit model (GPCM), which treats polytomous response categories (e.g., Likert scale data) as ordered performance levels.

3 Data Illustration: Depression Scores

To illustrate how the methodology can be used to analyze empirical twin data, twin responses to a seven item depression scale were analyzed to investigate if depression is etiologically different for older twins.

The scale was part of a telephone questionnaire that was administered in the National Survey of Midlife Development in the United States (MIDSU) in 1995–1996 under the auspices of the Inter-university Consortium for Political and Social Research [2].[1] Each item of this scale consists of a statement (e.g., "I lose interest in most things"), followed by the instruction of the interviewer to indicate if the respondent agreed with the statement or not. The data of twins with unknown zygosity (N twins = 25) were excluded from the analysis, resulting in a total sample size of 364 MZ twin pairs and 585 DZ twin pairs. One MZ twin pair and five DZ twin pairs with non-matching ages were excluded from the sample, further reducing the sample size to 363 MZ twin pairs and 580 DZ twin pairs. When the age of the first twin was missing, the age of the second twin was imputed (and vice versa). For a total of 98 twin pairs, the age was not recorded (missing). Mean age of the whole sample was 44 (SD = 12) and 45 (SD = 12) for MZ and DZ twin pairs respectively. For an easier interpretation, the continuous age variable was transformed into a dichotomous dummy variable that cut the sample in two distinct age groups, coded as 0 (all twin pairs younger than 60 years) and 1 (all twin pairs that were 60 years or older). 724 (77%) twin pairs were younger than 60 years and 121 (13%) 60 years old or older.

We fitted an ACE model with Rasch model that included all possible interactions between latent variables and age group (e.g., A×M, C×M and E×M, where M refers to the dichotomous variable that divides twins in the two age groups). We used independent normal distributions for intercepts and interaction effects (e.g., β_{0a}, β_{0c} and $\beta_{0e} \sim N(-1, 2)$, β_{1a}, β_{1c} and $\beta_{1e} \sim N(0, 10)$) as well as for difficulty parameters (e.g., $\beta_k \sim N(0, 10)$) and the main effect of the moderator variable (e.g., $\beta_{1m} \sim N(0, 10)$). As there were twin pairs with unknown age, independent Bernoulli distributed prior distributions were defined (e.g., for every twin family i, $\boldsymbol{M}_i \sim \boldsymbol{Bernoulli}(\pi)$). On the probability, π, separate Beta distributed hyperpriors were used for MZ and DZ twins respectively (e.g., $\pi_{mz} \sim \boldsymbol{Beta}(1, 1)$ and $\pi_{dz} \sim \boldsymbol{Beta}(1, 1)$).

After a burn-in period of 15,000 iterations, the posterior distribution was based on an additional 20,000 iterations from 1 Markov chain. For data handling and MCMC estimation, the open-source software packages R [8] and JAGS [7] were used. For all parameters of interest, posterior means and standard deviations were calculated as was the 95% highest posterior density (HPD) interval. When the HPD does not contain zero, the influence of a parameter can be regarded as significant. This however does not hold for the variance components of this particular application, as these are bounded at zero due to a very low phenotypic variance.

[1] The opinions expressed in this article are those of the authors and do not necessarily reflect the views of the ICPSR.

Table 1 Data illustration: depression data. Estimates of variance components and heritability, based on the ACE model with integrated Rasch model. Total phenotypic variance was 4.963. HPD refers to highest posterior density interval

	Posterior point estimate (SD)	95% HPD
β_{1s}	0.019 (0.001)	[0.017; 0.021]
$\exp(\beta_{0a})$	2.231 (0.605)	[1.002; 3.266]
$\exp(\beta_{0c})$	1.572 (0.484)	[1.000; 2.508]
$\exp(\beta_{0e})$	1.160 (0.170)	[1.000; 1.532]
β_{1a}	−0.014 (0.006)	[−0.027; −0.002]
β_{1c}	−0.009 (0.006)	[−0.022; 0.001]
β_{1e}	0.003 (0.003)	[−0.004; 0.007]
h^2	0.44 (0.093)	[0.247; 0.608]

3.1 Results

The posterior means and standard deviations of all parameters as well as narrow-sense heritability can be found in Table 1. Narrow-sense heritability was defined as the relative magnitude of the total phenotypic variance that can be explained by additive genetic variance (e.g., $h^2 = \frac{\exp(\beta_{0a})}{\sigma_P^2} = \frac{\exp(\beta_{0a})}{(\exp(\beta_{0a})+\exp(\beta_{0c})+\exp(\beta_{0e}))}$). In the second column of the table, the 95% highest posterior density interval can be found. To save space, estimates of item parameters as well as convergence plots are not displayed here but can be obtained from the author.

It can be seen that most of the phenotypic variance can be explained by additive genetic influence. While common-environmental influences also contribute to phenotypic variance, the influence of unique-environmental influences is negligibly small. HPD intervals show that none of the interaction effects was significant.

4 Discussion

This contribution provides an overview of a Bayesian framework that makes it possible to estimate interactions in genetically-informative item-level twin data through estimating both genetic and measurement (IRT) model simultaneously. The incorporation of such a measurement model into genetic twin analyses is important, since it has been shown that analyzing an aggregated measure such as the sum-score can result in the spurious finding of an interaction effect due to measurement scale artefacts [5, 9, 11]. The models presented here can be extended to a longitudinal version when the phenotypic trait has been assessed at multiple occasions, which also enables inference about the nature (e.g., genetic or environmental) of the covariance among the different measurement points [12].

To illustrate the methodology, the scores on a 7 item depression scale of 364 MZ twin pairs and 585 DZ twin pairs were analyzed. Results showed that genetic influences were the most important source in explaining phenotypic variance. Furthermore, no interaction term was significant, implying that depression is not etiologically different in older people.

To make the methodology presented here more accessible to applied researcher in the behavior genetics community, the R package *BayesTwin* was developed, which includes all models presented here and some extensions. A tutorial for this package can be found elsewhere [10].

References

1. Boker, S.M., Neale, M.C., Maes, H.H., Wilde, M.J., Spiegel, M., Brick, T.R., Spies, J., Estabrook, R., Kenny, S., Bates, T.S., Mehta, P., Fox, J.: OpenMx: an open source extended structural equation modeling framework. Psychometrika **76**, 306–317 (2011)
2. Brim, O.G., Baltes, P.B., Bumpass, L.L., Cleary, P.D., Featherman, D.L., Hazard, W.R.: National survey of midlife development in the United States (MIDUS). Inter-university Consortium for Political and Social Research, Ann Arbor (2010)
3. Embretson, S.E., Reise, S.P.: Item Response Theory for Psychologists. Psychology Press, New Jersey (2009)
4. Jinks, J.L., Fulker, D.W.: Comparison of the biometrical genetical, mava, and classical approaches to the analysis of human behavior. Psych. Bull. **73**, 311–349 (1970)
5. Molenaar, D., Dolan, C.V.: Testing systematic genotype by environment interactions using item level data. Beh. Genet. **44**, 212–231 (2014)
6. Muthen, L., Muthen, B.O.: Mplus User's Guide, 6th edn. Los Angeles (1998–2017)
7. Plummer, M.: JAGS: a program for analysis of bayesian graphical models using Gibbs sampling. In: Proceedings of the 3rd International Workshop on Distributed Statistical Computing (2003)
8. R Development Core Team: R: A language and environment for statistical computing. Vienna, Australia (2008)
9. Schwabe, I., van den Berg, S.M.: Assessing genotype by environment interaction in case of heterogeneous measurement error. Behav. Genet. **44**, 394–406 (2014)
10. Schwabe, I.: An R package for bayesian inference of item-level twin data. JOSS. **5** (2017)
11. Schwabe, I., Boomsma, D.I., van den Berg, S.M.: IRT modelling to estimate genotype by environment interaction. Twin Res. Hum. Genet. **6**, 511–520 (2017)
12. Schwabe, I., Gu, Z., Tijmstra, J., Hatemi, P., Martin, N.G., Pohl, S.: Psychometric modelling of longitudinal genetically-informative twin data. Frontiers in Genetics, Accepted subject to minor revisions (2019)
13. van den Berg, S.M., Glas, C.A.W., Boomsma, D.I.: Variance decomposition using an IRT measurement model. Beh Genet. **37**, 604–616 (2007)

Applied Statistics

A Phase II Clinical Trial Design for Associated Co-primary Efficacy and Toxicity Outcomes with Baseline Covariates

Kristian Brock, Lucinda Billingham, Christina Yap and Gary Middleton

Abstract The experimental design presented here is motivated by a phase II clinical trial called PePS2, investigating the efficacy and safety of an immunotherapy called pembrolizumab in a specific subgroup of lung cancer patients. Previous trials have shown that the probability of efficacy is correlated with particular patient variables. There are clinical trial designs that investigate co-primary efficacy and toxicity outcomes in phase II, but few that incorporate covariates. We present here the approach we developed for PePS2, latterly recognised to be a special case of a more general method originally presented by Thall, Nguyen and Estey. Their method incorporates covariates to conduct a dose-finding study but has been scarcely used in trials. Dose-finding is not required in PePS2 because a candidate dose has been widely tested. Starting from the most general case, we introduce our method as a novel refinement appropriate for use in phase II, and evaluate it using a simulation study. Our method shares information across patient cohorts. Simulations show it is more efficient than analysing the cohorts separately. Using the design in PePS2 with 60 patients to test the treatment in six cohorts determined by our baseline covariates, we can expect error rates typical of those used in phase II trials. However, we demonstrate that care must be taken when specifying the models for efficacy and toxicity because more complex models require greater sample sizes for acceptable simulated performance.

Keywords Covariate · Efficacy · Phase ii · Toxicity · Trial

K. Brock (✉) · L. Billingham · C. Yap · G. Middleton
University of Birmingham, B15 2TT, Birmingham, UK
e-mail: k.brock@bham.ac.uk

L. Billingham
e-mail: l.j.billingham@bham.ac.uk

C. Yap
e-mail: c.yap@bham.ac.uk

G. Middleton
e-mail: g.middleton@bham.ac.uk

R. Argiento et al. (eds.), *Bayesian Statistics and New Generations*,
Springer Proceedings in Mathematics & Statistics 296,
https://doi.org/10.1007/978-3-030-30611-3_13

1 Introduction

There is a relative dearth of phase II clinical trial designs that incorporate patient covariates to assess efficacy and toxicity. We introduce a novel approach here.

Our motivation is a phase II trial called PePS2 that investigates an immunotherapy in a specific subgroup of lung cancer patients. We developed a Bayesian regression method that adjusts for predictive patient data available at trial commencement to investigate co-primary binary outcomes. We latterly learned that our design is a special case of Thall, Nguyen and Estey (TNE), a family of methods that perform dose-finding trials guided by efficacy and toxicity outcomes whilst accounting for baseline patient data [17]. Their design yields personalised dose recommendations.

PePS2 is not a dose-finding trial. Instead, it seeks to estimate the probabilities of efficacy and toxicity at a dose of pembrolizumab previously demonstrated to be safe and effective in a closely-related group of patients [9]. To acknowledge its heritage, we introduce our design as a novel simplification of TNE that removes the dose-finding components so that it may be used in phase II.

In Sect. 2, we describe the PePS2 trial and the pertinent clinical data from previous trials. In Sect. 3, we review the literature for suitable experimental designs. We describe our design in detail in Sect. 4 and evaluate it with a simulation study in Sect. 5. Finally, in Sect. 6, we describe future plans for this work.

2 The Clinical Trial Scenario

PePS2 is a phase II trial of pembrolizumab in non-small-cell lung cancer (NSCLC) patients with Eastern Cooperative Oncology Group performance status 2 (PS2). NSCLC is a common sub-type of lung cancer. Patients with PS2 are ambulatory and capable of self-care but typically too ill to work. Critically, it is doubtful that a PS2 patient could tolerate the toxic side effects of chemotherapy.

The primary objective of the trial is to learn if pembrolizumab is associated with sufficient disease control and tolerability to justify use in PS2 patients. The joint primary outcomes are (i) *toxicity*, defined as the occurrence of a treatment-related dose delay or treatment discontinuation due to adverse event related to pembrolizumab; and (ii) *efficacy*, defined as the occurrence of stable disease, partial response (PR) or complete response (CR), without prior progressive disease, at or after the second post-baseline disease assessment by version 1.1 of the *Response Evaluation Criteria In Solid Tumors* [8]. The second assessment is scheduled to occur at week 18.

Pembrolizumab inhibits the programmed cell death 1 (PD-1) receptor via the programmed death-ligand 1 (PD-L1) protein. It has been shown to be active and tolerable in patients with better performance status [9]. Overall, 19.4% of patients had an objective response (PR or CR) and 9.5% experienced a major adverse event, defined as an event of at least grade 3 by the *Common Terminology Criteria for Adverse Events*, v4.0. These statistics compare favourably to those typically seen in advanced NSCLC patients under chemotherapy [1, 13]. We foresee no reason why they should be materially different in PS2 patients.

Table 1 Objective response rates for the validation sample ($n = 204$) in [9]

PD-L1 Group	Criteria	Objective Response %, (95% CI)
Low	PD-L1 score $<$ 1%	10.7 (2.3, 28.2)
Medium	1% $\geq$ PD-L1 score $<$ 50%	16.5 (9.9, 25.1)
High	PD-L1 score $\geq$ 50%	45.2 (33.5, 57.3)

Table 2 Cohorts used in the PePS2 trial. x_i shows the predictive variable vector for patient i

Cohort	Previous treatment status	PD-L1 category	$x_i = (x_{1i}, x_{2i}, x_{3i})$
1	TN	Low	(0,1,0)
2	TN	Medium	(0,0,1)
3	TN	High	(0,0,0)
4	PT	Low	(1,1,0)
5	PT	Medium	(1,0,1)
6	PT	High	(1,0,0)

Garon et al. introduce the PD-L1 proportion score biomarker, defined as the percentage of neoplastic cells with staining for membranous PD-L1 [9]. Efficacy outcomes for the 204 patients in their validation group, summarised by PD-L1 score, are shown in Table 1. Objective responses are observed in all cohorts and the rate increases with PD-L1. Based on this information, we expect PD-L1 to be predictive of response in our PS2 population.

Furthermore, 24.8% of patients who had received no previous anti-cancer therapy (treatment-naive, TN) achieved a response, compared to 18.0% in the group that had been previously treated (PT) [9]. This represents a potentially small but important effect that should be considered when testing the treatment. We propose to investigate pembrolizumab by jointly stratifying by the three Garon PD-L1 groups, and PT and TN statuses. Each patient will belong to exactly one of six cohorts, as demonstrated in Table 2.

In phase II, there is strong motivation to deliver findings quickly to inform the next study phase. Recruitment of approximately 60 PS2 patients within one year would be feasible but accrual materially higher would be unlikely. Given the relative dearth of treatment alternatives, we seek to offer the trial to all PS2 patients and not stratify accrual. Pembrolizumab has not been investigated in PS2 patients so the clinical scenario requires a trial design that tests efficacy and toxicity. Given the evidence that PD-L1 and pretreatedness are associated with response, it is highly desirable to use a trial design that incorporates this predictive information. The next section describes our search for a clinical trial design to achieve these objectives.

3 Review of Available Trial Designs

We sought a clinical trial design that uses covariates to study co-primary binary outcomes. The well-known phase II design by Bryant and Day (BD) takes threshold rates of efficacy and toxicity and returns the number of events to approve the treatment

[4]. For given levels of significance and power, the thresholds identify the optimal trial of the competing outcomes. The design does not use covariates, assuming the population to be homogeneous. Parallel BD designs in our six cohorts would require a prohibitively large total sample size. Other phase II sequential designs with multiple outcomes [3, 6, 7, 11, 14, 15] generally focus on providing stopping rules rather than incorporating predictive information.

Several phase I dose-finding designs [2, 16, 19] use co-primary outcomes. These could potentially be adapted to our purpose, although they generally do not use covariates. A notable exception is TNE, an extension of EffTox [16] that adds patient covariates to analyse co-primary efficacy and toxicity at different doses. The objective of their Bayesian design is to recommend a personal dose of an experimental agent, after adjusting for baseline data. The design was used in a dose-finding study of PR104 in relapsed or refractory acute myeloid or lymphoblastic leukaemia [12]. We found no other examples of its use, and no suggestion that it had been adapted for the non-dose-finding context. Our proposed design can be considered as a simplification of TNE for use in phase II.

4 Assessing Efficacy and Toxicity and Adjusting for Covariates

In this section, we describe the statistical design used in PePS2, with the general TNE model as the starting point. We call this design P2TNE, for *Phase II Thall, Nguyen and Estey*. TNE present marginal probability models for an experimental treatment:

$$\operatorname{logit} \pi_k(\tau, x, y, \boldsymbol{\theta}) = f_k(\tau, \boldsymbol{\alpha}_k) + \boldsymbol{\beta}_k x + \tau \boldsymbol{\gamma_k} y \ , \tag{1}$$

where $k = E, T$ denote efficacy and toxicity respectively. τ is the given dose appropriately normalised; x and y are vectors of covariates, with y interacting with dose; $\boldsymbol{\theta}$ is a pooled vector of all parameters to be estimated; $f_k(\tau, \boldsymbol{\alpha}_k)$ characterise the dose effects; and $\boldsymbol{\beta}_k$ and $\boldsymbol{\gamma}_k$ are vectors of covariate effects and dose-covariate interactions. TNE also introduce similar models for the events under historical treatments by which informative data on dose and covariate effects can be incorporated.

The authors consider joint models for associating events. They present an example using the Gumbel model, as used in [16]:

$$\begin{aligned}\pi_{a,b}(\pi_E, \pi_T, \psi) =& (\pi_E)^a (1-\pi_E)^{1-a} (\pi_T)^b (1-\pi_T)^{1-b} \\ &+ (-1)^{a+b} (\pi_E)(1-\pi_E)(\pi_T)(1-\pi_T) \frac{e^{\psi}-1}{e^{\psi}+1} \ ,\end{aligned} \tag{2}$$

where a and b equal 1 when efficacy and toxicity occur in a given patient respectively, else 0. For $\psi \in \mathbb{R}$, the fractional term takes values on $(-1, 1)$, reflecting the correlation between the events. We refer to ψ as the association parameter.

To derive P2TNE, we remove all terms related to τ in (1) to reflect that dose is fixed. Furthermore in PePS2, we consider only the historic outcomes of the same single experimental treatment under a closely-related cohort of patients with NSCLC.

Let $x_i = (x_{1i}, x_{2i}, x_{3i})$ denote the covariate data and a_i, b_i the occurrence of efficacy and toxicity in patient i. For trial data:

$$\boldsymbol{X} = \{(x_1, a_1, b_1), ..., (x_n, a_n, b_n)\} ,$$

the aggregate likelihood function is

$$\mathscr{L}(\boldsymbol{X}, \boldsymbol{\theta}) = \prod_{i=1}^{n} \pi_{a_i,b_i}(\pi_E(x_i, \boldsymbol{\theta}), \pi_T(x_i, \boldsymbol{\theta}), \psi) .$$

Let $\boldsymbol{\theta}$ have prior distribution $f(\boldsymbol{\theta})$. For patients with covariate data x, the posterior expectation of the probability of efficacy under treatment is

$$\mathbb{E}(\pi_E(x, \boldsymbol{\theta})|\boldsymbol{X}) = \frac{\int \pi_E(x, \boldsymbol{\theta}) f(\boldsymbol{\theta}) \mathscr{L}(\boldsymbol{X}, \boldsymbol{\theta}) d\boldsymbol{\theta}}{\int f(\boldsymbol{\theta}) \mathscr{L}(\boldsymbol{X}, \boldsymbol{\theta}) d\boldsymbol{\theta}} ,$$

and the posterior probability that the rate of efficacy exceeds some threshold π_E^* is

$$\Pr(\pi_E(x, \boldsymbol{\theta}) > \pi_E^*|\boldsymbol{X}) = \frac{\int \mathbb{I}(\pi_E(x, \boldsymbol{\theta}) > \pi_E^*) f(\boldsymbol{\theta}) \mathscr{L}(\boldsymbol{X}, \boldsymbol{\theta}) d\boldsymbol{\theta}}{\int f(\boldsymbol{\theta}) \mathscr{L}(\boldsymbol{X}, \boldsymbol{\theta}) d\boldsymbol{\theta}} .$$

The treatment is acceptable in patients with covariate vector x if

$$\begin{aligned} \Pr(\pi_E(x, \theta) > \pi_E^*|\boldsymbol{X}) > p_E \\ \Pr(\pi_T(x, \theta) < \pi_T^*|\boldsymbol{X}) > p_T , \end{aligned} \tag{3}$$

where π_E^*, p_E, π_T^* and p_T are chosen by the trialists. The clinical investigator chose the values $\pi_E^* = 0.1$ and $\pi_T^* = 0.3$ to reflect that efficacy less than 10% or toxicity exceeding 30% would render the treatment unattractive for further study in this patient group. We derived $p_E = 0.7$ and $p_T = 0.9$ by simulation using the method described below. Our chosen models for marginal efficacy and toxicity are:

$$\begin{aligned} \text{logit}\, \pi_E(x_i, \boldsymbol{\theta}) &= \alpha + \beta x_{1i} + \gamma x_{2i} + \zeta x_{3i} \\ \text{logit}\, \pi_T(x_i, \boldsymbol{\theta}) &= \lambda , \end{aligned} \tag{4}$$

with the events associated by (2). Our efficacy model assumes that the event log-odds for PT patients in the PD-L1 categories are a common linear shift of those in TN patients, an assumption we call *piecewise parallelism*, broadly supported by [9]. The rate of toxicity is assumed uniform across groups, supported by the data in [9, 10]. We analyse more complex models that relax each of these assumptions.

5 Simulation Study

Choice of priors is contentious in clinical trials. We simulated performance under diffuse, regularising, and informative priors. Our diffuse priors are normal with $\mu = 0$ and $\sigma = 10$. Regularising priors expect event rates close to 20% in all cohorts, put the majority of prior predictive mass in the left tail, but admit that event rates can be high. Informative priors expect event rates similar to those observed in [9], modestly penalised to reflect PS2 patient prognosis.

Table 3 shows operating characteristics using 60 patients. We tuned p_E and p_T by simulation in key benchmark scenarios, requiring that the design approve in all cohorts: (i) with at least 80% probability in scenario 1; and (ii) with no more than 5% probability in scenario 2. These probabilities reflect typical values for frequentist power and significance in phase II trials. Starting with $p_T = p_T = 0.7$, we saw that the designs accepted too often in scenario 2. With patients potentially near end-of-life, we chose to adjust operating performance by increasing certainty when evaluating toxicity; $p_T = 0.8$ was still too permissive but $p_T = 0.9$ achieved our goal under the regularising and diffuse priors, and $p_T = 0.95$ under informative priors. Scenarios 4–6 show that performance is good in settings inspired by the reported data [9, 10]. Compared to diffuse priors, the regularising priors improve approval probability without pre-empting covariate effects like the informative priors.

Table 3 also shows performance of beta-binomial conjugate models applied to cohorts individually with $Beta(1, 1)$ priors, accepting if (3) is satisfied with $p_E = 0.7$ and $p_T = 0.9$. By incorporating baseline covariates, P2TNE considerably improves performance without erroneously inflating acceptance in scenarios 2 and 5.

The diffuse priors generate prior predictive distributions with most of the probability mass polarised close to events rates of 0 and 1, inconsistent with our beliefs and the published data. Coverage of posterior credible intervals was lowest and empirical standard error of estimates highest under the diffuse priors (data not shown).

Our model choices (4) imply fairly strong assumptions. We analyse model embellishments to infer the cost of greater model complexity. We relax the piecewise parallel assumption by adding interactions terms to the efficacy model. Under diffuse priors, approval probabilities and coverages decrease in our scenarios. An extra 20–40 patients are required to match performance of the simpler model under diffuse priors. To correctly improve the rejection probability in cohort 4 under scenarios 4–6, this model requires several times the initial sample size, an unjustifiable increase.

We relaxed the assumption that toxicity is uniform over groups by mirroring in the toxicity model the efficacy covariate terms in (4), yielding a model with nine parameters including ψ. The extra model complexity reduces approval probabilities and coverage. Poor coverage is a particular problem in the toxicity model in scenarios

Table 3 Summary of simulated trials. Sc is scenario number; Co is cohort number. Patient cohorts are defined in Table 2. PrEff and PrTox are true probabilities of efficacy and toxicity. OddsR shows ratio of odds of efficacy in patients that experience toxicity to those that do not. OddsR=1 reflects no association; OddsR < 1 implies efficacy is less likely when toxicity occurs. N shows mean number of patients; Eff and Tox the mean number of events. Inf is the approval probability under informative priors; Reg and Diff are the same under regularising and diffuse priors. BetaBin shows approval probability using cohort-specific beta-binomial models. 10,000 iterations used

Sc	Co	PrEff	PrTox	OddsR	N	Eff	Tox	Inf	Reg	Diff	BetaBin
1	1	0.300	0.1	1.0	9.3	2.8	0.9	0.883	0.896	0.878	0.540
	2	0.300	0.1	1.0	13.1	3.9	1.3	0.906	0.920	0.905	0.658
	3	0.300	0.1	1.0	7.5	2.3	0.8	0.980	0.909	0.816	0.473
	4	0.300	0.1	1.0	12.5	3.7	1.2	0.875	0.912	0.896	0.635
	5	0.300	0.1	1.0	10.8	3.2	1.1	0.873	0.909	0.890	0.590
	6	0.300	0.1	1.0	6.8	2.0	0.7	0.959	0.893	0.819	0.459
2	1	0.100	0.3	1.0	9.3	0.9	2.8	0.012	0.025	0.019	0.035
	2	0.100	0.3	1.0	13.1	1.3	3.9	0.013	0.028	0.023	0.032
	3	0.100	0.3	1.0	7.5	0.8	2.3	0.038	0.029	0.021	0.034
	4	0.100	0.3	1.0	12.5	1.2	3.7	0.009	0.024	0.021	0.034
	5	0.100	0.3	1.0	10.8	1.1	3.2	0.009	0.024	0.022	0.032
	6	0.100	0.3	1.0	6.8	0.7	2.0	0.027	0.025	0.019	0.041
3	1	0.300	0.1	0.2	9.3	2.8	0.9	0.884	0.897	0.879	0.562
	2	0.300	0.1	0.2	13.1	3.9	1.3	0.906	0.920	0.904	0.667
	3	0.300	0.1	0.2	7.5	2.3	0.8	0.981	0.909	0.818	0.494
	4	0.300	0.1	0.2	12.5	3.7	1.2	0.877	0.913	0.897	0.652
	5	0.300	0.1	0.2	10.8	3.2	1.1	0.874	0.908	0.889	0.605
	6	0.300	0.1	0.2	6.8	2.0	0.7	0.960	0.893	0.820	0.478
4	1	0.167	0.1	1.0	9.3	1.5	0.9	0.408	0.451	0.398	0.293
	2	0.192	0.1	1.0	13.1	2.5	1.3	0.651	0.690	0.633	0.432
	3	0.500	0.1	1.0	7.5	3.8	0.8	0.993	0.981	0.974	0.622
	4	0.091	0.1	1.0	12.5	1.1	1.3	0.208	0.277	0.215	0.131
	5	0.156	0.1	1.0	10.8	1.7	1.1	0.405	0.493	0.419	0.298
	6	0.439	0.1	1.0	6.8	3.0	0.7	0.961	0.930	0.931	0.581
5	1	0.167	0.3	1.0	9.3	1.5	2.8	0.027	0.063	0.039	0.071
	2	0.192	0.3	1.0	13.1	2.5	3.9	0.046	0.099	0.066	0.084
	3	0.500	0.3	1.0	7.5	3.8	2.3	0.071	0.141	0.102	0.159
	4	0.091	0.3	1.0	12.5	1.1	3.7	0.014	0.037	0.021	0.028
	5	0.156	0.3	1.0	10.8	1.7	3.2	0.030	0.071	0.045	0.065
	6	0.439	0.3	1.0	6.8	3.0	2.0	0.070	0.135	0.099	0.163
6	1	0.167	0.1	0.2	9.3	1.5	0.9	0.408	0.451	0.396	0.308
	2	0.192	0.1	0.2	13.1	2.5	1.3	0.651	0.689	0.633	0.447
	3	0.500	0.1	0.2	7.5	3.8	0.8	0.993	0.981	0.974	0.627
	4	0.091	0.1	0.2	12.5	1.1	1.3	0.208	0.278	0.212	0.139
	5	0.156	0.1	0.2	10.8	1.7	1.1	0.402	0.493	0.415	0.313
	6	0.439	0.1	0.2	6.8	3.0	0.7	0.962	0.929	0.930	0.589

where the event rate is 10%. For instance, the four-parameter model performs very poorly in scenarios 1 and 3, particularly in the smallest cohorts. Performance is better in scenario 2 where the true rate is 30%. This is notable because the published data [9, 10] suggest low toxicity. In scenarios not shown in Table 3, this model successfully identifies differential toxicity associated with covariates but requires a sample size exceeding 100 to do so with high probability. Weighing the extra demand in resource against the likely benefit, we prefer the simpler model.

Lastly, scenarios 3 and 6 show that model performance is seemingly unaffected by strong association in efficacy and toxicity events. We investigated a model variant that assumes independence by setting $\psi = 0$ in (2). Approval probability and precision were practically unchanged. This is understandable because ψ is absent from (4) and therefore does not affect (3). ψ is useful, however, in conditional inference. For example, the predicted distribution of unknown efficacy conditioned on observed toxicity is shifted lower by ψ given negative association prevailing in the collected trial data, and vice-versa. Given its useful role with no performance penalty, we retain ψ.

6 Further Work and Availability of Materials

Statisticians know that dichotomising continuous variables reduces information. We have used in this research the PD-L1 categorisation previously introduced and validated in NSCLC [9]. In ongoing work, we use the underlying continuous score in place of the categorisation. In this setting, further care must be taken when specifying the model form and the parameter priors. For instance, we expect overwhelmingly that the gradient term describing the sensitivity of efficacy with respect to PD-L1 score will be positive, so that higher scores are more likely to yield efficacy events. However, it is debatable whether our priors or model form should reflect that we expect greater or lesser efficacy-PD-L1 sensitivity in treatment naive or pretreated patients. A hierarchical approach has some merit, where PD-L1 gradients are interpreted as draws from some common distribution. This would allow heterogeneity to manifest in subgroups whilst discouraging over-fitting via shrinkage-based regularisation. Missing data is a perennial challenge in clinical trials. A hierarchical approach has the further benefit of pragmatically treating patients with unknown pretreatment status as a third cohort. Intuitively, we could interpret this group as behaving like an unknown mixture of pretreated and treatment-naive patients.

One of the focuses of this research has been the consideration of different models that could eventually be fit to the trial data. We approached the problem as if one candidate model had to be identified in advance in the analysis plan. An alternative is to specify a suite of models and then combine their inferences. For instance, in Bayesian model averaging, the response distributions generated by the candidate models are weighted together by their marginal posterior probabilities. In contrast, methods have been introduced that *stack* posterior predictive distributions, using the leave-one-out (LOO) predictor for each model and each data-point, deriving model weights that minimise the LOO mean squared error [18]. A method like this could allow us to combine models with markedly different features like simple and complex specifications for the toxicity sub-model in a data-oriented manner.

Models used in this research were implemented in Stan [5] and all materials are available on GitHub at https://github.com/brockk/bebop.

References

1. Borghaei, H., Paz-Ares, L., Horn, L., Spigel, D.R., Steins, M., Ready, N.E., et al.: Nivolumab versus docetaxel in advanced non-squamous non-small-cell lung cancer. N. Engl. J. Med. **373**, 123–135 (2015)
2. Braun, T.M.: The bivariate continual reassessment method: extending the CRM to phase I trials of two competing outcomes. Control. Clin. Trials. **23**, 240–256 (2002)
3. Brutti, P., Gubbiotti, S., Sambucini, V.: An extension of the single threshold design for monitoring efficacy and safety in phase II clinical trials. Stat. Med. **30**, 1648–1664 (2011)
4. Bryant, J., Day, R.: Incorporating toxicity considerations into the design of two-stage phase II clinical trials. Biometrics. **51**, 1372–1383 (1995)
5. Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., et al.: Stan: a probabilistic programming language. J. Stat. Softw. **76**, 1–32 (2017)
6. Conaway, M., Petroni, G.: Designs for phase II trials allowing for a trade-off between response and toxicity. Biometrics. **52**, 1375–1386 (1996)
7. Cook, R., Farewell, V.: Guidelines for monitoring efficacy and toxicity responses in clinical trials. Biometrics. **50**, 1146–1152 (1994)
8. Eisenhauer, E.A., Therasse, P., Bogaerts, J., Schwartz, L.H., Sargent, D., et al.: New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur. J. Cancer. **45** 228–2DD47 (2009)
9. Garon, E.B., Rizvi, N.A., Hui, R., Leighl, N., Balmanoukian, A.S., et al.: Pembrolizumab for the treatment of nonsmall-cell lung cancer. N. Engl. J. Med. **372**, 2018–2028 (2015)
10. Herbst, R.S., Baas, P., Kim, D., Felip, E., Perez-Gracia, J.L., et al.: Pembrolizumab versus docetaxel for previously treated, PD-L1-positive, advanced non-small-cell lung cancer (KEYNOTE-010): a randomised controlled trial. Lancet. **387**, 1540–1550 (2016)
11. Jin, H.: Alternative designs of phase II trials considering response and toxicity. Contemp. Clin. Trials. **109**, 525–536 (2007)
12. Konopleva, M., Thall, P.F., Arana Yi, C., Borthakur, G., Coveler, A., et al.: Phase I/II study of the hypoxia-activated prodrug PR104 in refractory/relapsed acute myeloid leukemia and acute lymphoblastic leukemia. Haematologica. **100**, 927–934 (2015)
13. Schiller, J.H., Harrington, D., Belani, C.P., Langer, C., Sandler, A., et al.: Comparison of four chemotherapy regimens for advanced non-small-cell lung cancer. N. Engl. J. Med. **346**, 92–98 (2002)
14. Thall, P.F., Simon, R.M., Estey, E.H.: New statistical strategy for monitoring safety and efficacy in single-arm clinical trials. J. Clin. Oncol. **14**, 296–303 (1996)
15. Thall, P.F., Sung, H.G.: Some extensions and applications of a Bayesian strategy for monitoring multiple outcomes in clinical trials. Stat. Med. **17**, 1563–1580 (1998)
16. Thall, P.F., Cook, J.D.: Dose-finding based on efficacy-toxicity trade-offs. Biometrics. **60**, 684–693 (2004)
17. Thall, P.F., Nguyen, H.Q., Estey, E.: Patient-specific dose finding based on bivariate outcomes and covariates. Biometrics. **64**, 1126–1136 (2008)
18. Yao, Y., Vehtari, A., Simpson, D., Gelman, A.: Using stacking to average Bayesian predictive distributions. Bayesian Anal. **13**, 917–1007 (2017)
19. Zhang, W., Sargent, D.J., Mandrekar, S.: An adaptive dose-finding design incorporating both toxicity and efficacy. Stat. Med. **25**, 2365–2383 (2006)

A Conditional Autoregressive Model for Estimating Slow and Fast Diffusion from Magnetic Resonance Images

Ettore Lanzarone, Elisa Scalco, Alfonso Mastropietro, Simona Marzi and Giovanna Rizzo

Abstract The Intra-Voxel Incoherent Motion (IVIM) model is largely adopted to estimate slow and fast diffusion parameters of water molecules in biological tissues, which are used as biomarkers for different diseases. However, the standard approach to obtain the maps of these parameters is based on a voxel-by-voxel estimation and neglects the spatial correlations, thus resulting in noisy maps. To get better maps, we propose a Bayesian approach that exploits a Conditional Autoregressive (CAR) prior density. We consider a pure CAR model and a mixture CAR model, and we compare the outcomes with two benchmark approaches. Results show better maps under the CAR models.

Keywords Conditional autoregressive model · Diffusion parameters · Intra-voxel incoherent motion · Magnetic resonance imaging · Spatial correlation

1 Introduction

Diffusion-Weighted Magnetic Resonance Imaging (DW-MRI) is a non-invasive technique that is largely employed to quantitatively characterize the diffusion properties of water molecules in biological tissues. The acquisition of DW-MRI images is based on matched dephasing and rephasing gradient envelopes, whose characteristics (magnitude, duration and time interval) are combined in a single parameter b that influences the signal attenuation associated with the diffusion properties.

The Intra-Voxel Incoherent Motion (IVIM) model describes the incoherent motion of the water molecules as a biexponential decay over b, with a slow and a fast

E. Lanzarone (✉)
Institute for Applied Mathematics and Information Technologies (IMATI), CNR, Milan, Italy
e-mail: ettore.lanzarone@cnr.it

E. Scalco · A. Mastropietro · G. Rizzo
Institute of Molecular Bioimaging and Physiology (IBFM), CNR, Lecco, Italy

S. Marzi
Medical Physics Laboratory, Regina Elena National Cancer Institute, Rome, Italy

R. Argiento et al. (eds.), *Bayesian Statistics and New Generations*,
Springer Proceedings in Mathematics & Statistics 296,
https://doi.org/10.1007/978-3-030-30611-3_14

diffusion component that depend on three parameters: two decay rate coefficients (slow diffusion D and fast pseudo-diffusion D^{*}) and the pseudo-diffusion volume fraction f [10].

This model is largely applied in clinics to estimate the parameters D, D^{*} and f based on patient-specific DW-MRI images acquired at different b values. As already reported in the literature, these parameters can be used as biomarkers for different diseases [17]. Moreover, their estimation over the voxels allows obtaining maps of the spatial distribution of the diffusion properties in Regions of Interest (ROIs). In the literature and the practice, the standard approach to obtain these maps is a nonlinear least-square method, which fits the biexponential function to the acquired DW-MRI images using the Levenberg Marquardt or the Trust Region algorithm [16, 17]. However, it generates noisy maps, especially for D^{*} and f [17], due to the voxel-by-voxel estimation that reduces the Signal to Noise Ratio (SNR) and neglects the spatial structure of the biological tissues.

Our aim is to develop an alternative approach to include the spatial correlation in the estimation procedure, in order to reduce the noise of estimated maps. To this end, we exploit a Conditional Autoregressive (CAR) specification of the prior density, and we consider both a pure CAR model and a mixture CAR model.

2 Literature Review

Bayesian approaches have been largely applied to medical imaging, especially for tissue classification [1, 5]. Functional MRI (fMRI) is the most relevant application field for Bayesian approaches, whose aim is to infer brain regions that exhibit a neuronal activity in response to a given stimulus, by detecting blood flow changes [19]. For example, Jeong et al. [8] applied linear regression models with long memory errors and discrete wavelet transforms, together with a Bayesian estimation of model parameters. Zhang et al. [18] applied a wavelet-based Bayesian nonparametric regression model, accounting for the spatial correlation structure of the tissue by means of a Markov random field prior. Zhang et al. [20] applied a spatio-temporal linear regression model that specifically accounts for the heterogeneity in neuronal activity between subjects via a spatially informed multi-subject nonparametric variable selection prior.

Bayesian approaches have been also proposed to estimate the IVIM parameters in the context of DW-MRI. A simple Bayesian approach with uninformative or minimally informative priors was firstly proposed by Neil and Bretthorst [12]. Barbieri et al. [2] found that this approach is associated with lower variability and higher precision and accuracy with respect to several least-square approaches. This simple voxel-by-voxel Bayesian model was also applied by Dyvorne et al. [3] to estimate the effects of diffusion gradient polarity and breathing acquisition scheme on image quality, SNR, IVIM parameters, and parameter reproducibility in the liver district. Orton et al. [13] proposed a shrinkage prior model with no user-defined parameters, and this approach was applied by Spinner et al. [14] in cardiovascular images. Freiman et al. [4] and Kurugol et al. [9] considered a spatial homogeneity prior in the form of a continuous Markov random field to obtain spatially regular parametric maps, and they proposed a *fusion bootstrap moves* algorithm to get the posterior density. In a recent work, While [17] compared several approaches to estimate the IVIM parametric maps, including a Bayesian model with a Gaussian shrinkage prior [13] and two Bayesian models with a spatial homogeneity prior [4]. He highlighted that the Bayesian approaches consistently outperformed the classical non-linear least square fitting, and that the use of a spatial homogeneity prior reduces the errors with respect to a Gaussian one. However, these approaches could mask some tissue features and, thus, they must be implemented with caution. For example, the prior distribution and the central tendency measure have an impact on the estimation results, as demonstrated by [7], who found that the best overall quality of the IVIM parameters is obtained with a lognormal prior. To the best of our knowledge, none of available works consider spatial autoregressive models.

3 Problem Description and Approach

Our dataset consists of several DW-MRI signal intensity images acquired at different b values ($b \in B$ starting from $b = 0\,\mathrm{s/mm^2}$). Each image is divided in voxels and we refer to a single layer of them, i.e., we deal with two-dimensional images for which the discrete coordinates of a voxel are (i, j), with $i \in I$ and $j \in J$.

The signal intensity in voxel (i, j) at b is denoted by $SI(i, j, b)$, and the intensity decay over b is described in each voxel by the IVIM model [17]:

$$SI\,(i,j,b) = SI\,(i,j,0)\left\{f\,(i,j)\,e^{-b\,D_{sum}(i,j)} + \left[1 - f\,(i,j)\right]e^{-b\,D(i,j)}\right\} \qquad \forall i \in I, j \in J, b \in B \setminus \{0\} \tag{1}$$

where $D_{sum}\,(i,j) = D\,(i,j) + D^*\,(i,j)$. Our goal is to estimate the coefficients $D(i,j)$, $D_{sum}(i,j)$ and $f\,(i,j)$ in each voxel. Then, we derive $D^*(i,j) = D_{sum}(i,j) - D(i,j)$.

We propose a pure CAR approach and a mixed CAR approach, in which the prior includes a CAR and a Gaussian component.

3.1 Likelihood Function

The parameters to estimate are included in set $\boldsymbol{\Theta}$, with $\boldsymbol{\Theta} = \boldsymbol{D} \cup \boldsymbol{D}_{sum} \cup \boldsymbol{f}$, $\boldsymbol{D} = \{D\,(i,j)\,, i \in I, j \in J\}$, $\boldsymbol{D}_{sum} = \{D_{sum}\,(i,j)\,, i \in I, j \in J\}$ and $\boldsymbol{f} = \{f\,(i,j)\,, i \in I, j \in J\}$.

Each parameter in $\boldsymbol{\Theta}$ is assumed to be a random variable. Thus, each decay Eq. (1) is a random process, and we may express the density of each $SI\ (i, j, b)$ as conditioned to $\boldsymbol{\Theta}$:

$$SI\ (i, j, b) \sim \mathscr{L}\left(SI\ (i, j, b)\ |\boldsymbol{\Theta}, SI\ (i, j, 0)\right) \quad \forall i \in I, j \in J, b \in B \setminus \{0\} \tag{2}$$

where $\mathscr{L}$ denotes the conditional probability law. Moreover, we assume that each observation SI^{obs}_{ijb} (with $b \in B \setminus \{0\}$) is subject to a measurement error and that the errors are independent, based on the fact that each image is separately acquired.

Thus, we model the observations as stochastic variables centered on the value computed from the model, which are conditionally independent given the model parameters. Indeed, we consider a Normal distribution $\mathscr{N}$ with mean value $SI\ (i, j, b)$ and standard deviation σ_{obs}:

$$SI^{obs}_{ijb} \sim \mathscr{N}\left(SI\ (i, j, b)\ , \sigma^2_{obs}\right) \quad \forall i \in I, j \in J, b \in B \setminus \{0\} \tag{3}$$

Actually, a Rician distribution should be used for low SNR values, while the Rician distribution is well approximated by the Gaussian one [6] for high SNR values. Anyway, the Gaussian and the Rician likelihood functions give similar results for a large range of SNR values [7].

The combination of (2) and (3) gives the conditional law of each observation $SI^{obs}_{i,j,b}$, while their product over $i \in I, j \in J$ and $b \in B \setminus \{0\}$ gives the likelihood function of the overall set of observations given $\boldsymbol{\Theta}$.

3.2 Prior Density

We assume a priori independence between $\boldsymbol{D}$, $\boldsymbol{D}_{sum}$ and $\boldsymbol{f}$. Then, separately for each parameter $\lambda\ (i, j)$, where λ generically denotes $D\ (i, j)$, $D_{sum}\ (i, j)$ or $f\ (i, j)$, we consider two alternative priors. The former is a CAR specification, while the latter is a mixture model with two components: the same CAR specification and a simple Gaussian density.

Moreover, the standard deviation σ_{obs} follows an independent Gamma density:

$$\sigma_{obs} \sim Gamma\left(\frac{\sum_{i \in I, j \in J} SI^{obs}_{ij0}}{2\,|I \times J|}, 1\right)$$

The mean value is the half of the average value of SI^{obs}_{ijb} at the initial $b = 0$, to scale the mean value based on the specific image set, and the variance is equal to the mean value.

3.2.1 CAR Specification

We refer to [11] and we assume the following intrinsic CAR model for each parameter $\lambda\,(i,j)$ given the rest of the values:

$$\lambda\,(i,j)\,|\boldsymbol{\lambda}_{i,j}^{c} \sim \mathscr{N}\left(\frac{\sum_{\alpha\in I,\beta\in J} w(\alpha,\beta,i,j)\lambda\,(\alpha,\beta)}{\sum_{\alpha\in I,\beta\in J} w(\alpha,\beta,i,j)}, \frac{\sigma_\lambda^2}{\sum_{\alpha\in I,\beta\in J} w(\alpha,\beta,i,j)}\right) \tag{4}$$

where $\boldsymbol{\lambda}_{i,j}^{c} = \boldsymbol{\lambda} \setminus \{\lambda\,(i,j)\}$ and $w(\alpha,\beta,i,j)$ denotes the spatial neighborhood matrix. In particular, we assume $w(\alpha,\beta,i,j) = 1$ for the voxels (α,β) bordering (i,j), and 0 elsewhere:

$$w(\alpha,\beta,i,j) = \begin{cases} 1 & \alpha = \{i-1, i+1\} \text{ and } \beta = \{j-1, j, j+1\} \\ 1 & \alpha = i \text{ and } \beta = \{j-1, j+1\} \\ 0 & \text{otherwise} \end{cases}$$

With respect to [11], we consider the intrinsic model by setting the spatial autocorrelation coefficient $\rho = 1$; possible instabilities are fixed by the mixture model.

Finally, the priors for the standard deviations are $\sigma_f \sim \mathit{Gamma}\,(0.25, 1)$, $\sigma_{D_{sum}} \sim \mathit{Gamma}\,(0.01, 1)$ and $\sigma_D \sim \mathit{Gamma}\,(0.001, 1)$, with mean value in agreement with the literature and variance equal to the mean value.

3.2.2 Mixture Model

We assume the following mixture distribution for each parameter $\lambda\,(i,j)$:

$$f\,(\lambda\,(i,j)\,|\boldsymbol{\lambda}_{i,j}^{c}) = w_C f_C\left(\lambda\,(i,j)\,|\boldsymbol{\lambda}_{i,j}^{c}\right) + w_G f_G\,(\lambda\,(i,j)) \tag{5}$$

where f denotes the mixture density, f_C the density of the CAR component, as in (4), and f_G the density of the Gaussian component, which assumes the following form for each parameter $\lambda\,(i,j)$:

$$\lambda\,(i,j) \sim \mathscr{N}\left(\mu_\lambda, \sigma_{0\lambda}^2\right) \tag{6}$$

with $\mu_f = \sigma_{0f} = 0.1$, $\mu_{D_{sum}} = \sigma_{0D_{sum}} = 0.01$, $\mu_D = 0.001$ and $\sigma_{0D} = 0.0001$.

The weights are set as follows: $w_C = 0.75$ and $w_G = 0.25$.

3.3 *Posterior Estimates*

The Bayesian models are implemented in R with package RSTAN [15]; the code is freely available upon request. Samples from the posterior densities are obtained with 1000 iterations after a warm up of 1000 iterations, which guarantee convergence.

4 Application to Test Images and Results

We consider two test images, the former related to the head-and-neck district (HN) and the latter to the shoulder (SH). In both cases, $B = \{0, 25, 50, 75, 100, 150, 300, 500, 800\}$ s/mm^2. We compare four approaches (our proposals and two benchmarks):

1. *Mixture*: the proposed approach with the mixture prior;
2. *CAR*: the proposed approach with the CAR prior;
3. *Gauss*: a simple Bayesian approach with the Gaussian prior in (6);
4. *Segm*: the standard segmented approach with a non-linear least square fitting, using the Trust Region optimization algorithm [17].

We first evaluate the quality of the maps. Then, we perform a quantitative comparison in some ROIs manually identified (tumor and muscle in both cases; parotid gland in HN). We consider the Coefficient of Variation (CV) for each parameter within each area, i.e., the ratio between the standard deviation and the mean value of the parameter over the voxels in the area [9]. As for the Bayesian approaches, we consider the maps of the posterior mean in each voxel, and mean and standard deviations for the CVs are computed with the posterior means in the voxels.

The maps of f, D and $D^* = D_{sum} - D$ in Figs. 1 and 2 show that *Mixture* and *CAR* outperform the other approaches. *Segm* maps are the noisiest, while *Gauss* maps are qualitatively too smooth, especially for D where contrasts are lost and compromise the estimation of the other coefficients. CVs in Table 1 show the highest values in *Segm* and the lowest in *Gauss*, as expected (noisy images and lost contrasts, respectively). CVs of *Mixture* and *CAR* approaches are fair and similar. However, as shown in Figs. 1 and 2, *Mixture* provides a better estimation of the D^* map.

5 Discussions and Conclusion

We apply for the first time in the literature a Bayesian CAR approach to estimate the IVIM model coefficients.

Results have confirmed that the maps estimated with the *Segm* approach are the noisiest, as reported in the literature. Also the maps estimated with the *Gauss* approach are not acceptable, even though the CV is low, because of the considerable underestimation of D and D^* and the overestimation of f (data not shown). *Mixture* and *CAR* approaches generate acceptable maps with similar CVs and realistic parameter values in each ROI. However, the pure *CAR* approach presents saturation problems, especially for D^*. These problems are mitigated by the *Mixture* approach, which regularizes local criticalities.

Future work will be devoted to further improve the estimation approach, e.g., exploiting the segmentation used in [16, 17] also in our Bayesian CAR framework.

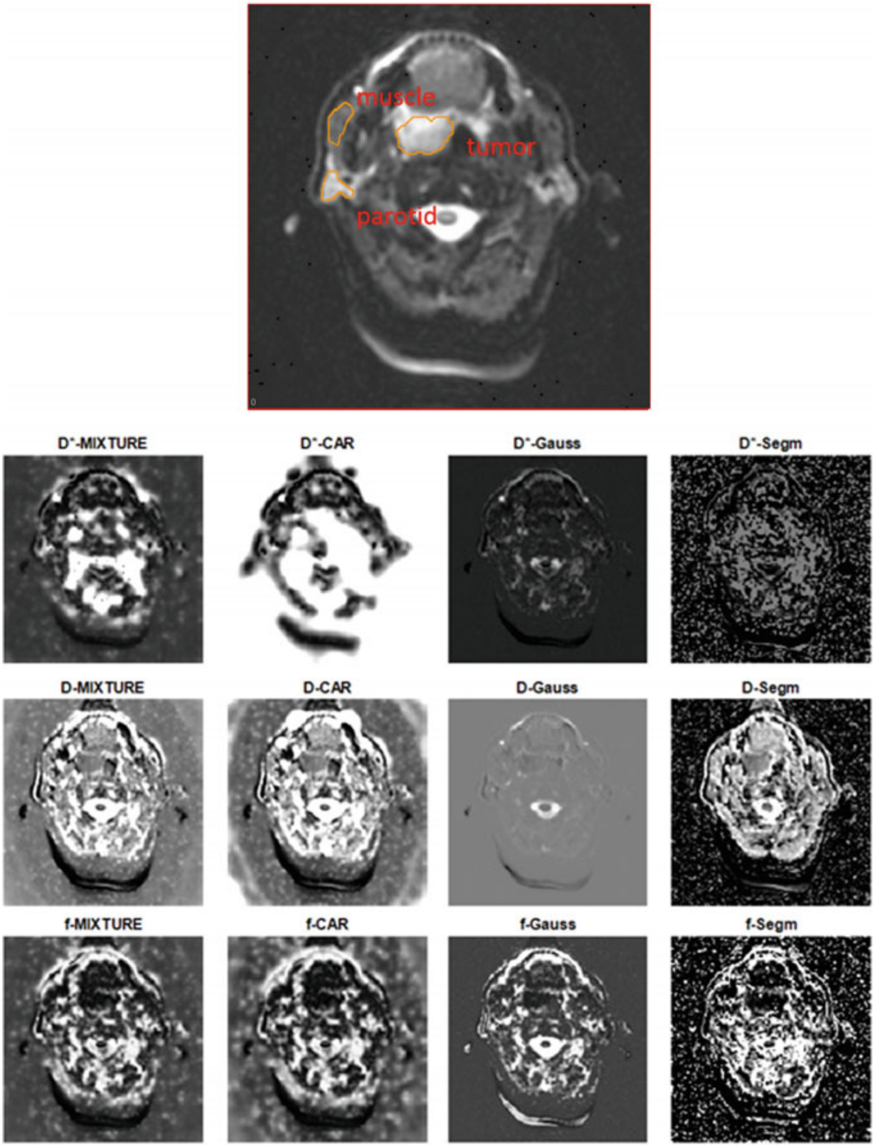


Fig. 1 HN case. DW-MRI image acquired at $b = 0\ \mathrm{s/mm^2}$ with the three ROIs (muscle, tumor and parotid) delineated in yellow (top image). Estimated maps of D^*, D and f under the four approaches: mixture (first column), CAR (second column), Gauss (third column), Segm (fourth column)

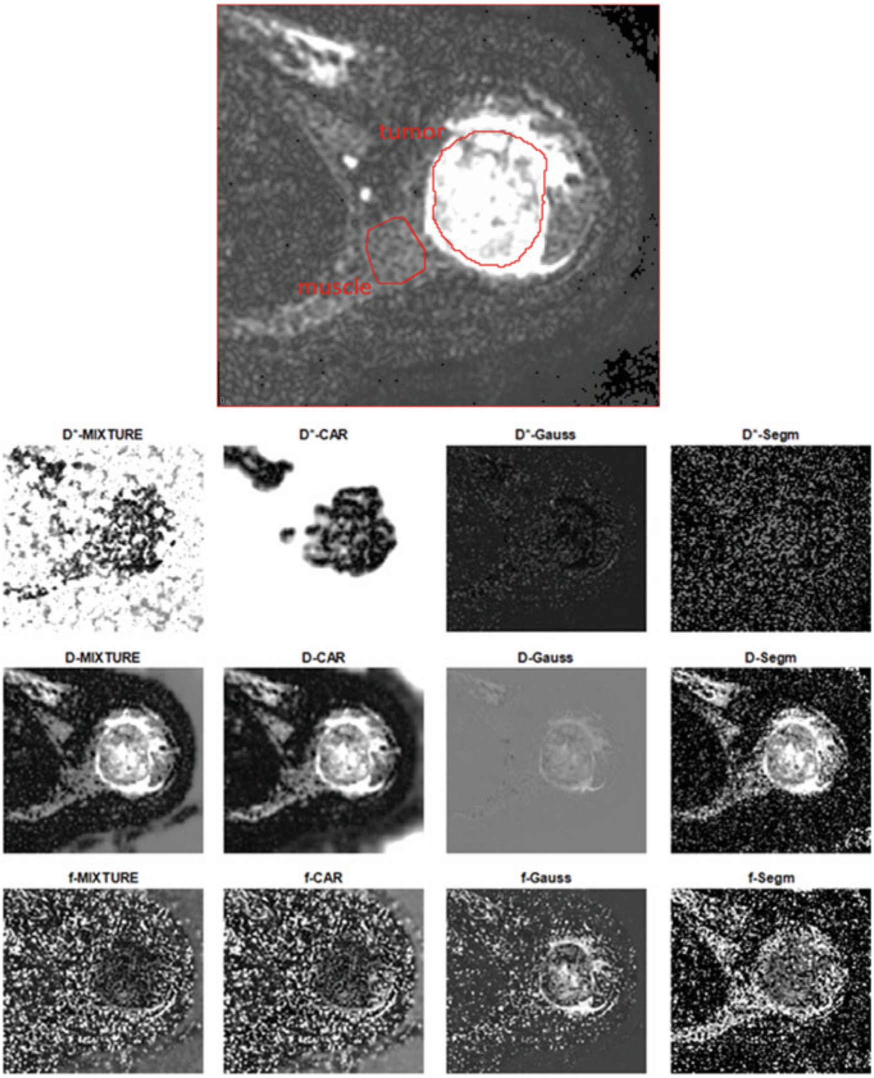


Fig. 2 SH case. DW-MRI image acquired at $b = 0$ s/mm^2 with the two ROIs (muscle and tumor) delineated in red (top image). Estimated maps of D^*, D and f under the four approaches: mixture (first column), CAR (second column), Gauss (third column), Segm (fourth column)

Table 1 CVs of parameters D^*, D and f in the ROIs, for both the HN and SH case

			Mixture	CAR	Gauss	Segm
D^*	HN	Tumor	0.532	0.394	0.379	0.750
		Parotid	0.324	0.500	0.426	0.425
		Muscle	0.530	0.505	0.355	1.182
	SH	Tumor	0.694	0.648	0.504	0.504
		Muscle	0.674	0.636	0.391	1.053
D	HN	Tumor	0.204	0.215	0.053	0.247
		Parotid	0.490	0.676	0.095	0.355
		Muscle	0.179	0.190	0.026	0.162
	SH	Tumor	0.258	0.255	0.072	0.293
		Muscle	0.236	0.228	0.029	0.452
f	HN	Tumor	1.008	1.067	0.671	1.570
		Parotid	0.540	0.522	0.437	0.588
		Muscle	0.830	0.902	0.584	1.223
	SH	Tumor	0.532	0.529	0.434	0.630
		Muscle	0.590	0.703	0.546	0.627

References

1. Alston, C.L., Mengersen, K.L., Thompson, J.M., Littlefield, P.J., Perry, D., Ball, A.J.: Extending the Bayesian mixture model to incorporate spatial information in analysing sheep CAT scan images. Aust. J. Agr. Res. **56**, 373–388 (2005)
2. Barbieri, S., Donati, O.F., Froehlich, J.M., Thoeny, H.C.: Impact of the calculation algorithm on biexponential fitting of diffusion-weighted MRI in upper abdominal organs. Magn. Reson. Med. **75**, 2175–2184 (2016)
3. Dyvorne, H.A., Galea, N., Nevers, T., Fiel, M.I., Carpenter, D., Wong, E., Orton, M., de Oliveira, A., Feiweier, T., Vachon, M.L., Babb, J.S., Taouli, B.: Diffusion-weighted imaging of the liver with multiple b values: effect of diffusion gradient polarity and breathing acquisition on image quality and intravoxel incoherent motion parameters - a pilot study. Radiology **266**, 920–929 (2013)
4. Freiman, M., Perez-Rossello, J.M., Callahan, M.J., Voss, S.D., Ecklund, K., Mulkern, R.V., Warfield, S.K.: Reliable estimation of incoherent motion parametric maps from diffusion-weighted MRI using fusion bootstrap moves. Med. Image Anal. **17**, 325–336 (2013)
5. Feng, D., Tierney, L., Magnotta, V.: MRI tissue classification using high-resolution Bayesian hidden Markov normal mixture models. J. Am. Stat. Assoc. **107**, 102–119 (2012)
6. Gudbjartsson, H., Patz, S.: The Rician distribution of noisy MRI data. Magn. Reson. Med. **34**, 910–914 (1995)
7. Gustafsson, O., Montelius, M., Starck, G., Ljungberg, M.: Impact of prior distributions and central tendency measures on Bayesian intravoxel incoherent motion model fitting. Magn. Reson. Med. **79**, 674–1683 (2018)
8. Jeong, J., Vannucci, M., Ko, K.: A wavelet-based Bayesian approach to regression models with long memory errors and its application to fMRI data. Biometrics **69**, 184–196 (2013)
9. Kurugol, S., Freiman, M., Afacan, O., Perez-Rossello, J.M., Callahan, M.J., Warfield, S.K.: Spatially-constrained probability distribution model of incoherent motion (SPIM) for abdominal diffusion-weighted MRI. Med. Image Anal. **32**, 173–183 (2016)

10. Le Bihan, D.: Intravoxel incoherent motion imaging using steady-state free precession. Magn. Reson. Med. **7**, 346–351 (1988)
11. Leroux, B.G., Lei, X., Breslow, N.: Estimation of disease rates in small areas: a new mixed model for spatial dependence. In: Halloran, M.E., Berry, D. (eds); Statistical Models in Epidemiology, the Environment, and Clinical Trials; the IMA Volumes in Mathematics and its Applications **116**, 179–191 (2000)
12. Neil, J.J., Bretthorst, G.L.: On the use of Bayesian probability theory for analysis of exponential decay date: an example taken from intravoxel incoherent motion experiments. Magn. Reson. Med. **29**, 642–647 (1993)
13. Orton, M.R., Collins, D.J., Koh, D.M., Leach, M.O.: Improved intravoxel incoherent motion analysis of diffusion weighted imaging by data driven Bayesian modeling. Magn. Reson. Med. **71**, 411–420 (2014)
14. Spinner, G.R., von Deuster, C., Tezcan, K.C., Stoeck, C.T., Kozerke, S.: Bayesian intravoxel incoherent motion parameter mapping in the human heart. J. Cardiovasc. Magn. Reson. **19**, 85 (2017)
15. Stan Development Team: Stan modeling language users guide and reference manual (version 2.9.0), http://mc-stan.org (2015)
16. Suo, S., Lin, N., Wang, H., Zhang, L., Wang, R., Zhang, S., Hua, J., Xu, J.: Intravoxel incoherent motion diffusion-weighted MR imaging of breast cancer at 3.0 Tesla: comparison of different curve-fitting methods. J. Magn. Reson. Imaging **42**, 362–370 (2015)
17. While, P.T.: A comparative simulation study of Bayesian fitting approaches to intravoxel incoherent motion modeling in diffusion-weighted MRI. Magn. Reson. Med. **78**, 2373–2387 (2017)
18. Zhang, L., Guindani, M., Versace, F., Vannucci, M.: A spatio-temporal nonparametric Bayesian variable selection model of fMRI data for clustering correlated time courses. Neuroimage **95**, 162–175 (2014)
19. Zhang, L., Guindani, M., Vannucci, M.: Bayesian models for functional magnetic resonance imaging data analysis. WIREs Comp. Stat. **7**, 21–41 (2015)
20. Zhang, L., Guindani, M., Versace, F., Vannucci, M.: A spatiotemporal nonparametric Bayesian model of multi-subject fMRI data. Ann. Appl. Stat. **10**, 638–666 (2016)

Simulation Study of HIV Temporal Patterns Using Bayesian Methodology

Diana Rocha, Manuel Scotto, Carla Pinto, João Tavares and Sónia Gouveia

Abstract Viral load values and CD4$^+$T cells count are markers currently evaluated in the clinical follow-up of HIV/AIDS patients. In this context, it is relevant to develop methods that provide a more complete temporal description of these markers, e.g. in between clinical appointments. To this end, we combine a mathematical model and a Bayesian methodology to estimate trajectories from a set of observed values. Also, we construct a variation band containing the most central trajectories for one patient, by exploring the range of values in the a posteriori distributions. The methods are illustrated with simulated data.

Keywords Bayesian statistics · Human immunodeficiency virus (HIV) · Mathematical models · Nonlinear programming · Parameter estimation

D. Rocha
Center for R&D in Mathematics and Applications (CIDMA), University of Aveiro (UA), Aveiro, Portugal

M. Scotto
CEMAT and Department of Mathematics, IST, University of Lisbon, Lisbon, Portugal

C. Pinto
School of Engineering, Polytechnic of Porto, Porto, Portugal

Centre of Mathematics of the University of Porto - CMUP, Porto, Portugal

J. Tavares
Centre of Mathematics of the University of Porto - CMUP, Porto, Portugal

S. Gouveia (✉)
Center for R&D in Mathematics and Applications (CIDMA), University of Aveiro (UA), Aveiro, Portugal
e-mail: sonia.gouveia@ua.pt

Institute of Electronics and Informatics Engineering of Aveiro, IEETA, UA, Aveiro, Portugal

R. Argiento et al. (eds.), *Bayesian Statistics and New Generations*,
Springer Proceedings in Mathematics & Statistics 296,
https://doi.org/10.1007/978-3-030-30611-3_15

1 Introduction and Methods

Viral dynamic models can be formulated through a system of nonlinear ordinary differential equations, which enables to describe the temporal evolution of the clinical parameters of a HIV patient. To this extent, several HIV dynamic models have been proposed in the literature [9]. An instance of that are the two versions of a HIV model based on the interactions between uninfected, latently infected and actively infected CD4^{+}T cells together with free virus, introduced by Perelson et al. [10]. Moreover, Bonhoeffer et al. [1] discuss a model that includes nonlinear interactions between virus and host cell populations, including the effect of a HIV therapy approach. Most of the statistical models used in HIV context are non-linear mixed effects and model longitudinal outcomes by accounting for within and between subject sources of variations [14]. In this work, we are solely interested in the within individual variation. We also assume that the within variation of a subject (i.e. its temporal trajectory) is well described by a system of ordinary differential equations (ODE), where its parameters reflect the intra-subject mechanisms related to the interaction between virus and immune system response. Thus, the statistical model is here defined from the numerical solution of the ODE system and an error term that represents measurement error and model misspecification [5, 6].

For the estimation of the parameters of a HIV model, the most commonly used approaches in practice are based either on nonlinear least squares (e.g. [7]) or on Bayesian estimation (e.g. [6]). The advantages of Bayesian approaches are well-known and include the simplicity of computational implementation and the generation of a posteriori distributions for the unknown parameters. Additionally, these methods offer the possibility of including prior knowledge in the model (e.g., the physiological variation interval for the parameters of the model).

This work aims at estimating the temporal trajectories of the clinical markers of a HIV patient, from a set of (sparse) observations over time. The methodology adopted in this work combines a Bayesian approach initial values obtained from nonlinear least squares to the experimental data. Furthermore, we introduce temporal trajectory bands based on the multivariate a posteriori distribution to characterize the variability of temporal trajectories for a given subject.

The rest of the paper is organized as follows: in Sect. 1.1, the dynamics of HIV/AIDS infection are described through a mathematical model. In Sect. 1.2, the Markov-Chain Monte Carlo (MCMC) methodology is introduced and explained in detail. Furthermore, the optimization procedure developed to obtain optimal initial estimates for the parameters of the model is also presented. These methods allow to build several trajectories for each patient, where each trajectory is obtained for a given solution in the a posteriori distributions. In Sect. 2, the methods are illustrated with simulated data that mimic experimental temporal trajectories and incorporates laboratory measurement errors of the clinical markers. Finally, Sect. 3 is devoted to conclusions.

1.1 Mathematical Model

We consider a modified version of the mathematical model in Stafford et al. [12] for the dynamics of HIV/AIDS infection. The model includes an additional parameter ε denoting the effectiveness of the antiretroviral therapy [8], and is represented as

$$\begin{aligned}\frac{dT(t)}{dt} &= \lambda - d_1T(t) - (1-\varepsilon)k_1T(t)V(t)\\ \frac{dT^*(t)}{dt} &= (1-\varepsilon)k_1T(t)V(t) - \delta T^*(t)\\ \frac{dV(t)}{dt} &= \pi_1T^*(t) - cV(t)\end{aligned} \tag{1}$$

with initial conditions $(T(0), T^*(0), V(0)) = (T_0, T_0^*, V_0)$. The state variables are the viral load $V(t)$ and the number of uninfected and infected CD4$^+$ T cells, respectively $T(t)$ and $T^*(t)$. Thus, the total number of CD4$^+$ T cells is $CD4(t) = T(t) + T^*(t)$. The model also incorporates parameters with clinical interpretation, namely $\boldsymbol{\theta} = (d_1, \varepsilon, k_1, \delta, \pi_1, c)$. The parameter d_1 is the difference between rate loss from cell death and rate gain due to cell division of CD4$^+$ T cells, and the term $\lambda = d_1T_0$ expresses the proliferation rate of uninfected target cells. The parameter $0 \leq \varepsilon \leq 1$ is the effectiveness of therapy, k_1 is the infectivity rate, δ is the death rate of infected cells, π_1 is the average number of virions produced by a single infected cell and c is the clearance rate of free virions [12].

1.2 Bayesian Approach for Parameter Estimation (MCMC)

The vector $\boldsymbol{\theta}$ is estimated from a set of $CD4(t)$ values observed for one patient at its clinical follow-up appointments. The Bayesian estimation approach used for the estimation of $\boldsymbol{\theta}$, presented in Sect. 1.2.1, incorporates non-informative prior distributions which require the setting of initial values in order to start the iterative parameters updates. In particular, there is a need to set the initial values for the parameters which are obtained by the optimization procedure described in Sect. 1.2.2.

1.2.1 MCMC Methodology

The repeated measurements $y(t_i) := CD4(t_i)$ of a subject at a treatment time t_i can be written as

$$y(t_i) = f(\boldsymbol{\theta}, t_i) + \boldsymbol{e}(t_i), i = 1, \ldots, n, \tag{2}$$

where $f(\boldsymbol{\theta}, t_i) := \hat{y}(t_i)$ represents the CD4 value of the subject at treatment time t_i, provided by the mathematical model described in (1), by considering a set of fixed

Algorithm 1 MCMC Algorithm

```
1: Compute optimal $\widehat{\boldsymbol{\eta}}$ and set $\boldsymbol{\eta} = \widehat{\boldsymbol{\eta}}$.
2: **Initialization** ($k = 1$)
   2.1 Set hyper-parameters $a, b, \boldsymbol{\eta}, \boldsymbol{\Lambda}, \boldsymbol{\Omega}$ and $\upsilon$.
   2.2 Generate $\boldsymbol{\theta}^{(k)}$ for $k = 1$:
     • generate $\sigma^{-2(k)}$ following $\sigma^{-2} \sim G_a(a, b)$ where $G_a$ is the Gamma distribution.
     • generate $\boldsymbol{\mu}^{(k)}$ following $\boldsymbol{\mu} \sim N(\boldsymbol{\eta}, \boldsymbol{\Lambda})$ where $N$ is the Normal distribution.
     • generate $\boldsymbol{\Sigma}^{-1(k)}$ following $\boldsymbol{\Sigma}^{-1} \sim W_i(\boldsymbol{\Omega}, \upsilon)$ where $W_i$ is the Wishart distribution.
     • generate $\boldsymbol{\theta}^{(k)} = \boldsymbol{\eta} + \boldsymbol{b}$, $[\boldsymbol{b}|\boldsymbol{\Sigma}] \sim N(0, \boldsymbol{\Sigma}^{(k)})$.
   2.3 Accept $\boldsymbol{\theta}^{(k)}$ if $(0.01, 0, 10^{-11}, 0.24, 50, 2.39) \leq \boldsymbol{\theta}^{(k)} \leq (0.02, 1, 10^{-5}, 0.7, 10^4, 23)$ otherwise repeat from 2.2.
3: **Actualization** ($k > 1$)
   3.1 Generate candidate $\boldsymbol{\varphi}$ for $k > 1$:
     • generate $\sigma^{-2(k)}$ following $[\sigma^{-2}|\boldsymbol{\mu}^{(k-1)}, \boldsymbol{\Sigma}^{-1(k-1)}, \boldsymbol{\theta}^{(k-1)}, \boldsymbol{y}] \sim G_a(a + \frac{1}{2}, \boldsymbol{A}^{-1(k)})$,
           where $\boldsymbol{A}^{(k)} = b^{-1} + \frac{1}{2}\sum_{i=1}^{n}[y(t_i) - f(\boldsymbol{\theta}^{(k-1)}, t_i)]^2$.
     • generate $\boldsymbol{\mu}^{(k)}$ following $[\boldsymbol{\mu}|\sigma^{-2(k-1)}, \boldsymbol{\Sigma}^{-1(k-1)}, \boldsymbol{\theta}^{(k-1)}, \boldsymbol{y}] \sim N(\boldsymbol{B}^{-1(k)}\boldsymbol{C}^{(k)}, \boldsymbol{B}^{-1(k)})$,
           where $\boldsymbol{B}^{(k)} = \boldsymbol{\Sigma}^{-1(k-1)} + \boldsymbol{\Lambda}^{-1}$ and $\boldsymbol{C}^{(k)} = \boldsymbol{\Sigma}^{-1(k-1)}\boldsymbol{\theta}^{(k-1)} + \boldsymbol{\Lambda}^{-1}\boldsymbol{\eta}$.
     • generate $\boldsymbol{\Sigma}^{-1(k)}$ following $[\boldsymbol{\Sigma}^{-1}|\sigma^{-2(k-1)}, \boldsymbol{\mu}^{(k-1)}, \boldsymbol{\theta}^{(k-1)}, \boldsymbol{y}] \sim W_i(\boldsymbol{D}^{-1(k)}, 1 + \upsilon)$,
           where $\boldsymbol{D}^{(k)} = \boldsymbol{\Omega}^{-1} + (\boldsymbol{\theta}^{(k-1)} - \boldsymbol{\mu}^{(k-1)})(\boldsymbol{\theta}^{(k-1)} - \boldsymbol{\mu}^{(k-1)})^T$.
     • generate candidate $\boldsymbol{\varphi} = \boldsymbol{\mu}^{(k)} + \boldsymbol{b}$, $[\boldsymbol{b}|\boldsymbol{\Sigma}^{(k)}] \sim N(0, \boldsymbol{\Sigma}^{(k)})$.
   3.2 Keep $\boldsymbol{\varphi}$ if $(0.01, 0, 10^{-11}, 0.24, 50, 2.39) \leq \boldsymbol{\varphi} \leq (0.02, 1, 10^{-5}, 0.7, 10^4, 23)$ otherwise repeat from 3.1.
4: **Accept or reject candidate $\boldsymbol{\varphi}$**
   4.1 Define $\pi(\boldsymbol{x}|\sigma^{-2(k)}, \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{-1(k)}, \boldsymbol{y}) = exp\{-\frac{\sigma^{-2(k)}}{2}\sum_{i=1}^{n}[y(t_i) - f(\boldsymbol{x}, t_i)]^2 - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}^{(k)})^T\boldsymbol{\Sigma}^{-1(k)}(\boldsymbol{x} - \boldsymbol{\mu}^{(k)})\}$.
   4.2 Evaluate the acceptance probability $\alpha(\boldsymbol{\varphi}|\boldsymbol{\theta}^{(k-1)}) = min\left(1, \frac{\pi(\boldsymbol{\varphi}|\sigma^{-2(k)}, \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{-1(k)}, \boldsymbol{y})}{\pi(\boldsymbol{\theta}^{(k-1)}|\sigma^{-2(k)}, \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{-1(k)}, \boldsymbol{y})}\right)$.
   4.3 Generate $u$ following $u \sim U(0, 1)$ where $U$ is the Uniform distribution.
   4.4 Assign $\boldsymbol{\theta}^{(k)} = \boldsymbol{\varphi}$ and move to $k + 1$ if $\alpha(\boldsymbol{\varphi}|\boldsymbol{\theta}^{(k-1)}) \geq u$ otherwise $\boldsymbol{\theta}^{(k)} = \boldsymbol{\theta}^{(k-1)}$ and repeat from 3.
```

values for the components in $\boldsymbol{\theta}$. Finally, $\boldsymbol{e}(t)$ is a zero-mean error with variance σ^2 [5]. In accordance with other literature studies, we considered log_{10}-transformations to ensure the positiveness of the dynamic parameters and to stabilize the $CD4(t)$ variance [5].

The MCMC methodology proceeds as presented in Algorithm 1. There is an initialization step where the hyper-parameters of the a priori distributions are set. We consider $\sigma^{-2} \sim G_a(a, b)$, $\boldsymbol{\mu} \sim N(\boldsymbol{\eta}, \boldsymbol{\Lambda})$ and $\boldsymbol{\Sigma}^{-1} \sim W_i(\boldsymbol{\Omega}, \upsilon)$ with parameters $a = 4, b = 8, \boldsymbol{\Lambda} = (0.4, 0.01, 0.3, 0.1, 100, 0.1), \boldsymbol{\Omega} = (2.0, 0.5, 2.0, 2.0, 1.25, 2.5)$ and $\upsilon = 5$ as in previous studies [5, 6]. Also, $\boldsymbol{\eta} = log_{10}(d_{1_0}, \varepsilon_0, k_{1_0}, \delta_0, \pi_{1_0}, c_0)$ is set as the vector of optimal parameters estimates resulting from the nonlinear programming algorithm described in Sect. 1.2.2.

After initialization, the Gibbs sampling steps update σ^{-2}, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}^{-1}$, whereas the Metropolis-Hasting algorithm updates $\boldsymbol{\theta}$. The values at state k, i.e. $\sigma^{-2(k)}$, $\boldsymbol{\mu}^{(k)}$ and $\boldsymbol{\Sigma}^{-1(k)}$, are used to generate a candidate for $\boldsymbol{\theta}^{(k)}$, denoted as $\boldsymbol{\varphi}$. Then, the acceptance probability is evaluated and, if higher than a uniform (0, 1) generated random number, the candidate $\boldsymbol{\varphi}$ is accepted and $\boldsymbol{\theta}^{(k)} = \boldsymbol{\varphi}$. Note that there is an implicit relation between σ, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, and the model in (2) through the evaluation of π in the acceptance probability. It is clear that π depends of $y(t_i) - f(., t_i)$, which

Table 1 Values for the parameters $\boldsymbol{\theta} = (d_1, \varepsilon = 0, k_1, \delta, \pi_1, c)$ used for the simulation of 100 replicas per patient [12]. See [4, 12] for further details on parameters range

Patient	d_1 (day^{-1})	k_1 (ml day^{-1})	δ (day^{-1})	π_1 (day^{-1})	c (day^{-1})
1	0.013	0.46×10^{-6}	0.40	980	3
2	0.012	0.75×10^{-6}	0.39	790	3
3	0.017	0.80×10^{-6}	0.31	730	3
Range	0.01–0.02	10^{-11}–10^{-5}	0.24–0.70	50–10^4	2.39–23

is the error term $e(t_i)$ in model in 2. While the computation of $f(\boldsymbol{\varphi}, t_i)$ indirectly depends of $\sigma^{-2(k)}$, $\boldsymbol{\mu}^{(k)}$ and $\boldsymbol{\Sigma}^{-1(k)}$, as $\boldsymbol{\varphi}$ is generated from $\sigma^{-2(k)}$, $\boldsymbol{\mu}^{(k)}$ and $\boldsymbol{\Sigma}^{-1(k)}$, the computation of $f(\boldsymbol{\theta}^{(k-1)}, t_i)$ indirectly depends of $\sigma^{-2(k-1)}$, $\boldsymbol{\mu}^{(k-1)}$ and $\boldsymbol{\Sigma}^{-1(k-1)}$, as $\boldsymbol{\theta}^{(k-1)}$ was generated from $\sigma^{-2(k-1)}$, $\boldsymbol{\mu}^{(k-1)}$ and $\boldsymbol{\Sigma}^{-1(k-1)}$, and accepted in the previous state of the chain. Afterwards, the chain starts over in the new state $k+1$ until 6000 candidates are accepted, i.e., at the end of this process, the algorithm produces $\boldsymbol{\theta}^{(k)}$, $k = 1, 2, \ldots, 6000$ vectors, to be used in subsequent analyses. In the initialization ($k = 1$) and actualization ($k > 1$) steps, the (physiological) validity of the model is tested by checking whether $\boldsymbol{\theta}^{(k)} = (d_1, \varepsilon, k_1, \delta, \pi_1, c)$ falls within the range of values in Table 1.

1.2.2 Optimal $\boldsymbol{\eta}$ Estimate

As stressed above, an acceptable $\boldsymbol{\eta}$ value for the MCMC method is obtained through the nonlinear programming algorithm. As before, $CD4(t_i)$ is the observed number of CD4$^+$T cells at time t_i, $i = 1, 2, \ldots, n$, and $f(\boldsymbol{\eta}, t_i) = \widehat{CD4}(t_i) = T(t_i) + T^*(t_i)$ is the value of $CD4(t_i)$ provided by the model (1) for a given $\boldsymbol{\eta}$. The *optimal* parameter estimates, say $\widehat{\boldsymbol{\eta}}$, can be obtained by minimizing the sum of squared errors between the model estimates and the observed CD4 values. Thus, the nonlinear programming algorithm can be formulated as

$$\begin{aligned} \text{minimize} \quad & g(\boldsymbol{\eta}) = \sum_{i=1}^{n} (\widehat{CD4}(t_i) - CD4(t_i))^2 \\ \text{subject to} \quad & \sum_{i=1}^{n} (\widehat{CD4}(t_i) - CD4(t_i)) = 0 \\ \text{and} \quad & \boldsymbol{lb} \leq \boldsymbol{\eta} \leq \boldsymbol{ub} \end{aligned} \tag{3}$$

where the restriction guarantees that the numerical solution $\widehat{\boldsymbol{\eta}}$ verifies that property of the minimum least squares method (i.e. equal contribution of negative and positive deviations from observations). Finally, $\widehat{\boldsymbol{\eta}}$ is restricted to physiological lower and upper bounds, respectively $\boldsymbol{lb} = (0.01, 0, 10^{-11}, 0.24, 50, 2.39)$ and $\boldsymbol{ub} = (0.02, 1, 10^{-5}, 0.7, 10000, 23)$ [4, 12].

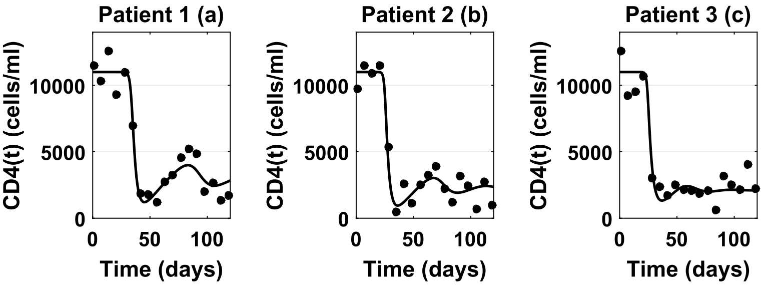


Fig. 1 **a**, **b**, **c**: $CD4(t)$ trajectory over time of the patients with the $\boldsymbol{\theta}$ parameters in Table 1. The circles represent the observations obtained for one replica of each patient

The optimization procedure was implemented with MATLAB™ function *fmincon*, that starts at an initial solution $\boldsymbol{\eta}^*$ to find a minimizer $\widehat{\boldsymbol{\eta}}$ of $g(\boldsymbol{\eta})$ subject to the above-mentioned restrictions and bounds. The initial solution $\boldsymbol{\eta}^*$ is obtained as that minimizing $g(\boldsymbol{\eta})$ in a set of 1000 candidates randomly generated from a multivariate uniform distribution on $\boldsymbol{lb}$ and $\boldsymbol{ub}$. The optimization procedure in (3) was first introduced by [11] and validated through a simulation study. We concluded that the parameters $\boldsymbol{\eta}$ (in Table 1) used for data simulation can be properly replaced by the estimates $\widehat{\boldsymbol{\eta}}$ obtained from the optimization procedure, as such replacement does not lead to a relevant impact on the optimum $g(\cdot)$ value. Moreover the results in [11] showed that $\widehat{\boldsymbol{\eta}}$ provides a better fit to the data than $\boldsymbol{\eta}$ because the optimization procedure produces a fine tuning to the data. Finally, it is important to stress that the replacement of $\boldsymbol{\eta}$ by $\widehat{\boldsymbol{\eta}}$ is mandatory when dealing with real data, because reference parameters $\boldsymbol{\eta}$ are unknown in such cases. The algorithms and other software used in this work were implemented in MATLAB™ (version R2015a), The Mathworks Inc., MA, USA.

2 Simulated Data and Results

The methods described above were tested on simulated data. Equally spaced $CD4(t)$ and $V(t)$ observations were generated in the interval [0, 120](days), by numerical Runge–Kutta integration of (1). We reproduced the evolution of three HIV patients with parameters presented in Table 1 and initial conditions $(T_0, T_0^*, V_0) = (11 \times 10^3, 0, 10^{-6})$. In this way, we obtain a set of $n = 18$ observations representing the temporal trajectory of a patient in clinical follow-up every 7 days. Afterwards, 100 replicas of the trajectory are randomly generated, by adding an error to the CD4$^+$T values following a zero mean Normal distribution with standard deviation equal to 20% of the measured value [13]. Figure 1a–c shows one replica of each patient and highlights the similarities and differences among patients. After estimation, the chains associated with the model's parameters were analyzed with respect to autocorrelation and convergence. As expected, the chains exhibited serial dependence in all cases (replicas and parameters). Therefore, for each replica, we considered the lowest sampling interval of the chain values (up to 6) for which the sample autocorrelation was within the 95% confidence band for a null autocorrelation. With this procedure, the initial 6000 samples were reduced to e.g. 2000, 1500, 1200 or 1000, depending of the replica. Finally, the Geweke and the Gelman and Rubin tests were performed to analyze convergence [2, 3]. Convergence was not rejected for all cases, as results in Table 2 show the G test statistics inside the interval $(-1.96, 1.96)$ of 95% confidence for equality of the means (of the first 10% and last 50% of each Markov chain) and the GR factors close to 1. Figure 2 presents the a posteriori distributions of the parameters $\boldsymbol{\theta}$, for the replicas presented in Fig. 1. In particular, the distributions associated with d_1 hold clear similarities for patients 1 and 2, and both show large differences to that of patient 3. This result was expected due to the d_1 values associated with each patient: similar d_1 values for patients 1 and 2, and distinct from that of patient 3 (see Table 1). For the remaining parameters, the differences between distributions are barely perceptible. Therefore, we considered the multivariate parameter $d = \sum_{i=1}^{n}(\widehat{CD4}(t_i) - CD4(t_i))$ to enhance inter-subject differences. The parameter d allows the simultaneous analysis of $\boldsymbol{\theta}$ values, as d is evaluated from $\widehat{CD4}(t_i) = T(t_i) + T^*(t_i)$ with a certain combination of values in $\boldsymbol{\theta}$. Therefore, each combination of values in $\boldsymbol{\theta}$ has associated just one value of d. Figure 3 resumes the analysis for patient 1. The distribution of d is obtained by evaluation of $\widehat{CD4}$ for all combinations of $\boldsymbol{\theta}$ values obtained from the multivariate a posteriori distribution. The vertical line locates $d = 0$ i.e. the combination with null deviation from observed to estimated data. The corresponding temporal trajectory is well adjusted to the simulated data and exhibits clear similarities to the temporal trajectory from the simulation parameters (Fig. 1). Also, the percentiles 25 and 75 of d distribution constitute a variation band containing 50% of the most central deviations and allow to obtain a variation band containing the most *central* temporal trajectories for each patient. Note that the 50% band of the patient trajectories is constructed from the distribution of d and, as expected, $d_{0.25} < d_{0.75}$ where d_p is the *pth*-percentile of d. However, note that the $\widehat{CD4(t_i)}$ values evaluated for $\boldsymbol{\theta}_{0.25}$ and $\boldsymbol{\theta}_{0.75}$ do not necessarily maintain the same ordering. Instead, these curves constitute the range of variation of the curves with lowest squared errors. In order to compare different antiretroviral therapy (ART) conditions, we further consider simulation cases with $\varepsilon \neq 0$ together with the parameters presented in Table 1, for all patients. Note that high values of ε mimic the effect of highly effective ART where the CD4$^+$T trajectories tend to be constant for longer time periods and, at the limiting value $\varepsilon = 1$, it reproduces a constant line. Therefore, we consider $\varepsilon \in \{0.1, 0.25, 0.35\}$ so to reproduce a variety of CD4$^+$T trajectories with an accentuated decay after the constant initial step. However, the set of simulation parameters theta [see table 1] and $\varepsilon \neq 0$ does not guarantee that the simulation case reproduces a real condition. Therefore, these simulation cases are presented solely to evaluate the estimation performance when $\varepsilon \neq 0$. Figure 4 resumes the results for patient 1 and shows that the performance in estimation for $\varepsilon \in \{0.1, 0.25, 0.35\}$ is fairly similar to that obtained for the case

Table 2 Chains' convergence analysis from Geweke test (G test statistics, first 3 lines of the table) and Gelman and Rubin diagnostic test (GR factor, last 3 lines of the table). The values are mean ± standard deviation evaluated for 100 replicas of each patient

Patient	$log_{10}(d_1)$	$log_{10}(k_1)$	$log_{10}(\delta)$	$log_{10}(\pi_1)$	$log_{10}(c)$
1	-0.041 ± 0.936	0.056 ± 1.031	-0.082 ± 1.114	-0.018 ± 1.086	-0.081 ± 1.016
2	0.011 ± 0.959	-0.132 ± 1.102	-0.047 ± 1.002	0.092 ± 1.127	0.125 ± 0.908
3	-0.051 ± 1.099	-0.124 ± 1.126	0.043 ± 1.121	0.060 ± 1.160	-0.059 ± 1.025
1	1 ± 0.001	1 ± 0.001	1 ± 0.001	1 ± 0.001	1 ± 0.001
2	1 ± 0.001	1 ± 0.001	1 ± 0.001	1 ± 0.001	1 ± 0.001
3	1 ± 0.002	1 ± 0.001	1 ± 0.001	1 ± 0.002	1 ± 0.001

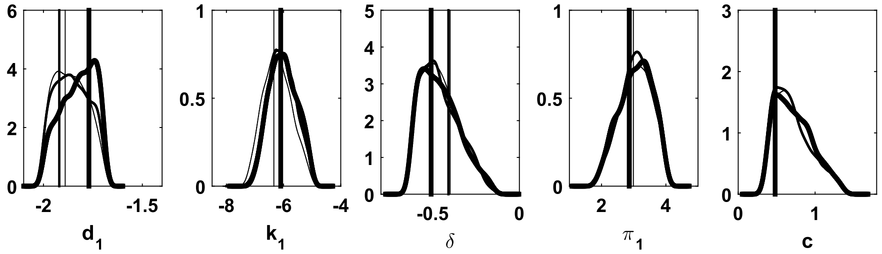


Fig. 2 A posteriori distributions for the $\boldsymbol{\theta} = (d_1, \varepsilon, k_1, \delta, \pi_1, c)$ parameters obtained for the replicas presented in Fig. 1. The vertical lines locate the simulation parameters displayed in Table 1

$\varepsilon = 0$. In particular, the central trajectories are well aligned with the simulated data and the variation bands reproduce adequate temporal trajectories of the patient. Note that the simulated data seems to exhibit a larger variability for increasing values of ε. This pattern is properly reproduced by the width of the variation bands which seem to be wider for increasing values of ε. With respect to the shape of the CD4$^+$T curve, its variation is explained by the distinct parameter values (ε included), which are estimated by the statistical method, and then sent to the ODE model to generate the trajectories. Thus, the shape reflects the change in several parameter values and not only in one.

3 Conclusion

In this work, the estimation of the parameters of a mathematical model is carried out by a Bayesian approach with initial conditions obtained from an optimization procedure based on nonlinear programming. The results suggest that the proposed method allows to obtain trajectories and variation bands that adequately describe the simulated data. The MCMC based approaches are known to be computationally expensive and largely dependent on the initial values of the hyper-parameters. We

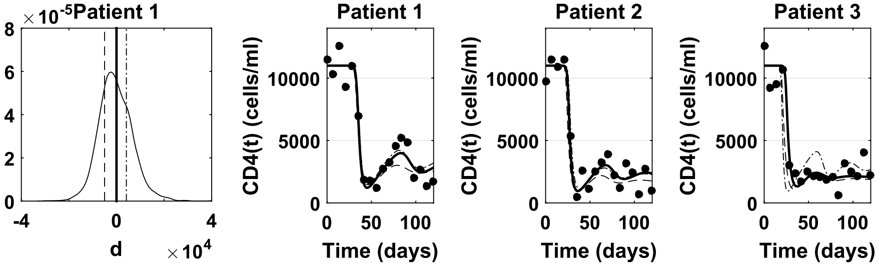


Fig. 3 Summary of results for the replica of patient 1 (see Fig. 1): distribution of d (with location of $d = 0$, and percentiles 25 and 75 of the distribution) and corresponding CD4$^+$T trajectories superimposing the simulated data (circles). Same curves representation for patients 2 and 3

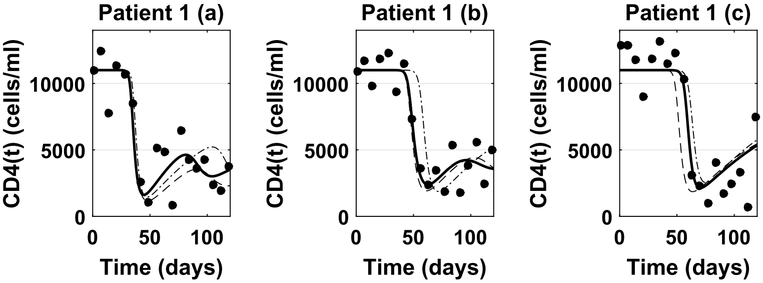


Fig. 4 CD4$^+$T trajectories for percentiles 25 and 75 of the distribution d superimposing the simulated data (circles) for the simulations of patient 1 for **a** $\varepsilon = 0.1$. **b** $\varepsilon = 0.25$. **c** $\varepsilon = 0.35$

deal with the latter limitation by providing optimal initial values to the Bayesian approach. Even so, the computational efficiency can still be improved by adopting e.g. Approximate Bayesian Computation (ABC) schemes that avoid the evaluation of the likelihood function. Future developments will also consider the evaluation of the proposed method in real clinical data, obtained from the follow-up of HIV infected patients.

Acknowledgements This work was partially funded by the Foundation for Science and Technology, FCT, through national (MEC) and European structural (FEDER) funds, in the scope of UID/MAT/04106/2019 (CIDMA/UA), UID/CEC/00127/2019 (IEETA/UA) and UID/MAT/00144/2019 (CMUP/UP) projects. Diana Rocha acknowledges the FCT grant (ref. SFRH/BD/107889/2015). This work was also partially funded by Portugal 2020 under the Competitiveness and Internationalization Operational Program, and by the European Regional Development Fund through project SOCA-Smart Open Campus (CENTRO-01-0145-FEDER-000010).

References

1. Bonhoeffer, S., May, R.M., Shaw, G.M., Nowak, M.A.: Virus dynamics and drug therapy. Proc. Natl. Acad. Sci. USA **94**, 6971–6976 (1997)
2. Gelman, A., Rubin, D.B.: Inference from iterative simulation using multiple sequences. Stat. Sci. **169**, 457–472 (1992)
3. Geweke, J.: Evaluating the accuracy of sampling based approaches to the calculation of posterior moments. Bayesian Stat. **4**, 169–193 (1992)
4. Hadjiandreou, M.M., Conejeros, R., Wilson, D.I.: Long-term HIV dynamics subject to continuous therapy and structured treatment interruptions. Chem. Eng. Sci. **64**, 1600–1617 (2009)
5. Huang, Y., Liu, D., Wu, H.: Hierarchical bayesian methods for estimation of parameters in a longitudinal HIV dynamic system. Biometrics **62**, 413–423 (2006)
6. Huang, Y., Wu, H., Acosta, E.P.: Hierarchical bayesian inference for HIV dynamic differential equation models incorporating multiple treatment factors. Biom. J. **52**, 470–486 (2010)
7. Liang, H., Miao, H., Wu, H.: Estimation of constant and time-varying dynamic parameters of HIV infection in a nonlinear differential equation model. Ann. Appl. Stat. **4**, 460483 (2010)
8. Nowak, M.A., May, R.M.: Virus Dynamics: Mathematical Principles of Immunology and Virology. Oxford University Press, Oxford (2000)
9. Perelson, A.S.: Modelling viral and immune system dynamics. Nat. Rev. Immunol. **2**, 28–36 (2002)
10. Perelson, A.S., Kirschner, D.E., Boer, R.: Dynamics of HIV infection of CD4^{+}T cells. Math. Biosci. **114**, 81–125 (1993)
11. Rocha, D., Gouveia, S., Pinto, C., Scotto, M., Tavares, J.N., Valadas, E., Caldeira, L.F.: On the parameters estimation of HIV dynamic models. In: Proceedings of the III Portuguese-Galician Meeting of Biometry (EBIO2018), University of Aveiro, Portugal, June 28–30, pp. 57–60 (2018)
12. Stafford, M.A., Coreya, L., Caob, Y., Daardd, E.S., Hob, D.D., Perelson, A.S.: Modeling plasma virus concentration during primary HIV infection. J. Theor. Biol. **203**, 285–301 (2000)
13. Whitby, L., Whitby, A., Fletcher, M., Helbert, M., Reilly, J.T., Barnett, D.: Comparison of methodological data measurement limits in CD4^{+}T lymphocyte flow cytometric enumeration and their clinical impact on HIV management. Cytom. Part B (Clin. Cytom.) **84**, 248–254 (2013)
14. Wu, L.: A joint model for nonlinear mixed-effects models with censoring and covariates measured with error, with application to AIDS studies. J. Am. Stat. Assoc. **97**, 955–964 (2002)

Modelling with Non-stratified Chain Event Graphs

Aditi Shenvi, Jim Q. Smith, Robert Walton and Sandra Eldridge

Abstract Chain Event Graphs (CEGs) are recent probabilistic graphical modelling tools that have proved successful in modelling scenarios with context-specific independencies. Although the theory underlying CEGs supports appropriate representation of structural zeroes, the literature so far does not provide an adaptation of the vanilla CEG methods for a real-world application presenting structural zeroes also known as the non-stratified CEG class. To illustrate these methods, we present a non-stratified CEG representing a public health intervention designed to reduce the risk and rate of falling in the elderly. We then compare the CEG model to the more conventional Bayesian Network model when applied to this setting.

Keywords Bayesian networks · Bayesian statistics · Chain event graphs
Event tree · Public health intervention

1 Introduction

The development of Chain Event Graphs (CEGs) which were first introduced in [22] was motivated by the need for a probabilistic graphical modelling tool that can handle asymmetric information. Such asymmetries may present as context-specific

A. Shenvi (✉)
Centre for Complexity Science, University of Warwick, Coventry, UK
e-mail: a.shenvi@warwick.ac.uk

J. Q. Smith
Department of Statistics, University of Warwick, Coventry, UK
e-mail: j.q.smith@warwick.ac.uk

The Alan Turing Institute, London, UK

R. Walton · S. Eldridge
Centre for Primary Care and Public Health, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, UK
e-mail: rtwalton123@gmail.com

S. Eldridge
e-mail: s.eldridge@qmul.ac.uk

R. Argiento et al. (eds.), *Bayesian Statistics and New Generations*,
Springer Proceedings in Mathematics & Statistics 296,
https://doi.org/10.1007/978-3-030-30611-3_16

conditional independence relations and structural zeroes. It has been shown that CEGs can be elicited from expert opinion, admit fast conjugate estimation, closed-form learning and model selection under complete sampling [8, 11, 21]. Various conditional independence statements can be read from the topology of a CEG without reference to its embellished edge probabilities [22, 25]. This enables any elicitation or explanation of the model class to be translated to and from natural language—a vital property for efficient decision support.

CEGs are constructed from event trees through the process of creating stages and merging vertices whose rooted subtrees have isomorphic stage structures as described in Sect. 2. The event tree [19] for a real-world system may have some of its branches unpopulated. Zeroes present in the data could be one of two types: sampling zeroes or structural zeroes. While the former refers to unobserved values due to sampling limitations, the latter refers to a logical impossibility of observing a non-zero value. Event trees can represent such information succinctly by simply deleting the unpopulated edges where the absence of a non-zero value can be justified to be a logical constraint. The CEG, being a function of its underlying event tree, inherits the property of embedding such information directly in its structure.

When it comes to modelling asymmetries, CEGs have proven to be more successful than other graphical models such as Bayesian Networks (BNs) which were popularised by Pearl as a tool for causal analysis [17] and for reading conditional independence statements directly from the topology of the graph through the process of d-separation [27]. Whilst the BN is a powerful modelling tool, it is unable to embed context-specific conditional independence relations directly into its graphical representation. In order to exploit such information from a BN, we need to make adjustments to the model using methods such as those proposed in [3, 12, 18]; none of which result in a unified graphical model for encapsulating these context-specific details. Whereas CEGs have been useful in modelling context-specific conditional independencies in domains such as public health and security [2, 7].

Another shortcoming of the BN is its inability to express structural zeroes in its topology. We've already stated how CEGs, through their relationship with event trees, are equipped to handle structural zeroes. Such CEGs are called non-stratified CEGs. To the best of our knowledge, the current literature does not demonstrate this property of the CEG. The aim of this paper is to represent a real-world process with structural zeroes using a non-stratified CEG and to compare it to a BN for the same process. We illustrate this by modelling a public health intervention to alleviate the risk of falls in the elderly based on the work developed in [9].

2 Preliminary Concepts

In this section we review the definitions of an event tree, a CEG and a BN as well as other associated concepts. For a detailed description of CEGs see [22]. For more details on BNs see [14, 17].

Event Tree: An event tree $\mathcal{T}$ = (V($\mathcal{T}$), E($\mathcal{T}$)) is an acyclic, connected, directed graph. The vertices represent events that may occur to a unit during the unfolding of the process being modelled. There exists only one vertex with no parents, this vertex is the root vertex s_0. All other vertices have exactly one parent. The vertices with no children are called *leaves*. Let L($\mathcal{T}$) be the set of leaves. The non-leaf vertices are called *situations*. Let S($\mathcal{T}$) form the set of situations where S($\mathcal{T}$) = V($\mathcal{T}$)\L($\mathcal{T}$). Let $\mathbb{X}(s_i)$ be the set of children of situation s_i.

Floret: A floret $\mathcal{F}$ of situation s_i is defined as $\mathcal{F}(s_i) = (V(\mathcal{F}(s_i)), E(\mathcal{F}(s_i)))$ where $V(\mathcal{F}(s_i)) = s_i \cup \mathbb{X}(s_i)$ and $E(\mathcal{F}(s_i))$ is the set of edges induced by $V(\mathcal{F}(s_i))$ in $\mathcal{T}$.

Stage: Two situations s_i and s_j in $\mathcal{T}$ are said to be in the same stage u if and only if there exists a bijection ψ_u under which X(s_i) and X(s_j) have the same distribution given by $\psi_u(s_i, s_j) : \mathbb{X}(s_i) \rightarrow \mathbb{X}(s_j)$, where ψ_u leads to a mapping of the edge labels which is meaningful for the real-world application.

Staged Tree: In event tree $\mathcal{T}$, situations in the same stage are given the same colour. A coloured event tree $\mathcal{T}$ is called a staged tree $\mathcal{ST}$.

Position: Two situations s_i and s_j in $\mathcal{T}$ are said to be in the same position w if and only if the staged trees $\mathcal{ST}_{s_i}$ and $\mathcal{ST}_{s_j}$ rooted at s_i and s_j respectively are isomorphic in the sense of isomorphism between coloured graphs.

Chain Event Graph: A Chain Event Graph is obtained from its underlying staged tree by collapsing situations in the same position into one vertex and adding an additional vertex w_∞ into which all the leaf vertices are collected. Only the stages containing situations which are in the same stage but not in the same position retain their colouring in the CEG.

Bayesian Network: A BN X is represented by a directed acyclic graph (DAG) G = (V, E) where the nodes represent random variables. The lack of an edge between two nodes of G represents conditional independence between them while a directed edge encodes information about the conditional dependencies between them. For Pa$_i$ denoting parents of X_i, a BN on variables X_i, $i = 1, \ldots, n$ has joint probability distribution given by $\mathbb{P}(X_1, \ldots, X_n) = \prod_{i=1}^{n} \mathbb{P}(X_i|\text{Pa}_i)$.

3 Intervention

Falls-related injuries and fatalities are a serious problem among the elderly. According to NICE guidelines [10], 30% of people older than 65 and 50% of people older than 80 fall at least once a year. Eldridge et al. [9] modelled an intervention to reduce falls in the elderly using a probability tree for short-term analysis and a Markov model for long-term effectiveness. The intervention was designed to enhance assessment, referral pathways and treatment for high-risk individuals aged over 65 years who have a substantial risk of falling. After assessment, individuals are classified as high-risk or low-risk of falling as per the recommendations in FRAT (Falls Risk Assessment Tool) [15]. In our model, we have additionally classified individuals by their type of residence as we learned from domain literature that the fall rates are higher for individuals living in institutionalised care [16]. We incorporate this information by setting the proportion of high-risk individuals in communal establishments higher than in the community. The variables used to study this intervention are $\mathcal{X} = \{X_A, X_R, X_T, X_F\}$. Here X_A indicates whether the individual aged over 65 resides in the community or in a communal establishment (such as nursing homes, care homes, hospitals) and whether they have been assessed or not; X_R indicates the risk level of the individual as high or low; X_T indicates whether the individual has been referred & treated, not referred & treated or not treated; and X_F indicates whether the individual suffered from a fall or not. As per the design of this intervention, all referred persons are treated and all those who fall under the not treated category have not been referred. Low-risk individuals do not receive referrals for treatment and hence we only classify them as treated or not treated.

4 Methods

We begin by constructing a staged tree for the intervention as shown in Fig. 1. This tree is non-stratified as the treatment variable X_T has no logical interpretation for individuals who have not been assessed as by intervention design they cannot receive any treatment. This staged tree is our data generating model. We simulated data for 50,000 individuals passing through this system by forward sampling. The numbers along the edges represent the observations along each branch. Observe that several of the branches are sparsely populated. For instance, there are only two observations along the edge indicating falls suffered by assessed low-risk individuals in communal establishments who received treatment. Sparsely populated branches may pose a problem for model selection. We discuss this further in Sect. 6.

We use a weakly informative equivalent sample size (α) of 4 and set the parameters of the Dirichlet prior on each situation of the tree using the mass conservation property as described in [6]. All CEG structures are assumed a priori equally likely. We use an adapted form of the greedy Agglomerative Hierarchical Clustering (AHC) algorithm developed in [11] to fit a CEG to the data and evaluate it using the Bayesian Dirichlet equivalent uniform (BDeu) score developed in [4]. The AHC algorithm in [11] sequentially merges situations into stages by combining the two situations at every step which give the highest improvement to the BDeu score of the graph. This was further refined by the introduction of the concept of hyperstages in [5]. A hyperstage consists of sets of situations such that two situations can be merged into a single stage if and only if they belong to the same set in the hyperstage. This allows further information about the domain to be embedded into the model and can also greatly reduce the dimension of the model search space. Under this adaptation, we used the domain knowledge to set the hyperstage structure and used this structure to run the AHC algorithm to output the best fitting CEG to the data.

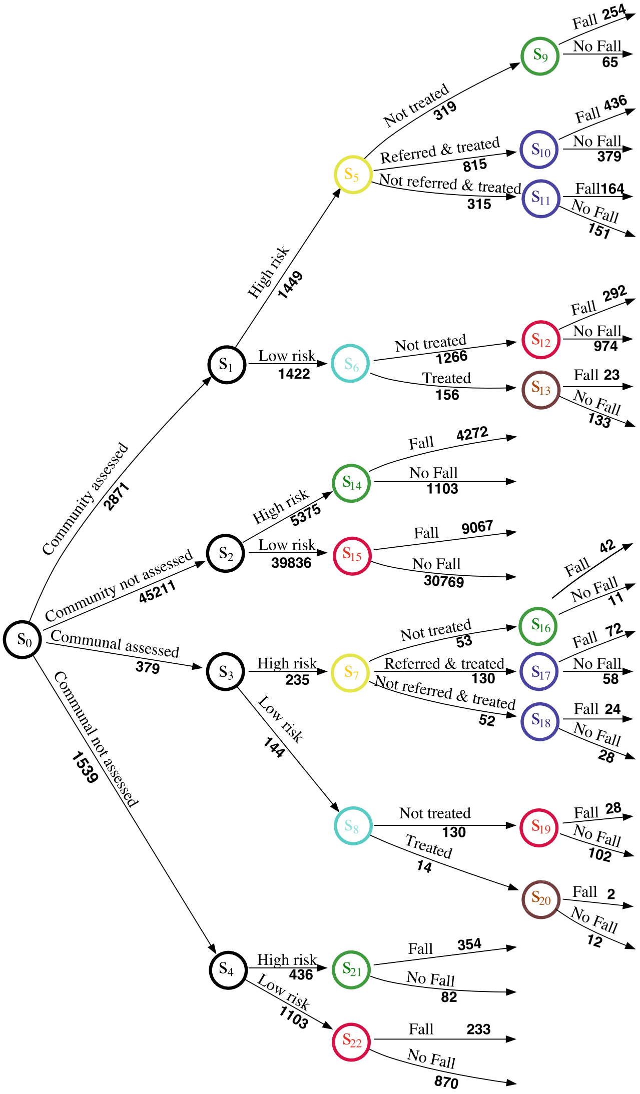


Fig. 1 Staged tree representing the data generating model

To learn the BN structure, we use the Hill-Climbing algorithm from the R package **bnlearn**. We compared the two graphical models based on their BDeu score as well as their ability to express context-specific information and structural zeroes which are essential in assessing the effectiveness of the intervention. The CEG model is then analysed further for its robustness and ability to incorporate causal manipulations.

5 Results

The Hill-Climbing algorithm outputs the BN in Fig. 2a. As the intervention gives rise to a total order of $X_A \preceq X_R \preceq X_T \preceq X_F$, certain edges must be suppressed in order for the BN to be representative of the real-world application. For instance, the directed edge from *Treatment* to *Risk* is not permissible given the total order. This gives rise to the BN in Fig. 2b. The BDeu score of this BN structure is -68709.99.

We set up the hyperstage structure for the AHC algorithm as $[\{s_1, s_2, s_3, s_4\}, \{s_5, s_7\}, \{s_6, s_8\}, \{s_9, s_{10}, s_{11}, s_{14}, s_{16}, s_{17}, s_{18}, s_{21}\}, \{s_{12}, s_{13}, s_{15}, s_{19}, s_{20}, s_{22}\}]$. The best-fitting CEG returned by the AHC is given in Fig. 3 when we assign the situation priors using α set to 4. Note that the stage structure of this CEG is exactly the same as the stage structure of the data generating tree given in Fig. 1. The BDeu score of this CEG is -68671.59, thus giving an extremely high Bayes Factor of 4.7523×10^{16} in favour of the CEG model. By Kass and Raftery's interpretation [13], this indicates *very strong evidence* in favour of the CEG model being a better fit to the data.

Additionally, from the topology of the CEG in Fig. 3, we can read the following context-specific conditional independence statements using *cuts* as defined in [22]:

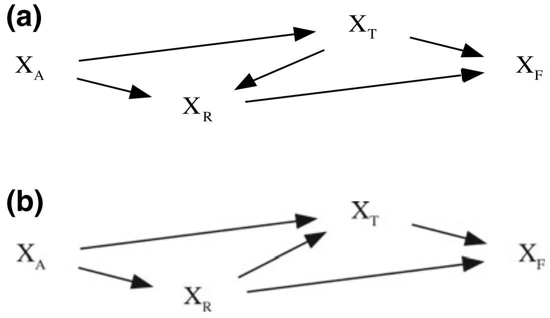


Fig. 2 **a** Original BN returned using the Hill-Climbing algorithm; **b** Best-fitting BN which admits the total order of $X_A \preceq X_R \preceq X_T \preceq X_F$

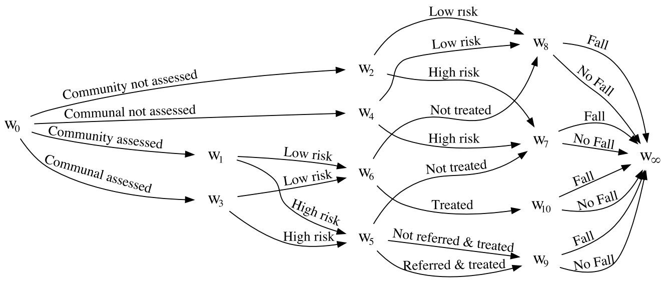


Fig. 3 CEG returned by the AHC algorithm

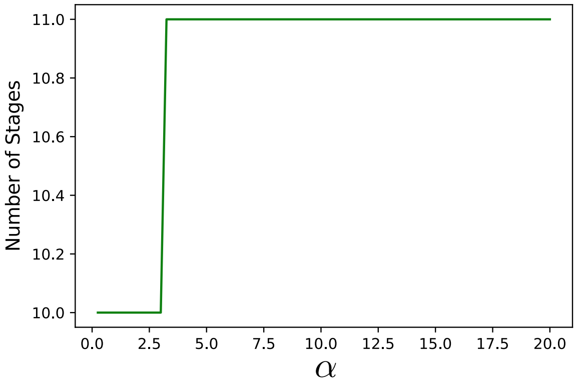


Fig. 4 The CEG model is robust given varying values of α

$$X_T \perp\!\!\!\perp X_R \mid \{X_A = \text{Not Assessed}\}, \tag{1}$$

$$X_T \not\perp\!\!\!\perp X_R \mid \{X_A = \text{Assessed}\}, \tag{2}$$

whereas Fig. 2b fails to show these contextual independencies.

Sensitivity to the choice of priors is one of the factors influencing the robustness of the CEG. The parameter α determines the prior assigned to each situation of the event tree. Thus to assess the robustness of the CEG, we fit the data using AHC to varying values of the equivalent sample size α and compare the number of stages at each value of α between 0.25 and 20 with increments of 0.25 as shown in Fig. 4. The number of stages remains steadily at 11 for α greater than three. Moreover, we observe that the stage structure also remains the same for α greater than 3.

CEGs admit exploration of causal hypotheses through manipulations under certain conditions as described in detail in [2, 24, 26]. Such CEGs are called causal CEGs. Manipulations in CEGs can be asymmetric as it is possible to intervene in certain positions and not necessarily on the entire variable. For instance, assuming our non-stratified CEG in Fig. 4 is also a causal CEG, we may wish to examine the effect of treatment given to all assessed high-risk individuals irrespective of their type of residence. This would result in the deletion of edge w_5 to w_7 and the edge labelled "not referred & treated" from w_5 to w_9. The edge probabilities remain unchanged except that the probability of traversing the remaining w_5 to w_9 edge conditional on reaching w_5 is one.

6 Discussion

We observed in Sect. 5 that the CEG is robust against varying values of α. For α greater than three, the resultant CEG was the data generating model. The smaller values of α were unable to return this due to sparsity along some of the branches of the event tree. In the falls intervention scenario we had domain literature to support the veracity of the staging produced by the AHC. In the absence of sufficient domain information, situations whose emanating edges contain sparse data-points should be maintained as singletons in the hyperstage structure. Failing this, spurious stages may be merged by the AHC resulting in an unreliable CEG.

In this paper, we compared two types of graphical models—Bayesian Networks and Chain Event Graphs—demonstrating the superiority of CEGs in modelling structural zeroes and context-specific independencies by modelling a falls intervention. In [9], which provided the falls intervention design, the effectiveness of the intervention was assessed using a probability tree for the short-term and a Markov model for the long-term. We briefly outline the advantages of CEGs over these modelling techniques when presented with a scenario such as the falls intervention.

A probability tree is the same as an uncoloured event tree. Unlike the BN, probability trees and Markov models can satisfactorily express asymmetric information introduced by structural zeroes. However, an essential property which these models lack is the ability to read conditional independence statements from their topology. They also do not admit causal manipulations. The colouring of the staged tree, followed by the collapsing of vertices in the CEG not only provide a succinct and complete description of the various paths that an individual may traverse as they pass through the system but also allow us to read context-specific conditional independence relationships and under a legitimate causal setting, can be subject to a causal analysis. These properties are particularly useful for modelling multi-factorial interventions where there are several different components of the intervention whose contributions and effects may not be trivially quantified or analysed.

It is important to note that the Markov model in [9] was for assessing long-term effectiveness. The CEG described in this paper caters to short-term analysis. Dynamic variants of the CEG have been developed in [1, 7]. For applications such as the falls intervention, we observe that individuals take varying amounts of time to move from certain states. For instance, individuals living in the community who have been assessed and have been referred and treated will not all fall. Also, those who do suffer from a fall would do so after varying amounts of time since they received their treatment. In such settings, it is also typical to record observations when a transition occurs rather than recording them at regular intervals. For instance, we would note that an individual has suffered a fall and the time that has elapsed since their treatment but we are unlikely to record every day or every month that an individual has not suffered a fall. This type of setting corresponds more closely to a semi-Markov process rather than a Markov process. However, like a Markov model, a semi-Markov model will not allow reading of conditional independence statements. For this purpose, we have developed a dynamic variant of the CEG that has an underlying semi-Markov structure. This is called the Reduced Dynamic Chain Event Graph (RDCEG). An early application of this class of models can be found in [23] and an associated technical paper is in preparation [20].

Acknowledgements Jim Q. Smith was supported by the Alan Turing Institute and funded by the EPSRC [grant number EP/K03 9628/1].

References

1. Barclay, L.M., Collazo, R.A., Smith, J.Q., Thwaites, P.A., Nicholson, A.E.: The dynamic chain event graph. Electron. J. Stat. **9**(2), 2130–2169 (2015)
2. Barclay, L.M., Hutton, J.L., Smith, J.Q.: Refining a Bayesian network using a chain event graph. Int. J. Approx. Reason. **54**(9), 1300–1309 (2013)
3. Boutilier, C., Friedman, N., Goldszmidt, M., Koller, D.: Context-specific independence in Bayesian networks. In: Proceedings of the Twelfth international conference on uncertainty in artificial intelligence, pp. 115–123. Morgan Kaufmann Publishers Inc. (1996)
4. Buntine, W.: Theory refinement on Bayesian networks. In: Proceedings of the seventh conference on uncertainty in artificial intelligence, pp. 52–60. Morgan Kaufmann Publishers Inc. (1991)
5. Collazo, R.A.: The dynamic chain event graph. Ph.D. Thesis, University of Warwick (2017)
6. Collazo, R.A., Görgen, C., Smith, J.Q.: Chain Event Graphs. Chapman & Hall/CRC, Boca Raton (2017)
7. Collazo, R.A., Smith, J.Q.: An N time-slice dynamic chain event graph. arXiv:1808.05726 (2018)
8. Cowell, R.G., Smith, J.Q.: Causal discovery through MAP selection of stratified chain event graphs. Electron. J. Stat. **8**(1), 965–997 (2014)
9. Eldridge, S., Spencer, A., Cryer, C., Parsons, S., Underwood, M., Feder, G.: Why modelling a complex intervention is an important precursor to trial design: lessons from studying an intervention to reduce falls-related injuries in older people. J. Health Serv. Res. Policy **10**(3), 133–142 (2005)
10. Falls in older people: assessing risk and prevention. In: Guidance and Guidelines — NICE. https://www.nice.org.uk/guidance/cg161 (2013)
11. Freeman, G., Smith, J.Q.: Bayesian MAP model selection of chain event graphs. J. Multivar. Anal. **102**(7), 1152–1165 (2011)
12. Friedman, N., Goldszmidt, M.: Learning Bayesian networks with local structure. In: Learning in Graphical Models, pp. 421–459. Springer, Dordrecht (1998)
13. Kass, R.E., Raftery, A.E.: Bayes factors. J. Am. Stat. Assoc. **90**(430), 773–795 (1995)
14. Korb, K.B., Nicholson, A.E.: Bayesian Artificial Intelligence. CRC Press, Boca Raton (2010)
15. Nandy, S., Parsons, S., Cryer, C., Underwood, M., Rashbrook, E., Carter, Y., Eldridge, S., Close, J., Skelton, D., Taylor, S.: Development and preliminary examination of the predictive validity of the Falls Risk Assessment Tool (FRAT) for use in primary care. J. Public Health **26**(2), 138–143 (2004)
16. Nurmi, I., Lüthje, P.: Incidence and costs of falls and fall injuries among elderly in institutional care. Scand. J. Prim. Health Care **20**(2), 118–122 (2002)
17. Pearl, J.: Causality. Cambridge University Press, Cambridge (2009)
18. Poole, D., Zhang, N.L.: Exploiting contextual independence in probabilistic inference. J. Artif. Intell. Res. **18**, 263–313 (2003)
19. Shafer, G.: The Art of Causal Conjecture. MIT press, Cambridge (1996)
20. Shenvi, A., Smith, J.Q.: The reduced dynamic chain event graph. arXiv:1811.08872 (2018)
21. Silander, T., Tze-Yun L.: A dynamic programming algorithm for learning chain event graphs. In: International Conference on Discovery Science, pp. 201–216. Springer, Berlin (2013)
22. Smith, J.Q., Anderson, P.E.: Conditional independence and chain event graphs. Artif. Intell. **172**(1), 42–68 (2008)
23. Smith, J.Q., Shenvi, A.: Assault crime dynamic chain event graphs. University of Warwick repository. http://wrap.warwick.ac.uk/104824/ (2018)
24. Thwaites, P.: Causal identifiability via chain event graphs. Artif. Intell. **195**, 291–315 (2013)
25. Thwaites, P., Smith, J.Q.: A separation theorem for chain event graphs. arXiv:1501.05215 (2015)
26. Thwaites, P., Smith, J.Q., Riccomagno, E.: Causal analysis with chain event graphs. Artif. Intell. **174**(12–13), 889–909 (2010)
27. Verma, T., Pearl, J.: Causal networks: semantics and expressiveness. In: Machine Intelligence and Pattern Recognition, vol. 9, pp. 69–76. North-Holland (1990)

Modelling Career Trajectories of Cricket Players Using Gaussian Processes

Oliver G. Stevenson and Brendon J. Brewer

Abstract In the sport of cricket, variations in a player's batting ability can usually be measured on one of two scales. Short-term changes in ability that are observed during a single innings, and long-term changes that are witnessed between matches, over entire playing careers. To measure long-term variations, we derive a Bayesian parametric model that uses a Gaussian process to measure and predict how the batting abilities of international cricketers fluctuate between innings. The model is fitted using nested sampling given its high dimensionality and for ease of model comparison. Generally speaking, the results support an anecdotal description of a typical sporting career. Young players tend to begin their careers with some raw ability, which improves over time as a result of coaching, experience and other external circumstances. Eventually, players reach the peak of their career, after which ability tends to decline. The model provides more accurate quantifications of current and future player batting abilities than traditional cricketing statistics, such as the batting average. The results allow us to identify which players are improving or deteriorating in terms of batting ability, which has practical implications in terms of player comparison, talent identification and team selection policy.

Keywords Cricket · Gaussian processes · Nested sampling

1 Introduction

As a sport, cricket is a statistician's dream. The game is steeped in numerous statistical and record-keeping traditions, with the first known recorded scorecards dating as far back as 1776. Given the statistical culture that has developed with the growth of cricket, using numeric data to quantify individual players' abilities is not a new

O. G. Stevenson (✉) · B. J. Brewer
University of Auckland, Auckland, New Zealand
e-mail: o.stevenson@auckland.ac.nz

B. J. Brewer
e-mail: bj.brewer@auckland.ac.nz

R. Argiento et al. (eds.), *Bayesian Statistics and New Generations*,
Springer Proceedings in Mathematics & Statistics 296,
https://doi.org/10.1007/978-3-030-30611-3_17

concept. However, despite the abundance of data available, cricket has only recently attracted the attention of statistical analysts in the public realm. This is potentially due to many previous academic studies being focused on the likes of achieving a fair result [6, 8, 11] and match outcome prediction [1, 3, 4, 17], rather than statistical applications that measure and predict individual player abilities and performances.

For as long as the modern game has existed, a player's batting ability has primarily been recognized using the *batting average*; in general, the higher the batting average, the better the player is at batting. However, the batting average fails to tell us about variations in a player's batting ability, which can usually be attributed to one of two scales: (1) short-term changes in ability that are observed *during* or *within* a single innings, due to factors such as adapting to the local pitch and weather conditions (commonly referred to as 'getting your eye in' within the cricketing community), and (2) long-term changes that are observed *between* innings, over entire playing careers, due to the likes of age, experience and fitness levels.

Early studies provided empirical evidence to support the claim that a batsman's score could be modelled using a geometric progression, suggesting players bat with a somewhat constant ability during an innings [10]. However, it has since been shown that the geometric assumptions do not hold for many players, due to the inflated number of scores of 0 that are present in many players' career records [2, 12]. Rather than model batting scores, Kimber and Hansford [12] and Cai et al. [5] used nonparametric and parametric hazard functions respectively, to measure how dismissal probabilities change with a batsman's score. Estimating a batsman's hazard function, $H(x)$, which represents the probability of getting out on score x, allows us to observe how a player's ability varies over the course of an innings. Both studies found that batsmen appeared to be more likely to get out while on low scores—early in their innings—than on higher scores, supporting the idea of 'getting your eye in'.

In order to quantify the effects of 'getting your eye in', Stevenson and Brewer [16] proposed an alternative means of measuring how player ability varies during an innings. The authors use a Bayesian parametric model to estimate the hazard function, allowing for a smooth transition in estimated dismissal probabilities between scores, rather than the sudden, unrealistic jumps seen in [12] and to a lesser extent [5]. For the vast majority of past and present international Test players, Stevenson and Brewer [16] found overwhelming evidence to suggest that players perform with decreased batting abilities early in an innings and improve as they score runs, further supporting the notion of 'getting your eye in'.

1.1 Modelling Between-Innings Changes in Batting Ability

While there is plenty of evidence to suggest that players do not bat with some constant ability during an innings, it is also unlikely that a player bats with some constant ability throughout their entire career. Instead, variations in a player's underlying ability are likely to occur *between* innings, due to factors such as how well the player has been performing recently (referred to as 'form' in cricket).

If batting form were to have a significant impact on player performance, we should be able to identify extended periods of players' careers with sequences of high scores (indicating the player was 'in' form) and sequences of low scores (indicating the player was 'out of' form). On the contrary, Durbach and Thiart [9] found little empirical evidence to support this idea. Instead, for the majority of players analyzed in the study, the authors suggest that public perceptions of batting form tend to be overestimated, with many players' scores able to be modelled using a random sequence.

Within a Bayesian framework, Koulis et al. [13] employed the use of a hidden Markov model to determine whether a batsman is in or out of form. The model estimates a number, K, of 'underlying batting states' for each player, including the expected number of runs to be scored when in each of the K states. Parameters that measure: *availability* (the probability a batsman is in form for a given match), *reliability* (the probability a batsman is in form for the next n matches) and *mean time to failure* (the expected number of innings a batsman will play before he is out of form), were also estimated for each batsman. However, a drawback of this approach is that the model requires an explicit specification of what constitutes an out of form state. The authors specify a batting state that has a posterior expected median number of runs scored of less than 25, as being out of form, and all other states as being in form. While in the context of one day or Twenty20 cricket this is not necessarily an unreasonable specification, there are numerous arguments that could be made to justify a low score, scored at a high strike rate, as a successful innings.

In this paper, we extend the Bayesian parametric model detailed in [16], such that we can not only measure and predict how player batting abilities fluctuate during an innings, but also between innings, over the course of entire playing careers. This allows us to treat batting form as continuous, rather than binary; instead of defining players as 'in' or 'out' of form, we can describe players as improving or deteriorating in terms of batting ability. At this stage our focus is on longer form test and first-class cricket, as limited overs cricket introduces a number of match-specific complications [7].

2 Model Specification

The derivation of the model likelihood follows the method detailed in [16]. If $X \in \{0, 1, 2, 3, \ldots\}$ is the number of runs a batsman is currently on, we define a *hazard function*, $H(x) \in [0, 1]$, as the probability a batsman gets out on score x. Assuming a functional form for $H(x)$, conditional on some parameters θ, we can calculate the probability distribution for X as follows:

$$P(X = x) = H(x) \prod_{a=0}^{x-1} [1 - H(a)] . \tag{1}$$

For any given value of x, this can be thought of as the probability of a batsman surviving up until score x, then being dismissed. However, in cricket there are a number of instances where a batsman's innings may end without being dismissed (referred to as a 'not out' score). Therefore, in the case of not out scores, we compute $P(X \geq x)$ as the likelihood, rather than $P(X = x)$. Comparable to right-censored observations in the context of survival analysis, this assumes that for not out scores the batsman would have gone on to score some unobserved score, conditional on their current score and their assumed hazard function.

Therefore, if I is the total number of innings a player has batted in and N is the number of not out scores, the probability distribution for a set of conditionally independent 'out' scores $\{x_i\}_{i=1}^{I-N}$ and 'not out' scores $\{y_i\}_{i=1}^{N}$ can be expressed as

$$p(\{x\}, \{y\}) = \prod_{i=1}^{I-N} \left(H(x_i) \prod_{a=0}^{x_i - 1} [1 - H(a)] \right) \times \prod_{i=1}^{N} \left(\prod_{a=0}^{y_i - 1} [1 - H(a)] \right). \tag{2}$$

When data $\{x, y\}$ are fixed and known, Eq. (2) gives the likelihood for any proposed form of the hazard function, $H(x; \theta)$. Therefore, conditional on the set of parameters θ governing the form of $H(x)$, the log-likelihood is

$$\log\Big(L(\theta)\Big) = \sum_{i=1}^{I-N} \log H(x_i) + \sum_{i=1}^{I-N} \sum_{a=0}^{x_i - 1} \log[1 - H(a)] + \sum_{i=1}^{N} \sum_{a=0}^{y_i - 1} \log[1 - H(a)]. \tag{3}$$

2.1 Parameterizing the Hazard Function

The model likelihood in Eq. (3) depends on the parameterization of the hazard function, $H(x)$. As per [16], we parameterize the hazard function in terms of an *effective average function*, $\mu(x)$, which represents a player's ability on score x, in terms of a batting average. Given the prevalence of the batting average in cricket, it is far more intuitive for players and coaches to think of ability in terms of batting averages, rather than dismissal probabilities. The hazard function can then be expressed in terms of the effective average function, $\mu(x)$, as follows

$$H(x) = \frac{1}{\mu(x) + 1} \tag{4}$$

where the effective average contains three parameters, $\theta = \{\mu_1, \mu_2, L\}$, and takes the following functional form

$$\mu(x) = \mu_2 + (\mu_1 - \mu_2) \exp\left(\frac{-x}{L}\right). \tag{5}$$

Here, μ_1 represents a player's initial batting ability when beginning a new innings, while μ_2 is the player's 'eye in' batting ability once used to the specific match conditions. Both μ_1 and μ_2 are expressed in terms of a batting average. The timescale parameter L, measures the speed of transition between μ_1 and μ_2 and is formally the e-folding time. By definition the e-folding time, L, signifies the number of runs scored for approximately 63% (formally $1 - \frac{1}{e}$) of the transition between μ_1 and μ_2 to take place and can be understood by analogy with a 'half-life'. This model specification allows us to answer questions about individual players, such as: (1) how well players perform when they first arrive at the crease, (2) how much better players perform once they have 'got their eye in' and (3) how long it takes them to 'get their eye in'.

2.2 *Modelling Between-Innings Changes in Batting Ability*

To extend the model further, such that we can measure variations in player batting ability between innings, we use the same likelihood function in Eq. (3). However, we re-parameterize the effective average function to include a time component, t, such that

$$\mu(x, t) = \text{expected batting average on score } x \text{, in } t\text{th career innings.} \tag{6}$$

For clarity, we will refer to $\mu(x)$ as the 'within-innings' effective average (explaining how ability changes within an innings). By marginalizing over all scores, x, we obtain the 'between-innings' effective average, $\nu(t)$, which explains how ability changes between innings, across a playing career.

$$\nu(t) = \text{expected batting average in } t\text{th career innings.} \tag{7}$$

When estimating $\nu(t)$, we need to account for variations in ability due to external factors such as: recent form, general improvements/deterioration in skill and the element of randomness associated with cricket. This is achieved by fitting a μ_2 parameter for each innings in a player's career, where $\mu_{2,t}$ represents a player's 'eye in' batting ability, corresponding to their tth innings. We are then able to predict the expected batting average in each innings, $\nu(t)$, analytically using Eq. (5).

To afford a player's underlying batting ability a reasonable amount of flexibility, the set of $\{\mu_{2,t}\}$ terms are modelled using a Gaussian process. A Gaussian process is fully specified by an underlying mean value, m, and covariance function, $K(t_i, t_j)$, which will determine by how much a player's batting ability can vary from innings to innings [14]. Our choice of covariance function is the commonly used squared exponential covariance, which contains scale and length parameters σ and ℓ.

Therefore, the model contains the set of parameters $\theta = \{\mu_1, \{\mu_{2,t}\}, L, m, \sigma, \ell\}$. The model structure with respect to parameters μ_1, L, C and D follows the model specification detailed in [16], with the parameters assigned the following prior distributions.

$$\begin{aligned} \mu_1 &\leftarrow C\mu_2 & \log(\mu_{2,t}) &\sim \text{Gaussian process}(m, K(t_i, t_j; \sigma, \ell)) \\ L &\leftarrow D\mu_2 & m &\sim \text{Lognormal}(\log(25), 0.75^2) \\ C &\sim \text{Beta}(1, 2) & \sigma &\sim \text{Exponential}(10) \\ D &\sim \text{Beta}(1, 5) & \ell &\sim \text{Uniform}(0, 100) \end{aligned}$$

These priors are either non-informative or are relatively conservative, while loosely reflecting our cricketing knowledge. It is worth noting, that as we are measuring ability in terms of a batting average (which must be positive), we model $\log(\mu_{2,t})$, rather than just $\mu_{2,t}$, to ensure positivity in our estimates.

As the model requires a set of $\{\mu_{2,t}\}$ parameters to be fitted (one for each innings played), the model can contain a large number of parameters for players who have enjoyed long international careers. Therefore, to fit the model we employ a C++ implementation of the nested sampling algorithm [15], which uses Metropolis-Hastings updates and is able to handle both high dimensional and multimodal problems that may arise. The model output provides us with the posterior distribution for each of the model parameters, as well as the marginal likelihood, which makes for trivial model comparison. For each player analyzed, we initialize the algorithm with 1000 particles and use 1000 MCMC steps per nested sampling iteration.

3 Analysis of Individual Players

3.1 Data

The data we use to fit the model are simply the Test career scores of an individual batsman and are obtained from Statsguru, the cricket statistics database on the Cricinfo website.[1] As the model assumes that a player's ability is not influenced by the specific match scenario, it is best suited to longer form cricket, such as Test matches, where there is generally minimal external pressure on batsmen to score runs at a prescribed rate.

3.2 Modelling Between-Innings Changes in Batting Ability

To illustrate the practical implications of the model, let us consider the Test match batting career of current New Zealand captain, Kane Williamson. The evolution of Williamson's between-innings effective average, $\nu(t)$, is shown in Fig. 1 and suggests that early in his career, Williamson was not as good a batsman as he is today. In fact, it was not until playing in roughly 50 innings that he began to consistently bat *at least* as well as his current career average of 50.36. This is not surprising, as it is

[1] www.espncricinfo.com.

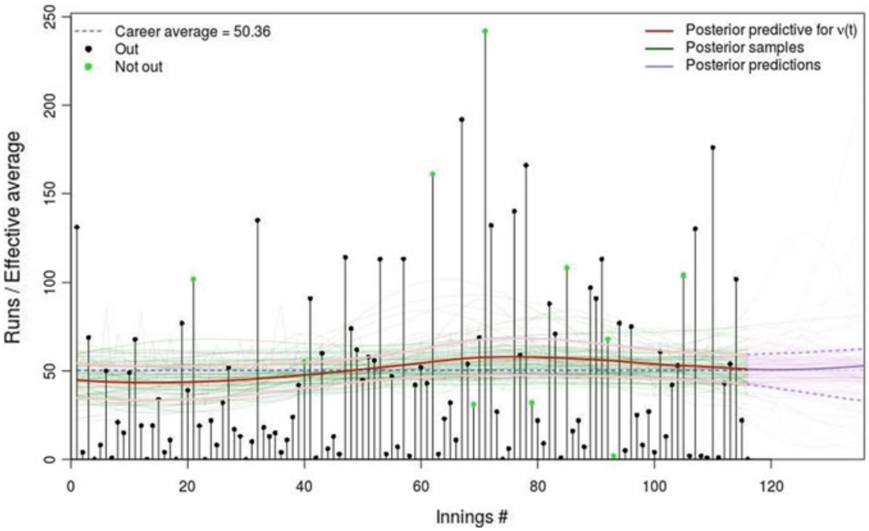


Fig. 1 Posterior predictive effective average function for $\nu(t)$ (red), fitted to Kane Williamson's test career data, including a subset of posterior samples (green), future predictions (purple) and a 68% credible interval (pink/dotted purple)

a commonly held belief that many players need to play in a number of matches to 'find their feet' at the international level, before reaching their peak ability.

To gain a better understanding of how Williamson compares to other batsmen globally, we can compare multiple players' effective average functions. Figure 2 compares the predictive effective average functions for the current top four batsmen in the world, as ranked by the official International Cricket Council (ICC) ratings.[2] As we might expect, all players appear to have improved in terms of batting ability since the start of their careers. Again, this supports the concept of 'finding your feet' at the international level, although different players appear to take different lengths of time to adjust to the demands of international cricket.

Table 1 shows each player's predicted effective average for their next innings, as well as their ICC rating. The order of these four players remains unchanged when ranking by predicted effective averages instead of ICC ratings, however, as we have computed the posterior predictive distributions for $\nu(t)$, our model has the added advantage of being able to quantify the differences in abilities between players. Rather than concluding 'Steve Smith is 26 rating points higher than Virat Kohli', we can make more useful statements such as: 'we expect Steve Smith to outscore Virat Kohli by 5.1 runs in their next respective innings' and 'Steve Smith has a 68.8% chance of outscoring Virat Kohli in their next respective innings'.

[2] As of 1st August, 2018: (1) Steve Smith, (2) Virat Kohli, (3) Joe Root and (4) Kane Williamson—commonly referred to as 'the big four'.

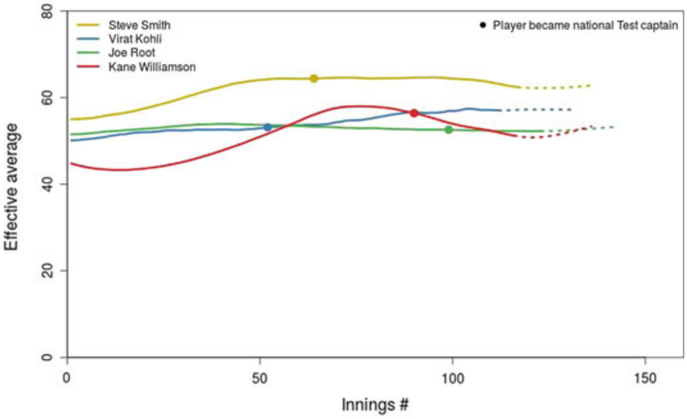


Fig. 2 Posterior predictive effective average functions, $\nu(t)$, for 'the big four', including predictions for the next 20 innings (dotted)

Table 1 Predicted effective averages, $\nu(t)$, for the next career innings for 'the big four'. The official ICC Test batting ratings (as of 1st August, 2018) are shown for comparison

Player	Career average	Predicted ν(next innings)	ICC rating (#)
S. Smith (AUS)	61.4	62.5	929 (1)
V. Kohli (IND)	53.4	57.4	903 (2)
J. Root (ENG)	52.6	52.6	855 (3)
K. Williamson (NZ)	50.4	51.2	847 (4)

4 Concluding Remarks and Future Work

We have presented a novel and more accurate method of quantifying player batting ability than traditional cricketing statistics, such as the batting average. The results provide support for the common cricketing belief of 'finding your feet', particularly for players beginning their international careers at a young age, with many batsmen taking a number of innings to reach their peak ability in the Test match arena. With respect to batting form, the model appears to reject the idea of recent performances having a significant impact on innings in the near future. In particular, it appears that the effect of recent form varies greatly from player to player.

A major advantage of the model is that we are able to maintain an intuitive cricketing interpretation, allowing for the results and implications to be easily digested by coaches and selectors, who may have minimal statistical training. Additionally, we are able to make probabilistic statements and comparisons between players, allowing us to easily quantify differences in abilities and predict the real life impacts of selecting one player over another. As such, the findings have practical implications in terms of player comparison, talent identification, and team selection policy.

It is worth noting that we have ignored important variables, such as the number of balls faced in each innings, as well as the strength of the opposition. Currently, the model treats all runs scored equally. Implementing a means of incorporating more in-depth, ball-by-ball data and including the strength of opposition bowlers will reward players who consistently score highly against world-class bowling attacks.

References

1. Bailey, M., Clarke, S.R.: Predicting the match outcome in one day international cricket matches, while the game is in progress. J. Sports Sci. Med. **5**(4), 480 (2006)
2. Bracewell, P.J., Ruggiero, K.: A parametric control chart for monitoring individual batting performances in cricket. J. Quant. Anal. Sports **5**(3) (2009)
3. Brooker, S., Hogan, S.: A Method for Inferring Batting Conditions in ODI Cricket from Historical Data (2011)
4. Brooks, R.D., Faff, R.W., Sokulsky, D.: An ordered response model of test cricket performance. Appl. Econ. **34**(18), 2353–2365 (2002)
5. Cai, T., Hyndman, R.J., Wand, M.: Mixed model-based hazard estimation. J. Comput. Graph. Stat. **11**(4), 784–798 (2002)
6. Carter, M., Guthrie, G.: Cricket interruptus: fairness and incentive in limited overs cricket matches. J. Oper. Res. Soc. **55**(8), 822–829 (2004)
7. Davis, J., Perera, H., Swartz, T.B.: A simulator for Twenty20 cricket. Aust. N. Z. J. Stat. **57**(1), 55–71 (2015)
8. Duckworth, F.C., Lewis, A.J.: A fair method for resetting the target in interrupted one-day cricket matches. J. Oper. Res. Soc. **49**(3), 220–227 (1998)
9. Durbach, I.N., Thiart, J.: On a common perception of a random sequence in cricket: application. S. Afr. Stat. J. **41**(2), 161–187 (2007)
10. Elderton, W., Wood, G.H.: Cricket scores and geometrical progression. J. Roy. Stat. Soc. **108**(1/2), 12–40 (1945)
11. Ian, P., Thomas, J.: Rain rules for limited overs cricket and probabilities of victory. J. Roy. Stat. Soc.: Ser. D: Stat. **51**(2), 189–202 (2002)
12. Kimber, A.C., Hansford, A.R.: A statistical analysis of batting in cricket. J. Roy. Stat. Soc. Ser. A: Stat. Soc., 443–455 (1993)
13. Koulis, T., Muthukumarana, S., Briercliffe, C.D.: A Bayesian stochastic model for batting performance evaluation in one-day cricket. J. Quant. Anal. Sports **10**(1), 1–13 (2014)
14. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. MIT Press (2006)
15. Skilling, J.: Nested sampling for general Bayesian computation. Bayesian Anal. **1**(4), 833–859 (2006)
16. Stevenson, O.G., Brewer, B.J.: Bayesian survival analysis of batsmen in test cricket. J. Quant. Anal. Sports **13**(1), 25–36 (2017)
17. Swartz, T.B., Gill, P.S., Muthukumarana, S.: Modelling and simulation for one-day cricket. Can. J. Stat. **37**(2), 143–160 (2009)

Ice Cores and Emulation: Learning More About Past Ice Sheet Shapes

Fiona Turner, Richard Wilkinson, Caitlin Buck, Julie Jones and Louise Sime

Abstract Creating more accurate reconstructions of past Antarctic ice sheet shapes allows us to better predict how they will vary in the changing climate and contribute to future sea level changes. In this research, we use expert elicitation to create a subjective prior distribution of the Antarctic ice sheets at the Last Glacial Maximum (LGM), 21Ka. A design of shapes from this distribution will be run through the global climate model HadCM3, providing us with output that we can compare with proxy data to find a better estimate of the ice sheet shape at the LGM.

Keywords Antarctic ice sheets · Expert elicitation · Principal component analysis · Subjective Bayesian methods

1 Background

1.1 *Environmental Research*

Extensive research has been done on reconstructing the ice sheets at both poles. This is done using mathematical computer models of climate combined with data and various physical constraints, providing us with estimates of the shape and size of the

F. Turner (✉) · R. Wilkinson · C. Buck
School of Mathematics and Statistics, University of Sheffield, Sheffield, UK
e-mail: feturner1@sheffield.ac.uk

R. Wilkinson
e-mail: r.d.wilkinson@sheffield.ac.uk

C. Buck
e-mail: c.e.buck@sheffield.ac.uk

J. Jones
Department of Geography, University of Sheffield, Sheffield, UK
e-mail: julie.jones@sheffield.ac.uk

L. Sime
British Antarctic Survey, Cambridge, UK
e-mail: lsim@bas.ac.uk

R. Argiento et al. (eds.), *Bayesian Statistics and New Generations*,
Springer Proceedings in Mathematics & Statistics 296,
https://doi.org/10.1007/978-3-030-30611-3_18

ice sheets. The focus is often on critical time periods, when the ice sheets were most in flux and there is greatest variance in estimates.

Data derived from indirect observations are known as proxy data. These come from such sources as coral reefs, tree rings and ice cores. They act as observations for time periods or locations when direct observations were not made, making them an essential tool in palaeo-climatology. In this research, we use ice cores drilled from the Antarctic ice sheets to learn more about the past size of the sheets. These ice cores contain the stable water isotopes deuterium and oxygen-18; the ratio of these to oxygen-16 can tell us the precipitation and temperature when the ice formed, amongst other things. Many of the ice cores contain hundreds of thousands of years worth of data, making them an invaluable set of observations in an area that was unstudied by humans until the twentieth century.

The Last Glacial Maximum (LGM) is a time period of significant interest to palaeo-climatologists. Taking place at 21Ka, it was the peak of the most recent ice age, when the ice sheets were at their greatest extent. Studying the transition of the ice sheets to present day can teach us a lot about how they respond as the climate warms. The relative recentness of this event also means there are many ice cores dating back to the LGM, providing us with a set of proxy observations to use in our analysis.

Sea levels were 125 m lower than present day during the LGM and have the potential to rise further still if both the Antarctic ice sheets melt completely. If we can estimate how the shape and size of the ice sheets vary in the changing climate, we can better predict how they will contribute to sea level changes.

1.2 Why Use Bayesian Methods?

Prior distributions created in Bayesian analysis represent the statisticians' own judgements. Eliciting expert opinions about the parameters we are trying to describe could allow us to reduce uncertainty around them. Expert elicitation is an important stage in creating a subjective prior distribution. Particularly when there is little information in the data being used, a prior distribution can be very influential on results.

The Antarctic ice sheets have been reconstructed at the LGM multiple times. We have collected forty of these shapes from the literature and will use them in our own analysis of the shape and size of the sheets. Disregarding the shapes when creating this reconstruction would result in a less informed model; we therefore will use them to form a subjective prior distribution. By incorporating the shapes in to our model, our prior distribution will better represent the uncertainty around the ice sheets' shapes at the LGM.

2 Building the Prior Model

We wish to create a prior model using the collected reconstructions. However, we first need to reduce the dimensions of our dataset; having forty variables would make running the model too computationally expensive. We therefore create a basis: a set of direction vectors that span the variation between and within the collected ice sheet shapes. These vectors will be used to create new synthetic ice sheet shapes based on the existing reconstructions.

The first four reconstructions we collected ([1, 3, 9, 11]) are the most widely used and accepted in the climatology community. We wish for these to influence our prior model more than the other thirty-six ice sheet shapes, so we split the collected ice sheet shapes into two sets and first create basis vectors from these four shapes using Principal Component Analysis (PCA). PCA reduces the dimensions of a data set whilst preserving as much of the information in the data as possible. It creates a set of orthogonal vectors by maximising

$$\boldsymbol{a}^T \Sigma \boldsymbol{a}, \tag{1}$$

where $\boldsymbol{a}$ is the vector we are trying to find and Σ is the covariance matrix of the data. These vectors are then ordered by how much of the variation in the data they represent. A cut off point for these vectors is decided on based on how much variation each vector represents. The vectors containing less than this cut off are then discarded. Performing PCA on the forty shapes altogether would result in smaller reconstruction errors overall. However, we want to ensure that the initial four shapes have reconstruction errors that are as small as possible, and that they influence the prior model more than the other thirty-six shapes. Splitting the shapes in to two sets allows us to create basis vectors from the first four ice sheet shapes and then find any variation in the second set that hasn't already been described. Performing PCA on the first four shapes gives us four principal components representing 50.45%, 34.48%, 9.23% and 5.84% of the variation respectively. We discard the fourth component as it represents only a small amount of the variation in the data; this leaves us with three basis vectors. When we reconstruct all forty shapes using these principal components, we find they represent 75.77% of all variation. We now have an initial set of basis vectors, which we collectively call A.

Having worked first with these four widely used reconstructions, a further literature review revealed thirty-six reconstructions that are also worthy of inclusion. These come from [2, 4–6, 10]. We use these thirty-six other shapes to create further basis vectors. These must be orthogonal to A, so that they do not describe any directions in the basis that A already represents. We find the new vectors by using the null space of A; the set of vectors for which

$$A\boldsymbol{x} = 0.$$

We call this set B, and wish to find vectors of the form $\boldsymbol{w}B$ for some $\boldsymbol{w}$.

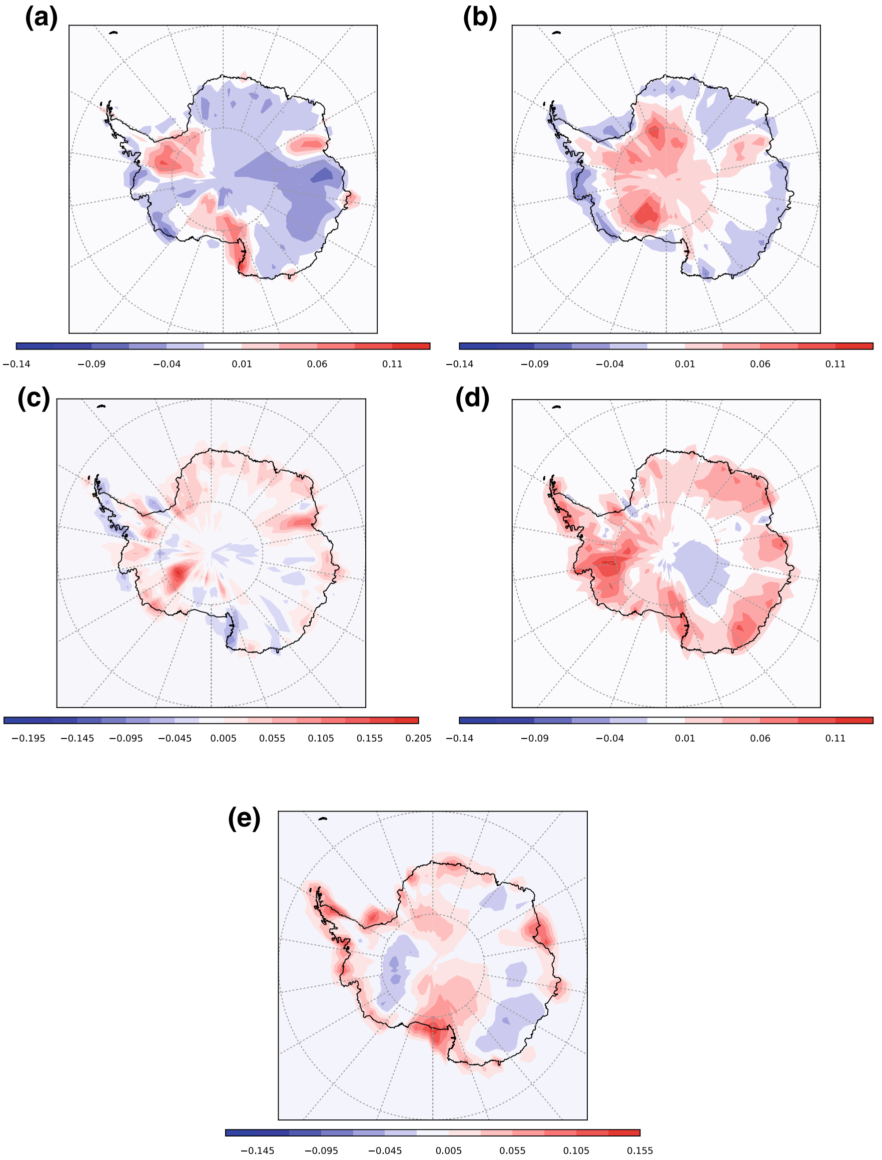


Fig. 1 Our five basis vectors. The black line represents present day Antarctica for reference

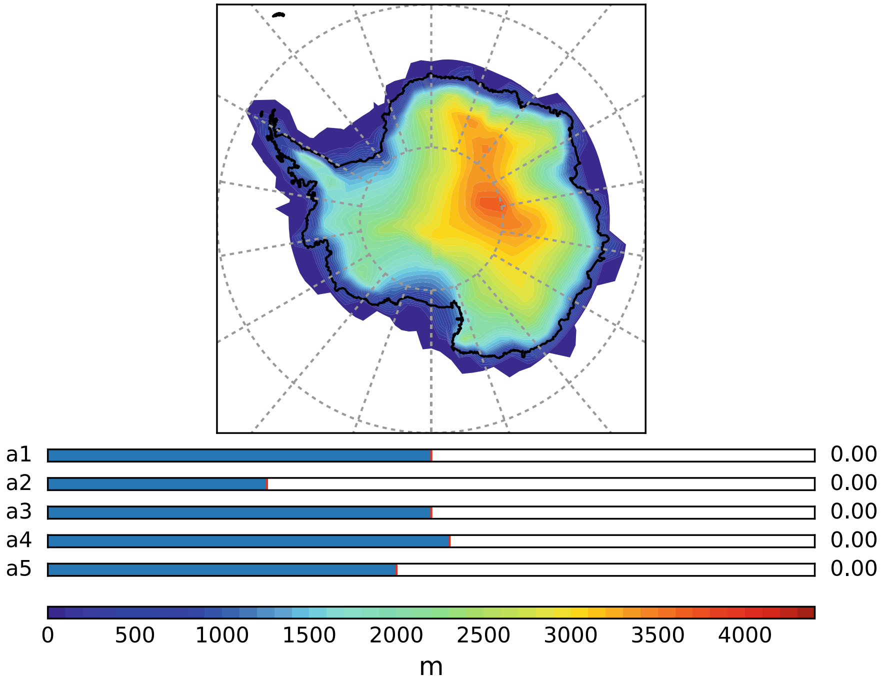


Fig. 2 An interactive plot of our prior model; the sliders below the plot control the five variables, allowing us to create new ice sheet shapes by adjusting the individual basis vectors. Currently the sliders are all set to zero; the above plot is of the mean ice sheet shape

In PCA we maximised $A^T \Sigma A$. Here, we are maximising $\boldsymbol{w}^T B^T \Sigma B \boldsymbol{w}$. As

$$\Sigma = X^T X,$$

where X is our data set of thirty-six shapes, and

$$B^T \Sigma B = B^T X^T X B = (X B)^T X B,$$

this is an eigenvalue problem that we can solve with the singular value decomposition, which factorises a $n \times p$ matrix in to the form $U \Sigma V^T$, where U is a $n \times n$ matrix, Σ is a diagonal $n \times n$ matrix of eigenvalues and V is a $p \times n$ matrix of corresponding eigenvectors. We take the first two column vectors of V as our new basis vectors. These represent 18.93% of variation in the forty shapes; our basis vectors now account for 94.7% of the variation in the dataset. This method essentially copies PCA for two sets of data. Other methods we had tried, including Newton-Raphson, proved too computationally expensive as our data set is too large to invert. We now have a model with five variables with which we can create synthetic ice sheet shapes. These five variables can be seen in Fig. 1. This takes the form

$$y = \alpha_1 x_1 + ... + \alpha_5 x_5 + \mu,$$

where the x_i are the basis vectors, μ is the mean shape that was removed from the data before analysis and the α_i are a set of constraints on the variables that allow us to control how much of each variable contributes to the model output. An interactive plot of this model can be seen in Fig. 2. The five sliders control the α_i; moving them left or right determines how much each of the five variables contributes to the ice sheet plotted above.

3 Expert Elicitation

We now wish to run a set of synthetic ice sheet shapes from our prior model through the global climate model, HadCM3. This will provide us with estimates of water isotope values that can be compared to data collected from ice cores. HadCM3 is both computationally and financially very expensive to run. We are therefore limited in how many jobs we can submit and any shapes we run through the model must be plausible for the LGM. We decide to gather expert judgements on our prior model to assist us with determining the direction of our future work; only shapes that the experts judge to have a high probability of occurring at the LGM will be run through HadCM3.

There is extensive literature on how to conduct a formal expert elicitation, such as the Sheffield Elicitation framework [7]; this a package on how to elicit probability distributions from multiple experts. We initially consulted [8] when planning our elicitation process. However, although we considered conducting a formal elicitation, it became clear we required something a lot simpler. Rather than a formal systematic approach, we organised a meeting with various experts at the British Antarctic Survey.

We shared our model with a group of five ice modellers and LGM experts at the British Antarctic Survey. These were Robert Arthern,[1] Richard Hindmarsh,[2] Dominic Hodgson,[3] Robert Mulvaney[4] and James Smith.[5] We held an informal meeting to explain the project and ask for their advice on it. From this meeting we gained constraints on the values the α_i's could take. We then created a design of shapes by building a maxi-min Latin Hypercube design for the α_i's. This gave us a set of eighty shapes that spanned the plausible basis we have created. We shared this with the experts as a follow-up to our meeting and asked them to recommend which of the shapes to discard and which to run through HadCM3, providing reasons for their recommendations. Using their judgements, we decided on a set of forty-nine synthetic ice sheet shapes that will be used for our analysis. Examples of the shapes

[1] Ice sheet modeller, rart@bas.ac.uk.

[2] Glaciologist, rcah@bas.ac.uk.

[3] Sedimentologist, daho@bas.ac.uk.

[4] Science leader of the Ice Dynamics and Palaeoclimate team, rmu@bas.ac.uk.

[5] Sedimentologist, jaas@bas.ac.uk.

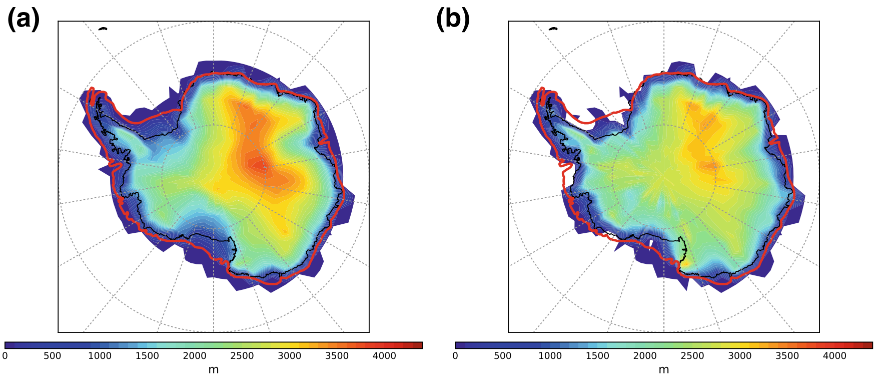


Fig. 3 Two of the shapes created and shown to the experts for their judgements; the shape on the left was kept whilst the shape on the right was discarded. The red line was added on their advice; it is the ice extent limit from [2], and was used a constraint on the shapes

they recommended to be kept or discarded are plotted in Fig. 3. The red line plotted over the ice sheet is the ice extent limit from [2]; they recommended this as a measure on the plausibility of each ice sheet.

4 Future Work

We will run this chosen set of synthetic ice sheet shapes through HadCM3. This will give us estimated values of water isotopes over the time period run. We will then build a Gaussian process emulator to interpolate between these values; the output of the emulator will be modelled as a likelihood, and compared to collected ice cores to see which input from the prior model provides isotope estimates that most closely matches these observations. We should then be able to determine on a more accurate estimate of the size and shape of the Antarctic ice sheets at the LGM.

References

1. Argus, D.F., Peltier, W.R., Drummond, R., Moore, A.W.: The Antarctica component of postglacial rebound model ICE-6G_C (VM5a) based on GPS positioning, exposure age dating of ice thicknesses, and relative sea level histories. Geophys. J. Int. **198**, 537–563 (2014)
2. Bentley, M.J., Cofaigh, C.A., Anderson, J.B., Conway, H., Davies, B., Graham, A.G.C., Hillenbrand, C.-D., Hodgson, D.A., Jamieson, S.S.R., Larter, R.D., Mackintosh, A., Smith, J.A., Verleyen, E., Ackert, R.P., Bart, P.J., Berg, S., Brunstein, D., Canals, M., Colhoun, E.A., Crosta, X., Dickens, W.A., Domack, E., Dowdeswell, J.A., Dunbar, R., Ehrmann, W., Evans, J., Favier, V., Fink, D., Fogwill, C.J., Glasser, N.F., Gohl, K., Golledge, N.R., Goodwin, I., Gore, D.B., Greenwood, S.L., Hall, B.L., Hall, K., Hedding, D.W., Hein, A.S., Hocking, E.P., Jakobsson,

M., Johnson, J.S., Jomelli, V., Jones, R.S., Klages, J.P., Kristoffersen, Y., Kuhn, G., Leventer, A., Licht, K., Lilly, K., Lindow, J., Livingstone, S.J., Masse, G., McGlone, M.S., McKay, R.M., Melles, M., Miura, H., Mulvaney, R., Nel, W., Nitsche, F.O., O'Brien, P.E., Post, A.L., Roberts, S.J., Saunders, K.M., Selkirk, P.M., Simms, A.R., Spiegel, C., Stolldorf, T.D., Sugden, D.E., van der Putten, N., van Ommen, T., Verfaillie, D., Vyverman, W., Wagner, B., White, D.A., Witus, A.E., Zwartz, D.: A community-based geological reconstruction of Antarctic ice sheet deglaciation since the Last Glacial Maximum. Quat. Sci. Rev. **100**, 1–9 (2014)

3. Briggs, R.D., Pollard, D., Tarasov, L.: A data-constrained large ensemble analysis of Antarctic evolution since the Eemian. Quat. Sci. Rev. **103**, 91–115 (2014)
4. de Boer, B., Haywood, A.M., Dolan, A.M., Hunter, S.J., Prescott, C.L.: The transient response of ice volume to orbital forcing during the warm late Pliocene. Geophys. Res. Lett. **44**, 10486–10494 (2017)
5. Golledge, N.R., Fogwill, C.J., Mackintosh, A.N., Buckley, K.M.: Dynamics of the last glacial maximum Antarctic ice-sheet and its response to ocean forcing. Proc. Natl. Acad. Sci. **109**, 16052–16056 (2012)
6. Golledge, N.R., Levy, R.H., McKay, R.M., Fogwill, C.J., White, D.A., Graham, A.G.C., Smith, J.A., Hillenbrand, C.-D., Licht, K.J., Denton, G.H., Ackert, R.P., Maas, S.M., Hall, B.L.: Glaciology and geological signature of the Last Glacial Maximum Antarctic ice sheet. Quat. Sci. Rev. **78**, 225–247 (2013)
7. Oakley, J.E., O'Hagan, A.: SHELF: the Sheffield elicitation framework (version 3.0). School of Mathematics and Statistics, University of Sheffield, UK (2016). (http://tonyohagan.co.uk/shelf)
8. 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, Hoboken (2006)
9. Peltier, W.R.: Global glacial isostasy and the surface of the ice-age earth: the ICE-5G (VM2) model and GRACE. Annu. Rev. Earth Planet Sci. **32**, 111–149 (2004)
10. Pollard, D., DeConto, R.M.: Modelling West Antarctic ice sheet growth and collapse through the past five million years. Nature **458**, 329–333 (2009)
11. Whitehouse, P.L., Bentley, M.J., Milne, G.A., King, M.A., Thomas, I.D.: A new glacial isostatic adjustment model for Antarctica: calibrated and tested using observations of relative sea-level change and present-day uplift rates. Geophys. J. Int. **190**, 1464–1482 (2012)

Author Index

R. Argiento et al. (eds.), *Bayesian Statistics and New Generations*,
Springer Proceedings in Mathematics & Statistics 296,
https://doi.org/10.1007/978-3-030-30611-3