Topics in computational statistics

Yu Yang

Supervised by:

Prof. Robert Kohn, Prof. Scott Sisson, Dr. Matias Quiroz

A thesis in fulfilment of the requirements for the degree of

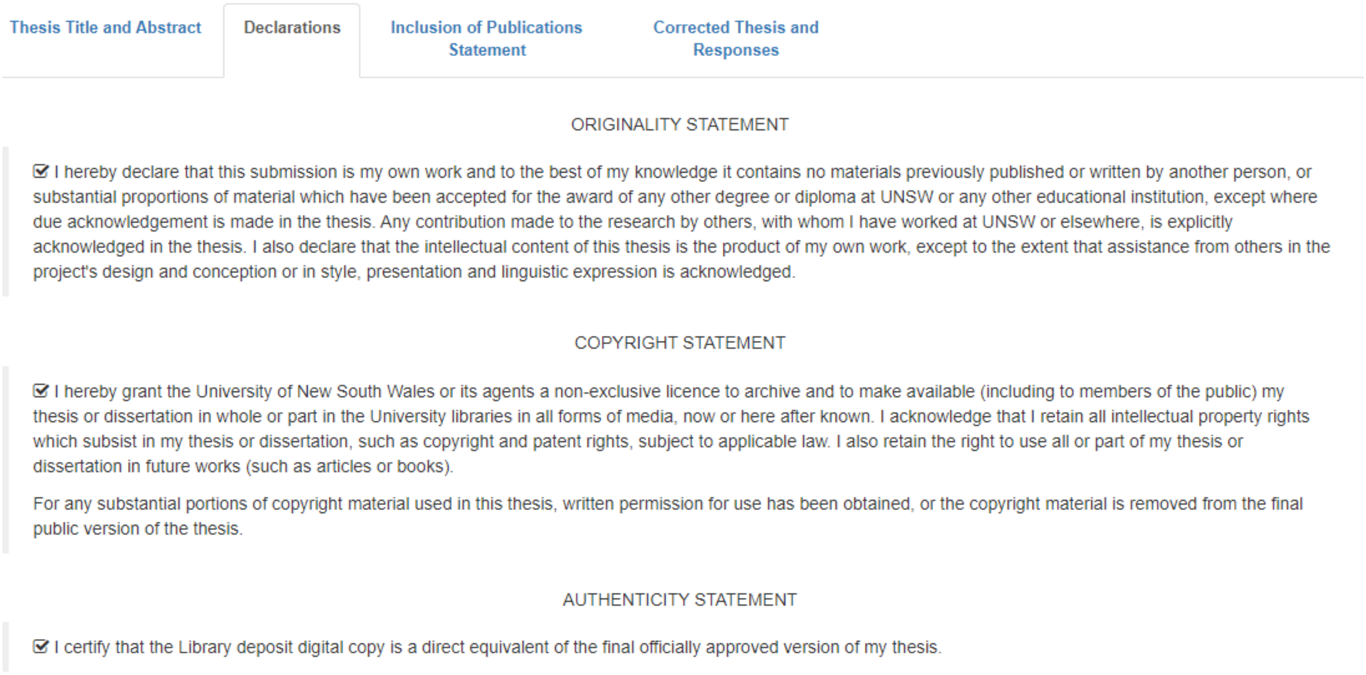
Doctor of Philosophy

School of Economics

UNSW Business School

The University of New South Wales

Mar 2022


Thesis Title and Abstract
Declarations
Inclusion of Publications Statement
Corrected Thesis and Responses

ORIGINALITY STATEMENT

☑ I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.

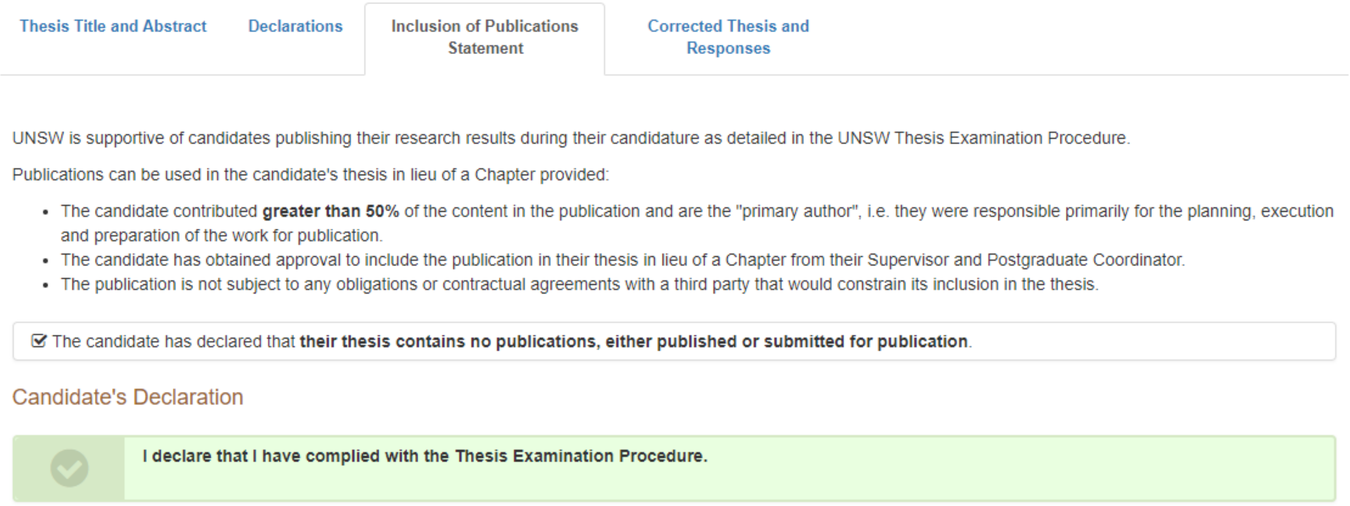
COPYRIGHT STATEMENT

☑ I hereby grant the University of New South Wales or its agents a non-exclusive licence to archive and to make available (including to members of the public) my thesis or dissertation in whole or part in the University libraries in all forms of media, now or here after known. I acknowledge that I retain all intellectual property rights which subsist in my thesis or dissertation, such as copyright and patent rights, subject to applicable law. I also retain the right to use all or part of my thesis or dissertation in future works (such as articles or books).

For any substantial portions of copyright material used in this thesis, written permission for use has been obtained, or the copyright material is removed from the final public version of the thesis.

AUTHENTICITY STATEMENT

☑ I certify that the Library deposit digital copy is a direct equivalent of the final officially approved version of my thesis.


Thesis Title and Abstract
Declarations
Inclusion of Publications Statement
Corrected Thesis and Responses

UNSW is supportive of candidates publishing their research results during their candidature as detailed in the UNSW Thesis Examination Procedure.

Publications can be used in the candidate's thesis in lieu of a Chapter provided:

- The candidate contributed **greater than 50%** of the content in the publication and are the "primary author", i.e. they were responsible primarily for the planning, execution and preparation of the work for publication.
- The candidate has obtained approval to include the publication in their thesis in lieu of a Chapter from their Supervisor and Postgraduate Coordinator.
- The publication is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in the thesis.

☑ The candidate has declared that **their thesis contains no publications, either published or submitted for publication**.

Candidate's Declaration

✓ **I declare that I have complied with the Thesis Examination Procedure.**

THE UNIVERSITY OF NEW SOUTH WALES
Thesis/Dissertation Sheet

Surname or Family name: **Yang**

First name: **Yu**

Abbreviation for degree as given in the University calendar: **PhD**

Faculty: **UNSW Business School**

School: **School of Economics**

Thesis Title: **Topics in computational statistics**

Abstract

Research in computational statistics develops numerically efficient methods to estimate statistical models, with Monte Carlo algorithms a subset of such methods. This thesis develops novel Monte Carlo methods to solve three important problems in Bayesian statistics. For many complex models, it is prohibitively expensive to run simulation methods such as Markov chain Monte Carlo (MCMC) on the model directly when the likelihood function includes an intractable term or is computationally challenging in some other way. The first two topics investigate models having such likelihoods. The third topic proposes a novel model to solve a popular question in causal inference, which requires solving a computationally challenging problem.

The first application is to symbolic data analysis, where classical data are summarised and represented as symbolic objects. The likelihood function of such aggregated-level data is often intractable as it usually includes a high dimensional integral with large exponents. Bayesian inference on symbolic data is carried out in the thesis by using a pseudo-marginal method, which replaces the likelihood function with its unbiased estimate.

The second application is to doubly intractable models, where the likelihood includes an intractable normalising constant. The pseudo-marginal method is combined with the introduction of an auxiliary variable to obtain simulation consistent inference. The proposed algorithm offers a generic solution to a wider range of problems, where the existing methods are often impractical as the assumptions required for their application do not hold.

The last application is to causal inference using Bayesian additive regression trees (BART), a non-parametric Bayesian regression technique. The likelihood function is complex as it is based on a sum of trees whose structures change dynamically with the MCMC iterates. An extension to BART is developed to estimate the heterogeneous treatment effect, aiming to overcome the regularisation-induced confounding issue which is often observed in the direct application of BART in causal inference.

Declaration relating to disposition of project thesis/dissertation

Signature **Yu Yang** Witness Date **Mar, 2022**

FOR OFFICE USE ONLY Date of completion of requirements for Award

Abstract

Research in computational statistics develops numerically efficient methods to estimate statistical models, with Monte Carlo algorithms a subset of such methods. This thesis develops novel Monte Carlo methods to solve three important problems in Bayesian statistics. For many complex models, it is prohibitively expensive to run simulation methods such as Markov chain Monte Carlo (MCMC) on the model directly when the likelihood function includes an intractable term or is computationally challenging in some other way. The first two topics investigate models having such likelihoods. The third topic proposes a novel model to solve a popular question in causal inference, which requires solving a computationally challenging problem.

The first application is to symbolic data analysis, where classical data are summarised and represented as symbolic objects. The likelihood function of such aggregated-level data is often intractable as it usually includes a high dimensional integral with large exponents. Bayesian inference on symbolic data is carried out in the thesis by using a pseudo-marginal method, which replaces the likelihood function with its unbiased estimate.

The second application is to doubly intractable models, where the likelihood includes an intractable normalising constant. The pseudo-marginal method is combined with the introduction of an auxiliary variable to obtain simulation consistent inference. The proposed algorithm offers a generic solution to a wider range of problems, where the existing methods are often impractical as the assumptions required for their application do not hold.

The last application is to causal inference using Bayesian additive regression trees (BART), a non-parametric Bayesian regression technique. The likelihood function is complex as it is based on a sum of trees whose structures change dynamically with the MCMC iterates. An extension to BART is developed to estimate the heterogeneous treatment effect, aiming to overcome the regularisation-induced confounding issue which is often observed in the direct application of BART in causal inference.

Acknowledgement

First and foremost, I would like to thank my supervisors: Professor Robert Kohn, Professor Scott Sisson and Dr. Matias Quiroz for their consistent support, guidance and inspiration throughout my PhD study. Each of them has great impacted on the way that I think about research and problem-solving. Due to the pandemic, almost one third of my PhD time was off-campus, which means I had to work from home and spend all day alone with the numerous problems encountered in the research. My supervisors paid great attention to my research and provided constant support through the weekly online meetings. Without their encouragement and help, I could not have completed this thesis.

I am also incredibly grateful for the faculty members in the UNSW School of Economics and School of Mathematics and Statistics. Among them, I want to acknowledge Dr. Boris Beranger for the invaluable advice and contribution for Chapter 3. I also owe thanks to Dr. Keiichi Kawai for providing financial funding for conferences. Also, special thanks to the Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS) for providing an excellent platform for students to network with senior researchers. I received valuable feedback for my ongoing research from the retreats organised by ACEMS.

To my fellow PhD students and friends, my PhD study would have been lonely and painful without you. Thank you all for helping and encouraging me in the completion of the first year coursework. As a student who did not know any economics before, your kindness meant a lot to me. I would like to say a big thank you to Ningyi Sun, who offered numerous tea break chats during intense study and patiently answered random questions I raised. My sincere gratitude also goes to Vincent Chin, Doan Khue Dung, Jaslene Lin, Hung Dao, Igor Balnozan, Dandan Yu, Kaiyin Hu, Yangqi Zhang and Di Tian.

Last but not least, I am grateful to my parents for the love and care throughout the journey. Your support and encouragement is my biggest motivation to complete this challenging journey. I also cannot fail to recognise my husband, Jialin Yang, who has always been at my side for every hard decision I made. I owe so much of my success to you.

Contents

List of Figures

List of Tables

Abbreviations

ABC	Approximate Bayesian computation
AIS	Annealed importance sampling
ATE	Average treatment effect
ATT	Average treatment effect on the treated group
BART	Bayesian additive regression tree
BCF	Bayesian causal forest
CART	Bayesian classification and regression tree
CATE	Conditional average treatment effect
ESS	Effective sample size
GRF	Generalised random forest
IACT	Integrated autocorrelation time
MAE	Mean absolute error
MAPE	Mean absolute percentage error
MCMC	Markov chain Monte Carlo
MET	Minimax-exponentially-tilted estimator
MH	Metropolis-Hasting Algorithm
MLE	Maximum likelihood estimator
OLS	Ordinary least squares
PM	Pseudo-marginal
RIC	Regularisation-induced confounding
RMSE	Root mean squared error

SDA	Symbolic data analysis
SMC	Sequential Monte Carlo
SOV	Separation of variables
SUTVA	The stable unit treatment value assumption

Chapter 1

Introduction

Statistical analysis develops models to investigate data generating processes involving observations together with domain knowledge. Advances in computing power enable the possibility of analysing complicated models on massive amounts of data. The main aim of computational statistics is to develop and understand computationally intensive statistical methods from both a computational science and statistics perspective. Computational statistics is the backbone of modern data science as conducting data analysis inevitably requires computing. This is particularly true for Bayesian analysis, where the output of the analysis is the posterior distribution which quantifies the uncertainty of the inference. In most applications, the posterior distribution is intractable and various algorithms are available to draw samples from it, but often have large computational requirements. The main aim of this thesis is to develop novel methods for Bayesian analysis with specific applications in symbolic data analysis (SDA), doubly intractable problems and causal inference. Some applications, such as doubly intractable problems and causal inference, are widely studied in the literature with ongoing developments. Some are relatively new branches in modern statistics, such as SDA, but with promising potential in the era of big data. The motivation is application specific, but the overall target is to develop efficient algorithms to cope with complex models. The remaining part of this chapter briefly introduces Bayesian statistics and its rising challenges.

The well-known Bayes' theorem is

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{p(\mathbf{y})}, \tag{1.1}$$

where $p(\mathbf{y}|\boldsymbol{\theta})$ is the likelihood function of parameter(s) $\boldsymbol{\theta}$ given data $\mathbf{y}$, $\pi(\boldsymbol{\theta})$ is the prior and $p(\mathbf{y})$ is often called the data evidence or the marginal likelihood.

A fundamental problem in computational statistics is to obtain the expectation of $E_{\pi(\boldsymbol{\theta}|\mathbf{y})}\left[\psi(\boldsymbol{\theta})\right]$ of a function ψ of $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ with respect to the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{y})$, which is equivalent to evaluating the integral

$$\int_{\boldsymbol{\theta}} \psi(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}.$$

A closed form solution for the integral above rarely exists for a high-dimensional vector $\boldsymbol{\theta}$, but it can be approximated by Monte Carlo integration, giving

$$E_{\pi(\boldsymbol{\theta}|\mathbf{y})}\left[\psi(\boldsymbol{\theta})\right] \approx \frac{1}{M}\sum_{m=1}^{M} \psi(\boldsymbol{\theta}^{(m)}),$$

where $\boldsymbol{\theta}^{(m)}$ is drawn from the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{y})$.

One of the challenges in Bayesian inference is dealing the unknown $p(\mathbf{y})$ defined as $p(\mathbf{y}) = \int_{\boldsymbol{\theta}} p(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$, which makes evaluating the posterior $\pi(\boldsymbol{\theta}|\mathbf{y})$ intractable in most applications. A popular method to conduct Bayesian analysis is Markov chain Monte Carlo (MCMC), which draws correlated samples of $\boldsymbol{\theta}$ from $\pi(\boldsymbol{\theta}|\mathbf{y})$ without needing to know $p(\mathbf{y})$. However, MCMC is not applicable for many inference problems. For example, it may not be possible to evaluate the likelihood function $p(\mathbf{y}|\boldsymbol{\theta})$ point-wise in $\boldsymbol{\theta}$. For some problems, pseudo-marginal (PM) methods replace the likelihood function with its unbiased estimator. The likelihood function can also be computationally demanding when it involves a variable number of parameters, which may lead to inefficiencies within MCMC algorithms. Chapters 3 and 4 investigate modelling with intractable likelihood functions in Bayesian analysis. In Chapter 5, the focus shifts to Bayesian non-parametric regression for causal inference, where the likelihood function is provided as a sum-of-trees model having no fixed structures.

Chapters 3, 4 and 5 are composed of three separate and self-contained articles. To maintain the consistency of the thesis, detailed reviews of the methodologies (MCMC and PM methods) are covered in Chapter 2. The chapter also covers the background knowledge for the applications in the remaining chapters. As a result, there is some overlapping material between the literature reviews and the technical chapters, with repeated definitions and explanation of notations when necessary.

Chapter 3 investigates models that often have an intractable likelihood function in the context of SDA, where the likelihood function is characterised by an integral raised to a large power, i.e., $p(\mathbf{y}|\boldsymbol{\theta}) \propto [\int_{\mathbf{y}} f(\mathbf{y}|\boldsymbol{\theta})d\mathbf{y}]^n$ with n a large number and $f(\mathbf{y}|\boldsymbol{\theta})$ a density function. This likelihood function is built using "symbolic data", which contain structural information about the data, derived from the original data set. The integral can often be intractable, although an unbiased estimator may be available. To overcome the intractability, we utilise the PM method. As the exact computation of the likelihood estimate is usually expensive, an approximate method is proposed to speed up the algorithm with minor difference in the results. The approximate method is applied to a factor model and a linear regression involving heteroscedasticity. It achieves significantly less computing time compared to that of the full data, with a tolerable difference in terms of the accuracy.

In Chapter 4, the likelihood function involves an unknown normalising constant that depends on the parameters of interest. This leads to the so-called doubly intractable problem in Bayesian analysis. The PM method is adapted again in combination with an auxiliary variable approach to generate simulation consistent results. Compared with existing methods in the literature, the proposed algorithm has favourable properties such as being more widely applicable and having available guidelines for hyperparameter tuning.

Chapter 5 focuses on the Bayesian additive regression tree (BART), a non-parametric Bayesian regression technique, which uses regression trees to fit a highly non-linear response surface with a variable number of parameters. To avoid overfitting, BART uses a regularisation prior so that each tree is able to explain part of the relationship between the dependent variable and the covariates. Even though BART is a flexible predictive model, its application in causal inference produces a bias in the estimated causal effects in the presence of confounding, where the treatment and the outcome are both affected by other variables (usually unobserved or omitted). The chapter extends the original BART model for estimating heterogeneous treatment effects of an observational study with binary treatments and continuous outcomes. Such an extension overcomes the regularisation-induced confounding issue and provides more accurate results compared with the existing tree-related methods.

Finally, Chapter 6 concludes and discusses potential future research work.

Chapter 2

Literature review

The literature review consists of two parts: the generic methods used in this thesis, and a brief background introduction for each of the technical chapters.

2.1 The Markov chain Monte Carlo simulation method

Bayesian methods target the posterior distribution of the parameter(s) $\boldsymbol{\theta}$ given the data $\mathbf{y}$, which is expressed as $\pi(\boldsymbol{\theta}|\mathbf{y}) = p(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})/p(\mathbf{y})$, where $p(\mathbf{y}|\boldsymbol{\theta})$ is the likelihood and $\pi(\boldsymbol{\theta})$ is the prior. The marginal likelihood $p(\mathbf{y}) = \int_{\boldsymbol{\theta}} p(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$, and this integral can be high-dimensional. For most complex models, the posterior distribution is unknown, which poses challenges for sampling directly from it.

Markov chain Monte Carlo (MCMC) (Brooks et al., 2011) is the most popular class of algorithms to draw samples from the posterior distribution. It produces correlated samples from the posterior distribution using a properly designed Markov chain which has $\pi(\boldsymbol{\theta}|\mathbf{y})$ as its invariant distribution. MCMC techniques are not only useful in Bayesian inference. They are often applied to solve integration and optimisation problems, having wide applications in statistical mechanics, physics, machine learning etc. (Andrieu et al., 2003; Jerrum and Sinclair, 1996).

2.1.1 The Metropolis-Hastings algorithm

The Metropolis-Hastings (MH) algorithm is an important MCMC algorithm (Hastings, 1970; Metropolis et al., 1953). It designs a Markov process by constructing the transition from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}'$ from the density $q(\boldsymbol{\theta}'|\boldsymbol{\theta})$, from which $\boldsymbol{\theta}'$ is drawn. The transition from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}'$ is accepted with probability:

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \min\left\{1, \frac{\pi(\boldsymbol{\theta}')p(\mathbf{y}|\boldsymbol{\theta}')q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})q(\boldsymbol{\theta}'|\boldsymbol{\theta})}\right\}; \tag{2.1}$$

the chain remains at $\boldsymbol{\theta}$, otherwise.

Algorithm 1 describes the MH approach which is frequently used in Bayesian inference due to its simplicity and general applicability. The marginal likelihood $p(\mathbf{y})$ is not required for MH because it cancels out in the MH acceptance ratio based on (2.1). The only requirement is that the likelihood function of the model is analytically tractable, i.e., it can be evaluated point-wise in $\boldsymbol{\theta}$.

Algorithm 1 The Metropolis-Hastings algorithm

```
1: Initialise $\boldsymbol{\theta}$ such that $\pi(\boldsymbol{\theta}|\mathbf{y}) > 0$.
2: Propose $\boldsymbol{\theta}'$ from $q(\boldsymbol{\theta}'|\boldsymbol{\theta})$.
3: Calculate the MH acceptance ratio:
       $\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}') = \min\left\{1, \frac{\pi(\boldsymbol{\theta}')p(\mathbf{y}|\boldsymbol{\theta}')q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})q(\boldsymbol{\theta}'|\boldsymbol{\theta})}\right\}.$
4: Generate $u_{ar} \sim \text{Uniform}(0, 1)$.
5: Update $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}'$ if $\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}') > u_{ar}$. Otherwise, $\boldsymbol{\theta}$ is unchanged.
6: Repeat Step 2-5 based on the updated $\boldsymbol{\theta}$ for a prefixed number of iterations.
```

The MH algorithm is designed to obey the detailed balance condition, which states that the probability of the chain being in a set A (under the stationary distribution) and moving to a general set B is the same with A and B reversed. To satisfy the condition, the proposal distribution must have a positive density function in the support of the parameter space of the posterior distribution. For example, a random walk proposal usually fulfils such requirement and is often used in the MH algorithm. A random walk proposal with the normal density for the scalar θ' is defined as $q(\theta'|\theta) = N(\theta'; \theta, \sigma^2)$, which is a normal distribution with mean θ and variance σ^2, and so θ' is sampled from a normal distribution with mean θ and variance σ^2. A large σ^2 results in too many rejections whereas a small σ^2 results in slow movement in the parameter space. Both cases lead to poor efficiency of the MH algorithm.

Theoretical results suggest that for a multi-dimensional posterior distribution formed from independent and identically distributed (iid) components, the acceptance ratio of a multivariate proposal should be around 0.234 to achieve the optimal asymptotic efficiency of the MH algorithm (Gelman et al., 1997; Neal and Roberts, 2006; Roberts and Rosenthal, 2001). However, a universal proposal distribution does not exist for several reasons. For example, multi-modality of the posterior distribution can lead to slow convergence of the chain. Adaptive MH algorithms (Roberts and Rosenthal, 2009) are popular methods for constructing the proposal distribution; they tune the proposal distribution automatically based on information obtained from previous draws. Numerous algorithms have been built for fast and reliable adaptive MH algorithms (Atchadé and Rosenthal, 2005; Giordani and Kohn, 2010; Haario et al., 2001). In this thesis, we adopt the approach proposed by Garthwaite et al. (2016), where a stochastic search algorithm based on the Robbins-Monro process (Robbins and Monro, 1951) is adapted, so that the scale parameter of a Gaussian random walk proposal is automatically tuned to target a prespecified value of the overall sampler acceptance probability. If a specific proposal is rejected, then the scalar σ decreases for the next iteration. Otherwise, a larger σ is used instead to encourage a larger step size. The changes in σ decrease in magnitude as the algorithm runs, so that σ converges. For a multivariate parameter, Garthwaite et al. (2016) combine the Robbins-Monro process with the strategy in Haario et al. (2001) to construct a positive-definite covariance matrix for the Gaussian proposal.

2.1.2 The Gibbs sampler

The Gibbs sampler is a method for sampling $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$ from $\pi(\boldsymbol{\theta}|\mathbf{y})$ over at least two dimensions, with θ_k a univariate parameter or a parameter block. The Gibbs sampler is proposed by Geman and Geman (1984) and formalised by Gelfand and Smith (1990). It assumes that the conditional distributions of $\theta_i, i = 1, \ldots, d$, are tractable to work with. Instead of sampling from $\pi(\boldsymbol{\theta}|\mathbf{y})$ which is impractical for many models, sampling from the full conditional distribution $\pi(\theta_i|\boldsymbol{\theta}_{j\neq i}, \mathbf{y})$ is often more straightforward. The notation $\boldsymbol{\theta}_{j\neq i}$ refers to all the parameters (parameter blocks) except for the ith parameter (parameter block). In each iteration of the Gibbs sampler, θ_i is updated in turn, with the remaining parameters $\boldsymbol{\theta}_{j\neq i}$ being fixed. The sampling order can be fixed before the Gibbs sampling begins. We also note that each conditional distribution may be for a parameter block, i.e., a vector parameter.

Algorithm 2 The Gibbs sampler

```
1: Initialise $\theta_i$, $i = 1, \ldots, d$, such that $\pi(\boldsymbol{\theta}|\mathbf{y}) > 0$ where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$.
2: Update $\theta_i$ from $\pi(\theta_i|\boldsymbol{\theta}_{j\neq i}, \mathbf{y})$ for $i = 1, \ldots, d$.
3: Repeat Step 2 for a fixed (in advance) number of iterations.
```

The Gibbs sampler (Algorithm 2) is a special case of the MH algorithm, with the proposal distribution $q(\boldsymbol{\theta}|\boldsymbol{\theta}')$ defined as $\pi(\theta_i|\boldsymbol{\theta}_{j\neq i}, \mathbf{y})$. The MH acceptance ratio is one in this case. Compared with the MH algorithm, Gibbs sampling updates $\boldsymbol{\theta}$ in each iteration and does not require a specially designed proposal distribution. The main drawback of Gibbs sampling is that it can have poor sampling efficiency if the blocks of $\boldsymbol{\theta}$ are correlated, i.e., the chain can traverse the parameter space slowly, especially in high dimensions. The Gibbs sampler is also limited in many cases as the full conditional distribution for each element of $\boldsymbol{\theta}$ may be unavailable. Running a MH update within the Gibbs sampler is an alternative to the full Gibbs sampler (Gilks et al., 1995).

2.2 The pseudo-marginal method

In many statistical models, the likelihood function $p(\mathbf{y}|\boldsymbol{\theta})$ is analytically or computationally intractable. One example is state space models, where the likelihood of the parameters after integrating out the states is often intractable, but can be estimated unbiasedly by particle filters (Shephard and Pitt, 1997). In such cases, the MH acceptance ratio (2.1) is computationally intractable as $p(\mathbf{y}|\boldsymbol{\theta})$ cannot be analytically evaluated. Beaumont (2003) replaces the likelihood function with its unbiased estimator. The idea is formally studied and described in Andrieu and Roberts (2009), who name the approach the pseudo-marginal (PM) method. Loosely speaking, the paper establishes the conclusion that when an unbiased and positive estimator for the likelihood is used inside the MH algorithm, the algorithm provides exact samples from the posterior distribution even though the likelihood is estimated.

Andrieu and Roberts (2009); Flury and Shephard (2011); Pitt et al. (2012) explain the PM method using an auxiliary variable representation. Suppose the likelihood estimator is expressed as $\widehat{p}(\mathbf{y}|\boldsymbol{\theta}, \mathbf{u})$, where $\mathbf{u}$ contains the random numbers involved in the construction of the estimate. Without loss of generality, $\mathbf{u}$ is assumed to be independent of $\boldsymbol{\theta}$ from here on. If we assume that $\widehat{p}(\mathbf{y}|\boldsymbol{\theta}, \mathbf{u})$ is an unbiased estimator of $p(\mathbf{y}|\boldsymbol{\theta})$, then

$$\int \widehat{p}(\mathbf{y}|\boldsymbol{\theta}, \mathbf{u}) p(\mathbf{u}) d\mathbf{u} = p(\mathbf{y}|\boldsymbol{\theta}).$$

The joint posterior density of $\mathbf{u}$ and $\boldsymbol{\theta}$ is defined as

$$\pi(\boldsymbol{\theta}, \mathbf{u}|\mathbf{y}) \propto \widehat{p}(\mathbf{y}|\boldsymbol{\theta}, \mathbf{u}) \pi(\boldsymbol{\theta}) p(\mathbf{u}), \tag{2.2}$$

and it is easy to verify that the density integrates to one as $\widehat{p}(\mathbf{y}|\boldsymbol{\theta}, \mathbf{u})$ is an unbiased estimator. Moreover, the marginal distribution of $\boldsymbol{\theta}$, $\int_{\mathbf{u}} \pi(\boldsymbol{\theta}, \mathbf{u}|\mathbf{y}) d\mathbf{u}$, is $\pi(\boldsymbol{\theta}|\mathbf{y})$, the posterior distribution of interest.

A key issue in implementing the PM method is controlling the variability of the likelihood estimator, specifically, the variance of the logarithm of the likelihood estimator. If the variance is too large, the chain is likely to get stuck at some local region due to overestimating the likelihood function at some point, leading to poor efficiency of the algorithm.

The overestimation can be caused by a set of random numbers used in generating the likelihood estimation, which results in the corresponding likelihood estimate being too large. As a consequence, the following proposal, which might be close to the current parameter value but with a "normal" likelihood estimate, is unlikely to be accepted. A small variance is desirable, but it may be too computationally costly to reduce the variance. The optimal variance of the logarithm of the likelihood estimator should be approximately in the range of 1 to 3 to achieve an optimal trade-off between sampling efficiency and computational cost (Doucet et al., 2015; Pitt et al., 2012; Schmon et al., 2021; Sherlock et al., 2015). When implementing the PM method, the number of particles (samples used in Monte Carlo integration) is set to target a variance in this range.

The original PM method uses independent sets of random numbers for generating likelihood estimates evaluated at the current and proposed parameters. For an estimator with a large variance (larger than 3), the PM method can be very inefficient. Several authors show that PM methods benefit from updating the random numbers in a way that induces a correlation between the logarithms of the estimators at the current and proposed draws. Deligiannidis et al. (2018) obtain this by correlating the random numbers used in constructing the estimators. A high correlation between these estimators significantly reduces the number of particles required for optimal implementation. Tran et al. (2016) propose an alternative approach called the block pseudo-marginal (BPM) method, which controls the correlation of the estimators more directly than Deligiannidis et al. (2018). The method is more efficient than that of Deligiannidis et al. (2018) for some problems. The BPM method divides the random numbers for $\mathbf{u}$ (in the numerator of the MH ratio) into blocks and updates $\boldsymbol{\theta}$ jointly with one block of $\mathbf{u}$ to form $\boldsymbol{\theta}'$ and $\mathbf{u}'$ in the denominator (see Algorithm 3). With many blocks, a high correlation is induced between the numerator and the denominator of the log MH acceptance ratio. Tran et al. (2016) also provide guidelines for selecting the optimal number of particles for an efficient implementation of the BPM algorithm.

The PM method discussed above assumes the existence of a positive and unbiased estimator for the likelihood function. If the estimator $\widehat{p}(\mathbf{y}|\boldsymbol{\theta}, \mathbf{u})$ is not necessarily positive, then

Algorithm 3 The block pseudo-marginal algorithm

```
1: Initialise $\boldsymbol{\theta}$ such that $\pi(\boldsymbol{\theta}|\mathbf{y}) > 0$ and generate a collection of random numbers $\mathbf{u} = (u_1, \ldots, u_B)$ with $B$ blocks.
2: Propose $\boldsymbol{\theta}'$ from $q(\boldsymbol{\theta}'|\boldsymbol{\theta})$.
3: Propose $\mathbf{u}'$ by randomly updating one block of $\mathbf{u}$. The proposal for $\mathbf{u}'$ is usually independent of $\boldsymbol{\theta}$, i.e., $q(\mathbf{u}'|\mathbf{u}, \boldsymbol{\theta}, \boldsymbol{\theta}') = q(\mathbf{u}'|\mathbf{u})$.
4: Calculate the acceptance ratio by
       $$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}') = \min\left\{1, \frac{\pi(\boldsymbol{\theta}')\widehat{p}(\mathbf{y}|\boldsymbol{\theta}', \mathbf{u}')q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta})\widehat{p}(\mathbf{y}|\boldsymbol{\theta}, \mathbf{u})q(\boldsymbol{\theta}'|\boldsymbol{\theta})}\right\}.$$
       ▷ Here we assume a symmetric distribution for $q(\mathbf{u}'|\mathbf{u})$, so that the two proposal density functions, $q(\mathbf{u}'|\mathbf{u})$ and $q(\mathbf{u}|\mathbf{u}')$, cancel out in the acceptance ratio.
5: Update $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}', \mathbf{u} \leftarrow \mathbf{u}'$ if $\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}') > u_{ar}$ with $u_{ar} \sim \text{Uniform}(0, 1)$. Otherwise, $\boldsymbol{\theta}, \mathbf{u}$ are unchanged.
6: Repeat Step 2-5 based on the updated $\boldsymbol{\theta}, \mathbf{u}$ for a fixed number of iterations.
```

the corresponding $\widehat{\pi}(\boldsymbol{\theta}, \mathbf{u}|\mathbf{y})$ is not guaranteed to be positive, as the PM method defines a density on the augmented space of $\mathbf{u}$ and $\boldsymbol{\theta}$ in (2.2). Given a negative value of $\widehat{p}(\mathbf{y}|\boldsymbol{\theta}, \mathbf{u})$, the posterior density function of $\mathbf{u}$ and $\boldsymbol{\theta}$, $\widehat{\pi}(\boldsymbol{\theta}, \mathbf{u}|\mathbf{y})$ is invalid. Lyne et al. (2015) propose a solution to the issue by running the PM sampler on the absolute value of the likelihood estimate, which leads to an unbiased estimator. An importance sampling correction is applied to consistently estimate the posterior mean with respect to $\pi(\boldsymbol{\theta}|\mathbf{y})$ of any function of the parameters. We call this approach "the signed PMMH algorithm". As Lyne et al. (2015); Quiroz et al. (2021) point out, even though the signed PMMH algorithm eliminates the effect of negative estimates, the posterior mean has a big Monte Carlo error if a large portion of the estimates are negative.

Algorithm 4 describes the signed PMMH algorithm. The BPM algorithm (Algorithm 3) can be easily incorporated into the algorithm; we omit the details here. Both Chapter 3 and Chapter 4 provide detailed steps of the signed block PMMH algorithm for specific problems.

Algorithm 4 The pseudo-marginal Metropolis-Hastings with importance sampling sign correction (signed PMMH) algorithm

```
1: Initialise $\boldsymbol{\theta}$ such that $\pi(\boldsymbol{\theta}|\mathbf{y}) > 0$.
2: Propose $\boldsymbol{\theta}'$ from $q(\boldsymbol{\theta}'|\boldsymbol{\theta})$.
3: Calculate the acceptance ratio by
       $$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}') = \min\left\{1, \frac{\pi(\boldsymbol{\theta}')|\widehat{p}(\mathbf{y}|\boldsymbol{\theta}')|q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta})|\widehat{p}(\mathbf{y}|\boldsymbol{\theta})|q(\boldsymbol{\theta}'|\boldsymbol{\theta})}\right\}.$$
4: Update $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}'$ if $\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}') > u_{ar}$ with $u_{ar} \sim$ Uniform$(0, 1)$. Otherwise, $\boldsymbol{\theta}$ is unchanged.
5: Track the sign of $\widehat{p}(\mathbf{y}|\boldsymbol{\theta})$ for the updated $\boldsymbol{\theta}$: $\text{sign}(\mathbf{y}|\boldsymbol{\theta}) = 1$ if $\widehat{p}(\mathbf{y}|\boldsymbol{\theta}) \geq 0$, otherwise $\text{sign}(\mathbf{y}|\boldsymbol{\theta}) = -1$.
6: Repeat Step 2-5 based on the updated $\boldsymbol{\theta}$ for a prefixed number of iterations $M$.
7: The posterior expectation of any function $h(\boldsymbol{\theta})$ is
       $$\int h(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} = \frac{\sum_{m=1}^{M} h(\boldsymbol{\theta}^{(m)})\text{sign}(\mathbf{y}|\boldsymbol{\theta}^{(m)})}{\sum_{m=1}^{M} \text{sign}(\mathbf{y}|\boldsymbol{\theta}^{(m)})}.$$
```

2.3 Symbolic data analysis (SDA)

This section first shows how to obtain "symbolic data" (termed "symbols") from "classical data" in Section 2.3.1, where the original data is summarised into a symbolic data set of manageable size. As a large proportion of the symbolic data are presented in the format of intervals, Section 2.3.2 explains interval-valued data with its associated statistics. The section also covers methods for the analysis of interval-valued random variables. Section 2.3.3 outlines the likelihood-based approaches which fit symbols with parametric models.

2.3.1 From classical to symbolic data

Modern statistical models often analyse "classical data" of which realisations are typically single points in Euclidean space $\mathcal{X} \subseteq \mathbb{R}^p$. A large data matrix of size $n \times p$ can be formed accordingly, where each row takes the values from one individual, with the p variables represented in columns. Denote the matrix as $\mathbf{X} = (X_{ij})$ where i represents the ith observation and j is the jth variable. Let x_{ij} denote the observed value of the variable X_{ij} and $\mathcal{X}_j$ denote the domain of the jth variable X_j. Then $\mathbf{X} = (X_1, \ldots, X_p)$ takes value in $\mathcal{X} = \times_{j=1}^{p} \mathcal{X}_j$. Table 2.1 shows a typical data set taken from Billard and Diday (2003). The variables X_{ij} can be quantitative, with the response being a continuous variable such as systolic blood pressure $X_{systolic}$ or the discrete variable X_{age}. The variables can also be categorical. For example, the variable X_{city} is a categorical variable with the domain being a collection of cities $\mathcal{X}_{city} = \{\text{Boston}, \text{Chicago}, \text{El Paso}, \ldots\}$. For the variable X_{cancer}, the domain is not uniform as it can take a binary value $\mathcal{X}_{cancer} = \{\text{Yes}, \text{No}\}$ or a disease name $\mathcal{X}_{cancer} = \{\text{lung}, \text{brain}, \text{breast}, \ldots\}$. What is crucial in the classical data setting is that for each observation X_{ij}, there is only one realisation associated with the variable.

			Age	Systolic pressure	Diastolic pressure	
i	City	Gender	(years)	(mmHg)	(mmHg)	Cancer
1	Boston	M	24	120	79	No
2	Boston	M	56	130	90	No
3	Chicago	M	48	126	82	Lung
4	El Paso	F	47	121	86	Yes
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Table 2.1: Sample data set: Classical data.

In contrast, symbolic data contain the internal variation and structural information extracted from the original data set. The idea is to condition on some variables taking specific values, e.g. X_{city} = Boston, and then collapse the other variables to a distributional form. For example, the collapsed variable for X_{age} given the conditioning variable X_{city} = Boston would have a discrete distribution over positive integers, with the mass on $X_{age} = k$ being proportional to the number of individuals in the data set with X_{city} = Boston and $X_{age} = k$. As introduced by Diday (1989), symbolic data represent and summarise the "classical data" in such a way that the resulting symbolic data set is of manageable size and retains as much information as possible from the original data set. The choice of conditioning variable(s) is related to the subject of the analysis.

Table 2.2 (extracted from Billard and Diday, 2003) provides more examples of symbolic data obtained from the data set in Table 2.1. The symbolic objects u refer to groups of individual classical data points. The number of symbols is usually less than the number of entities n, because each symbolic data point contains the information derived of a set of observations. In order to describe the variation of the collapsed variables within each symbol, the symbolic variable can take, e.g., an interval-valued format. For example, the first row of Table 2.2 represents (as the collapsed variable) a male in his 20s, living in Boston, who has a brain tumour with a blood pressure 120/79 mmHg (as the conditioning variables). The object u can be a collection of individuals who have the same characteristics, or it can be interpreted as a specific male individual followed over a 10-year period with the Age records falling in the interval [20,30). The interval-valued format is not the only option for symbolic data. Other common representations of symbols include histograms, distributions, set of categories, etc. For example, the 4th row of Table 2.2 may present the same individual (4th) in Table 2.1, with the type of cancer undetermined, or a group of people with different cancer types. Here a distribution is used to state the type of cancer: A probability p of having lung cancer, and $(1 - p)$ of having breast cancer.

u	Age	Blood pressure (mmHg)	City	Type of cancer	Gender
1	[20,30)	(79/120)	Boston	{brain tumor}	{Male}
2	[50.60)	(90/130)	Boston	{lung, liver}	{Male}
3	[45,55)	(80/130)	Chicago	{prostate}	{Male}
4	[47,47]	(86/121)	El Paso	{breast p, lung 1-p}	{Female}
⋮	⋮	⋮	⋮	⋮	⋮

Table 2.2: Sample data set: Symbolic data.

It is worth pointing out that the symbol types are not limited to the ones mentioned above. Researchers may extract information and formulate the symbols according to the requirements of their analyses. For instance, Diday and Vrac (2005) use the cumulative density function as the format of symbols and model a partition problem by copulas. In this thesis, we mainly focus on the symbols relating to interval-valued variables. Chapter 3 presents more details of this approach to symbol construction.

Symbolic data can arise naturally through the process of data collection and processing. The original data may be collected as lists, intervals, histograms, distributions etc. One naturally occurring example is blood pressure which changes continuously, and so naturally takes a range of values. Symbolic objects may also be constructed for large data sets where it is too expensive to conduct inference on the original data. Hence, SDA has great potential for inference and data storage in the era of big data, where large volumes of data are becoming commonplace.

2.3.2 Interval-valued symbolic data

A symbolic observation can take multiple forms such as an interval, a list, a histogram or even a distribution of values. Bock and Diday (1999) formally introduce the concept of symbolic data. Bertrand and Goupil (2000) derive descriptive statistics for symbolic data such as sample mean, standard deviation etc. Symbolic data can be classified into quantitative or qualitative types. The former includes, but is not limited to, interval-valued and histogram-valued symbols. The latter includes categorical multi-valued variables. This section focuses on interval-valued variables which are analysed in Chapter 3.

Following Bertrand and Goupil (2000), suppose an interval-valued symbol S_{ij} of the ith observation ($i = 1, \dots, n$) and the jth variable ($j = 1, \dots p$) follows a uniform distribution with an observed interval (a_{ij}, b_{ij}). Let W_j be a point in S_{ij}. Then, for an arbitrary w,

$$\Pr(W_j < w) = \begin{cases} 0, & \text{if } w < a_{ij} \\ (w - a_{ij})/(b_{ij} - a_{ij}), & a_{ij} \leq w < b_{ij}, \\ 1, & b_{ij} \leq w. \end{cases}$$

The symbolic sample mean $\overline{W}_j$ and the symbolic sample variance S_j^2 of W_j are

$$\begin{aligned} \overline{W}_j &= \frac{1}{2n} \sum_{i=1}^{n} (a_{ij} + b_{ij}), \\ S_j^2 &= \frac{1}{3n} \sum_{i=1}^{n} (a_{ij}^2 + a_{ij} b_{ij} + b_{ij}^2) - \frac{1}{4n^2} \left[\sum_{i=1}^{n} (a_{ij} + b_{ij}) \right]^2 . \end{aligned}$$

Billard (2007, 2008) further show that the symbolic sample variance (SST) can be decomposed into the sum of interval variation (SSW) and external variation (SSB).

$$\begin{aligned} nS_j^2 &= SST = SSB + SSW \\ SSB &= \sum_{i=1}^{n} (\overline{W}_{ij} - \overline{W}_j)^2 \\ SSW &= \sum_{i=1}^{n} (b_{ij} - a_{ij})^2/12, \end{aligned}$$

where $\overline{W}_{ij} = (a_{ij} + b_{ij})/2$ and $\overline{W}_j$ is the symbolic sample mean defined above. Note that the result for SSW is consistent with the variance expression under the assumption of uniformity. In this thesis, we relax such assumptions (in practice, the data between a_{ij} and b_{ij} are unlikely to be uniform); see Chapter 3 for more detail. The definition of symbols of other formats is in Bertrand and Goupil (2000); Bock and Diday (1999). It is worth pointing out that "classical data" is a special case of symbolic data, where the distribution relating to symbols, e.g., the uniform distribution on interval-valued variables, has probability mass one on a single point.

2.3.3 Methods for interval-valued symbolic data

Symbolic data have a special structure, i.e., they have a specific structure such as interval-valued variables, histogram-valued variables. Hence, methodologies developed for "classical data" are inappropriate. For example, conventional methods such as linear regression methods cannot be directly applied to interval-valued variables. A number of papers explore different methods for modelling interval-valued symbolic data.

Explanatory analysis: Cazes et al. (1997) and Chouakria et al. (1998) develop principle component methods for analysing symbolic variables. Gowda and Diday (1991) introduce a dissimilarity measure for Boolean symbolic objects and later Billard and Diday (2003) extend clustering methods for symbolic objects based on the criteria proposed by Chavent (1998). Chavent and Lechevallier (2002) and De Carvalho and Tenório (2010) propose clustering methods based on various criteria. Brito (2002) presents hierarchical and pyramid clustering methods for SDA.

Regression analysis: Linear regression analysis builds models describing relationships between variables. It has been intensively investigated in the SDA literature. The variable of interest is usually one-dimensional, known as the response variable (Y) and the other variables are called explanatory/independent variables, denoted as $X_1, \ldots, X_p$, with the observed values denoted as $x_1, \ldots, x_p$. The linear regression models assume that Y is a linear combination of $X_1, \ldots, X_p$, with the weights determined by the parameters $\beta_0, \ldots, \beta_p$, plus a noise term ϵ. Mathematically, the model of an observation Y_i, $i = 1, \ldots, n$, is expressed as

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i,$$

or equivalently in matrix form,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where the design matrix $\mathbf{X}$ is of size $n \times (p+1)$, which collects $x_1, \ldots, x_p$ as the column vectors. The first column of $\mathbf{X}$ is a column vector of 1, presenting the variable (a constant) associated with the intercept β_0. The least squares estimator for $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_p]^T$ can be obtained by minimising $\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \ldots \beta_p x_{ip})^2$. The solution is $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{Y}$ with $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \ldots, \widehat{\beta}_p)^\top$.

Bock and Diday (1999) develop the first linear regression model, known as the centre method, for symbolic data where the centre points of the intervals are used as single points. Loosely speaking, each independent variable is collapsed into a fixed number (n_c) of one-dimensional intervals and the centre points are calculated accordingly from the constructed intervals. For example, suppose the ith interval of X_j is $S_{ij} = [a_{ij}, b_{ij}]$. Then, the centre point is calculated by $S_{ij}^c = (a_{ij} + b_{ij})/2$. A similar rule applies to Y to obtain the centre point S_{yi}^c. The linear model regresses S_{yi}^c against S_{ij}^c, $i = 1, \ldots, n_c$, $j = 1, \ldots, p$, where n_c is the number of central points. The model is

$$\mathbf{S}_y^c = \mathbf{S}^c \boldsymbol{\beta}^c + \boldsymbol{\epsilon}^c,$$

with $\mathbf{S}_y^c$, $\mathbf{S}^c$, $\boldsymbol{\epsilon}^c$ conformal matrices and vectors, which are defined similarly to the regression model on single data points. The predicted value of the model is $\mathbf{S}^{new,c}\widehat{\boldsymbol{\beta}}^c$ and the prediction interval is $[\mathbf{S}_L^{new}\widehat{\boldsymbol{\beta}}^c, \mathbf{S}_U^{new}\widehat{\boldsymbol{\beta}}^c]$ for a new set of variables $\mathbf{S}^{new} = (S_1^{new}, \ldots, S_p^{new})$ with $S_j^{new} = [S_{jL}^{new}, S_{jU}^{new}]$. The construction rule of the interval S_j^{new} is identical to that of S_{ij}. Its lower and upper bounds are $S_{jL}^{new}, S_{jU}^{new}$ respectively.

As an improvement for the centre method, the centre and range method is proposed in De Carvalho et al. (2004); Neto et al. (2004). Two independent linear regression models are used in this approach. The first regression model is the same as the one used in the centre method. The second model is applied to the ranges of the interval-valued variables, where the range of the interval $S_{ij} = [a_{ij}, b_{ij}]$ is defined as $S_{ij}^r = b_{ij} - a_{ij}$. The matrix form of this model is

$$\mathbf{S}_y^r = \mathbf{S}^r\boldsymbol{\beta}^r + \boldsymbol{\epsilon}^r.$$

The predicted value is the same as that of the centre method, and the prediction intervals are constructed using both $\widehat{\boldsymbol{\beta}}^c$ and $\widehat{\boldsymbol{\beta}}^r$. More detail is provided in De Carvalho et al. (2004).

Both methods mentioned above have undesirable characteristics: the predicted upper bounds might be smaller than the lower bounds given negative slope estimates of some elements in $\boldsymbol{\beta}$. To overcome this, Neto and de Carvalho (2010) build a regression model under the condition that all parameters must be positive. The approach is generalised by considering lasso-based constraints on the coefficients (Giordani, 2015). However, forcing a non-negative constraint on parameters may distort the true relationship between variables. Xu (2010) puts a constraint on the prediction bound, which forces the lower bound to be smaller or equal to the upper one. In addition, the author uses the empirical covariance function defined in Billard (2007, 2008) to obtain confidence intervals for the parameters of interest. Lima Neto and dos Anjos (2015) use copulas to model the lower and upper bounds jointly as a random vector.

Time series: Time series analysis aims to model a set of correlated observations. Interval-valued variables arise naturally in time series models as observations are collected as an ordered sequence through time. Teles and Brito (2005) model an interval time series using an autoregressive moving average (ARMA) model with midpoints and ranges of interval-valued variables. Maia et al. (2008) adopt a similar strategy to fit an autoregressive (AR) and autoregressive integrated moving average (ARIMA) models. García-Ascanio and Maté (2010) apply vector autoregressive (VAR) forecasting models to the interval-valued time series data to predict electric power demand per hour in Spain for two years. Ai et al. (2008) investigate the high order interval autoregressive models to analyse the sterling-U.S. dollar exchange rate.

A likelihood-based approach: In the previous section, the regression analysis on interval-valued variables often uses centre points and ranges obtained from the original data, where the symbols (centre points and ranges) are treated as observed data without considering the variation of the data within the interval. As there is no parametric modelling on the symbols, it is usually difficult to carry out inference or conduct hypothesis testing for quantities related to the symbols. In terms of parametric modelling, Le-Rademacher and Billard (2011) propose the first likelihood-based approach for interval-valued and histogram-valued variables. In this approach, each symbol is mapped to a random vector, chosen so that it uniquely defines the symbol. Appropriate distributions are then specified to model the vector. For example, for the interval-valued symbol $S_j = [a_j, b_j]$, define $\mathbf{\Theta} = (\Theta_{j1}, \Theta_{j2})$, with the realisations $\theta_{j1} = (a_j + b_j)/2$ and $\theta_{j2} = (b_j - a_j)^2/12$ respectively. Le-Rademacher and Billard (2011) assume

$$\Theta_{j1} \sim N(\mu, \sigma^2), \quad \Theta_{j2} \sim \text{Exp}(\beta).$$

Then the joint likelihood function is

$$L(\mu, \sigma^2, \beta; \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_m) \propto \prod_{j=1}^{m} [p_1(\theta_{j1}; \mu, \sigma^2) p_2(\theta_{j2}; \beta)],$$

where $\boldsymbol{\theta}_j = (\theta_{j1}, \theta_{j2})$, $j = 1, \dots, m$, and $p_1(\cdot)$, $p_2(\cdot)$ denote the density functions of a normal and an exponential distribution respectively. Maximum likelihood estimates (MLE) for μ, σ^2, β can then be obtained. Brito and Duarte Silva (2012); Lin et al. (2017) take an alternative approach by modelling centre points and log-ranges of interval-valued variables by a multivariate normal or a skew normal distribution to derive estimates accounting for the dependence between the interval centre point and its log-range.

The above methods build parametric models for the parameters associated with interval-valued variables, for example, midpoints and ranges. However, most frameworks ignore the process by which the interval-valued data are constructed from the underlying classical data. Instead, the data are commonly assumed to be uniformly distributed within intervals, which is rarely satisfied in real examples. Consequently, it is unclear how to incorporate the knowledge about the underlying micro-data into the analysis of the symbols. In some cases, researchers are more interested in modelling parameters associated with the original data than those obtained by modelling midpoints and interval ranges. To address this issue, Zhang et al. (2020) develop an inferential framework for interval-valued symbols, which allows direct parametric modelling of the underlying real-valued data in the presence of interval-valued summaries. The approach provides more interpretable results compared with the methods mentioned above. Beranger et al. (2018) further extend the approach by deriving a new general construction rule used for building the likelihood functions for interval-valued and histogram-valued variables. However, the approach is limited to the case of tractable likelihood functions. This thesis extends the approach and uses PM methods to carry out Bayesian inference which facilitates the likelihood-based approach for models with intractable likelihoods. The relevant literature and the construction rule in Beranger et al. (2018) are covered in Chapter 3.

2.4 The doubly intractable problem

2.4.1 Introduction

Suppose that the likelihood function is expressed as

$$p(\mathbf{y}|\boldsymbol{\theta}) = \frac{f(\mathbf{y}|\boldsymbol{\theta})}{Z(\boldsymbol{\theta})}, \quad \text{with } Z(\boldsymbol{\theta}) = \int_{\mathbf{y}} f(\mathbf{y}|\boldsymbol{\theta}) d\mathbf{y},$$

where $f(\mathbf{y}|\boldsymbol{\theta})$ is computable and $Z(\boldsymbol{\theta})$ is an intractable normalising constant depending on the parameter(s) $\boldsymbol{\theta}$. Standard MCMC techniques cannot be applied in such a model, because the acceptance ratio (2.1) cannot be computed as the normalising constant $Z(\boldsymbol{\theta})$

does not cancel out. Models with intractable normalising functions arise in many applications, for instance, exponential random graph models (EGRMs) (Robins et al., 2007) to model social network problems, the Ising and autologistic models for lattice data (Besag, 1974; Hughes et al., 2011) and Gaussian Markov random field models (GMRFs) (Rue and Tjelmeland, 2002) in spatial statistics.

An example of social network data is given below. Figure 2.1 illustrates the friendship connections of 34 individuals in a karate club (Zachary, 1977). An ERGM can be applied to model the data set with the corresponding likelihood function defined as

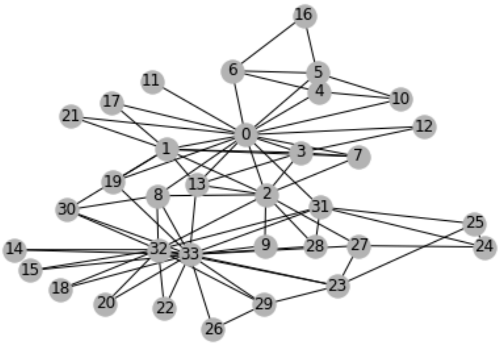


Figure 2.1: Zachary's karate club graph: A social network of 34 individuals with 78 edges.

$$f(\mathbf{y}|\boldsymbol{\theta}) = \frac{\exp\left(\sum_{i=1}^{d} \theta_i s_i(\mathbf{y})\right)}{Z(\boldsymbol{\theta})}, \text{ where } Z(\boldsymbol{\theta}) = \sum_{\mathbf{y} \in \mathcal{Y}} \left(\exp \sum_{i=1}^{d} \theta_i s_i(\mathbf{y}) \right).$$

The term $\mathcal{Y}$ refers to the set of all graphs. The terms $s_i(\mathbf{y})$ $i = 1, \ldots, d$, are usually sufficient statistics which are straightforwardly obtained from a realised graph. Suppose that an undirected graph has n nodes; then the constant $Z(\boldsymbol{\theta})$ for the model involves a sum over all $2^{n(n-1)/2}$ possible graphs, $\mathbf{y}$, making $Z(\boldsymbol{\theta})$ computationally intractable even for a relatively small n.

If a likelihood has an intractable normalising constant, then it is infeasible to calculate its gradient, and likelihood maximisation is then impractical. Besag (1974) introduces the idea of pseudolikelihood estimation, which involves summing over local data, and hence is tractable: the likelihood $f(\mathbf{y}|\boldsymbol{\theta})$ is approximated without $Z(\boldsymbol{\theta})$. However, when a data set has strong dependence and a small sample size, the maximum pseudolikelihood estimator may seriously overestimate the dependence. Geyer (1991) constructs Monte Carlo-based algorithms for approximating the MLE (MCMC-MLE). This approach has motivated the use of various simulation methods for EGRMs; see Snijders (2002) for a review.

2.4.2 Bayesian methods for the doubly intractable problem

Before introducing specific methods designed for doubly intractable problems, a number of approximate methods can be applied to the problem given that their associated assumptions are satisfied. For example, approximate Bayesian computation (ABC) algorithms (Marin et al., 2012) bypass evaluation of the likelihood function with a comparison between the observed data and simulated data. The intractable likelihood functions cancel out in the MH acceptance probability under ABC. Another example is the integrated nested Laplace approximation (INLA) (Rue et al., 2009), which approximates the posterior distribution by a normal density with the mean situated at the posterior mode and the variance obtained by the local curvature of the distribution evaluated at the mode.

In addition to approximate methods, there is a growing literature on Bayesian inference for doubly intractable problems with the aim of exact inference; see Park and Haran (2018) for a comprehensive review of such methods. The algorithms may be classified into two overlapping categories: auxiliary variable approaches and likelihood approximation approaches. Park and Haran (2018) also use the term “asymptotically exact” or “asymptotically inexact” to distinguish whether the stationary distribution of the generated Markov chain is equal to the desired posterior or not. Following Park and Haran (2018), we introduce both approaches below with one representative algorithm of each. Both algorithms are implemented in a simulation study in Chapter 4 to compare with our proposed method.

Auxiliary variable approaches: In general, auxiliary variable approaches cleverly choose the transition kernel of the parameter and the auxiliary variable so that the normalising constant cancels out in the MH acceptance probability. The exchange algorithm proposed by Murray et al. (2006), which includes the auxiliary variable $\mathbf{x}$ and the proposal

$\boldsymbol{\theta}'$ on an augmented space, is an example of such an approach. Let $f(\mathbf{x}|\boldsymbol{\theta}')/Z(\boldsymbol{\theta}')$ be the density function of $\mathbf{x}$ and $q(\boldsymbol{\theta}'|\boldsymbol{\theta})$ be the density function of $\boldsymbol{\theta}'$. Then, the joint density of $\mathbf{x}$ and $\boldsymbol{\theta}'$ conditional on $\mathbf{y}$, $\boldsymbol{\theta}$ is

$$p(\mathbf{x}, \boldsymbol{\theta}'|\mathbf{y}, \boldsymbol{\theta}) = p(\mathbf{x}|\boldsymbol{\theta}')q(\boldsymbol{\theta}'|\boldsymbol{\theta}) = \frac{f(\mathbf{x}|\boldsymbol{\theta}')}{Z(\boldsymbol{\theta}')}q(\boldsymbol{\theta}'|\boldsymbol{\theta}).$$

The posterior distribution on the augmented space is

$$\pi(\boldsymbol{\theta})\pi(\mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y}) \propto \pi(\boldsymbol{\theta})\frac{f(\mathbf{y}|\boldsymbol{\theta})}{Z(\boldsymbol{\theta})}q(\boldsymbol{\theta}'|\boldsymbol{\theta})\frac{f(\mathbf{x}|\boldsymbol{\theta}')}{Z(\boldsymbol{\theta}')}.$$

Since $\int_{\boldsymbol{\theta}'}\int_{\mathbf{x}} \pi(\mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})d\mathbf{x}d\boldsymbol{\theta}' = \pi(\boldsymbol{\theta}|\mathbf{y})$, the marginal distribution of this approach is the distribution of interest.

For each iteration, let $(\boldsymbol{\theta}, \boldsymbol{\theta}')$ be the current parameter setting for $(\mathbf{y}, \mathbf{x})$. Consider a symmetric parameter swapping proposal between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$, which is $(\boldsymbol{\theta}', \boldsymbol{\theta})$ for $(\mathbf{y}, \mathbf{x})$, then the MH acceptance ratio of $\boldsymbol{\theta}'$ together with the dependent auxiliary variable $\mathbf{x}$ is

$$\begin{aligned}
\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}') &= \min\left\{1, \frac{\pi(\boldsymbol{\theta}')\pi(\mathbf{x}, \boldsymbol{\theta}', \boldsymbol{\theta}|\mathbf{y})}{\pi(\boldsymbol{\theta})\pi(\mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})}\right\} \\
&= \min\left\{1, \frac{\pi(\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta})}\,\frac{f(\mathbf{y}|\boldsymbol{\theta}')\cancel{Z^{-1}(\boldsymbol{\theta}')}}{f(\mathbf{y}|\boldsymbol{\theta})\cancel{Z^{-1}(\boldsymbol{\theta})}}\,\frac{q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{q(\boldsymbol{\theta}'|\boldsymbol{\theta})}\,\frac{f(\mathbf{x}|\boldsymbol{\theta})\cancel{Z^{-1}(\boldsymbol{\theta})}}{f(\mathbf{x}|\boldsymbol{\theta}')\cancel{Z^{-1}(\boldsymbol{\theta}')}}\right\} \\
&= \min\left\{1, \frac{\pi(\boldsymbol{\theta}')f(\mathbf{y}|\boldsymbol{\theta}')q(\boldsymbol{\theta}|\boldsymbol{\theta}')f(\mathbf{x}|\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})f(\mathbf{y}|\boldsymbol{\theta})q(\boldsymbol{\theta}'|\boldsymbol{\theta})f(\mathbf{x}|\boldsymbol{\theta}')}\right\}.
\end{aligned}$$

Møller et al. (2006) introduce an alternative approach, where an auxiliary variable $\mathbf{x}$ is proposed with a predefined density conditional on $\boldsymbol{\theta}, \mathbf{y}$. The normalising constant cancels out by utilising the likelihood function as a part of the proposal for $\boldsymbol{\theta}, \mathbf{x}$.

Both algorithms require the implementation of perfect sampling (Propp and Wilson, 1996) from the likelihood function as the density function of $\mathbf{x}$ is also intractable. However, for some complex models, such as the Ising model on a large grid, perfect sampling is too costly.

Several papers propose methods for cases where it is infeasible to sample from the assumed data generating process. Caimo and Friel (2011) modify the original exchange algorithm, where the simulation of $\mathbf{x}$ from $f(\mathbf{x}|\boldsymbol{\theta}')$ is carried out by MCMC. A similar idea is explored by Liang (2010), named the double Metropolis-Hastings sampler, where the MH algorithm is used twice; once for generating $\boldsymbol{\theta}$ and once for generating the auxiliary variable $\mathbf{x}$. However, the algorithm is asymptotically inexact as the detailed balance condition does not hold for updating $\boldsymbol{\theta}$, unless the number of draws for $\mathbf{x}$ is infinite. To overcome the issue, Liang et al. (2016) propose the adaptive exchange algorithm, which can be an MCMC extension of the exchange algorithm. In this algorithm, two chains are run simultaneously. The auxiliary variables are generated via importance sampling from one Markov chain and the other chain adopts the exchange algorithm to generate posterior samples. The algorithm is asymptotically exact, but there are memory issues associated with the large number of intermediate variables needed to be stored within each iteration.

To conclude, auxiliary variable approaches avoid evaluation of the intractable constant directly. However, perfect sampling from the likelihood function is required. This can be relaxed, but this results either in inexact inference, or a computationally costly algorithm.

Approximate likelihood approaches: These approaches approximate the normalising constant, or its reciprocal, and substitute the approximation into the MH acceptance ratio. In this thesis, we consider the PM framework which replaces the likelihood function with an unbiased estimator. A representative method is the so-called Russian roulette (RR) method.

The RR method first appears in the physics literature (Carter and Cashwell, 1975) as a sampling technique for particle transport problems. Lyne et al. (2015) apply it to tackle the doubly intractable problem. The reciprocal of the normalising constant is expressed as a geometric series, and RR determines a truncation rule of the infinite sum. Specifically, $Z^{-1}(\boldsymbol{\theta})$ is estimated by

$$\widehat{Z^{-1}}(\boldsymbol{\theta}) = \frac{c(\boldsymbol{\theta})}{\tilde{Z}(\boldsymbol{\theta})}\left[1 + \sum_{n=1}^{\infty}\prod_{i=1}^{n}\widehat{k}_i(\boldsymbol{\theta})\right], \quad \text{with } \widehat{k}_i(\boldsymbol{\theta}) = 1 - c(\boldsymbol{\theta})\frac{\widehat{Z}_i(\boldsymbol{\theta})}{\tilde{Z}(\boldsymbol{\theta})}, \tag{2.3}$$

where $c(\boldsymbol{\theta})$ ensures $|1 - c(\boldsymbol{\theta})Z(\boldsymbol{\theta})/\tilde{Z}(\boldsymbol{\theta})| < 1$, and $\widehat{Z}_i(\boldsymbol{\theta})$ $(i = 1, \ldots, n)$ is an unbiased estimator for $Z(\boldsymbol{\theta})$. The ideal value of $\tilde{Z}(\boldsymbol{\theta})$ is the upper bound on $\widehat{Z}(\theta)$, i.e., $\sup_{i\geq 1}\widehat{Z}_i(\boldsymbol{\theta})$.

RR is then introduced to find a stopping rule for the summation (so that there is only a finite sum for n) such that $\widehat{Z^{-1}}(\boldsymbol{\theta})$ is an unbiased estimator of the likelihood $Z^{-1}(\boldsymbol{\theta})$.

Following the notation in Lyne et al. (2015), we rewrite the infinite sum in (2.3) (including the constant 1) as

$$S = \sum_{i=0}^{\infty} \phi_i(\boldsymbol{\theta}),$$

and define

$$S_k = \phi_0(\boldsymbol{\theta}) + \sum_{n=1}^{k} \frac{\phi_n(\boldsymbol{\theta})}{p_n},$$

where $S_0 = \phi_0(\boldsymbol{\theta})$ and $p_0 = 1$. The RR process is based on simulation of the stopping time τ according to the probabilities $\Pr(\tau \geq n) > 0$ for all $n \geq 0$. The stopping time is often chosen as

$$\tau = \inf\{k \geq 1 : U_k \geq q_k\},$$

where $U_k \overset{\text{iid}}{\sim} \text{Uniform}(0,1), q_k \in (0,1]$ and $p_n = \prod_{j=1}^{n-1} q_j$.

The RR estimator of S is $\widehat{S} = S_{\tau-1}$ in this case. It is easy to verify that $E(\widehat{S}) = S$. Lyne et al. (2015) further suggest $\phi_j(\boldsymbol{\theta}) = \phi^j(\boldsymbol{\theta})$ and $q_k = q = \phi(\boldsymbol{\theta})$ to ensure the variance of the RR estimator is finite. The RR estimator is then simplified as

$$S_{\tau-1} = 1 + \sum_{n=1}^{\tau-1} \frac{\prod_{i=1}^{n} \widehat{k}_i(\boldsymbol{\theta})}{\prod_{k=1}^{n-1} q_k}, \text{where } q_k = \prod_{i=1}^{k} \widehat{k}_i(\boldsymbol{\theta}).$$

The RR algorithm offers asymptotically exact inference and can be applied to a wider range of problems than the auxiliary variable approaches as perfect sampling from the likelihood function is not required. However, RR has two drawbacks. First, it is usually difficult to obtain a tight upper bound $\tilde{Z}(\boldsymbol{\theta})$ for $\widehat{Z}(\boldsymbol{\theta})$. If the upper bound is loose, then the convergence of the geometric series is slow and prevents the algorithm from running efficiently (in this case, $\widehat{Z}_i(\boldsymbol{\theta})/\tilde{Z}(\boldsymbol{\theta})$ is small and $\widehat{k}_i(\boldsymbol{\theta}) \approx 1$, leading to a large value of τ). Conversely, if the specified upper bound is not an actual upper bound, the estimator can produce negative values, which can be corrected by weighting the expectation as with the signed PMMH algorithm (see Section 2.2). However, the variance of the posterior estimate becomes large when too many negative estimates are observed. Second, the variability of the likelihood estimator cannot be controlled explicitly. Consequently, a significant amount of tuning is required for the algorithm to perform well.

Lyne et al. (2015) also propose an exponential auxiliary variable within the PM approach and claim a faster convergence rate of the corresponding geometric series. Suppose $\nu \sim \text{Expon}(-Z(\boldsymbol{\theta}))$, then the joint distribution of ν and $\boldsymbol{\theta}$ is

$$\begin{aligned}\pi(\boldsymbol{\theta}, \nu|\mathbf{y}) &= \pi(\boldsymbol{\theta})Z(\boldsymbol{\theta})\exp(-\nu Z(\boldsymbol{\theta}))\frac{f(\mathbf{y}|\boldsymbol{\theta})}{Z(\boldsymbol{\theta})}\frac{1}{p(\mathbf{y})} \\ &= \pi(\boldsymbol{\theta})\exp(-\nu Z(\boldsymbol{\theta}))f(\mathbf{y}|\boldsymbol{\theta})\frac{1}{p(\mathbf{y})} \\ &= \pi(\boldsymbol{\theta})\left[1 + \sum_{n=1}^{\infty}\frac{(-\nu Z(\boldsymbol{\theta}))^n}{n!}\right]f(\mathbf{y}|\boldsymbol{\theta})\frac{1}{p(\mathbf{y})}. \qquad (2.4)\end{aligned}$$

The series including the infinite sum in (2.4) can be estimated similarly to (2.3). However, the introduction of the auxiliary variable ν does not solve the issues as mentioned above. Chapter 4 adopts the same approach, but uses the block-Poisson estimator (Quiroz et al., 2021) to overcome the problem encountered by the RR method.

Other approaches besides the PM methods approximate likelihood functions with different targets. Atchadé et al. (2013) construct an adaptive MCMC method, which can be thought of as a Bayesian version of the MCMC-MLE approach (Geyer, 1991). The constant $Z(\boldsymbol{\theta})$ is approximated through importance sampling using the entire sample path. The proposed algorithm is not Markovian, but the marginal distribution of $\boldsymbol{\theta}$ generally converges to the posterior density. Similarly to the adaptive exchange algorithm, this algorithm requires many particles to cover the parameter space.

Park and Haran (2020) use a Gaussian process-based approximation to estimate the normalising constant, and provide a theoretical justification. The algorithm overcomes the expense in obtaining the normalising constant in other algorithms. The algorithm is effective for low-dimensional parameter spaces, which may involve high-dimensional data sets. The parameter dimensions considered in the examples of this paper are between 1 and 4. For a high-dimensional parameter space, the algorithm might be computationally impractical.

Stoehr et al. (2019) propose a noisy Hamiltonian Monte Carlo (HMC) algorithm for intractable distributions. The method relies on Hamiltonian dynamics to propose a large transition across the parameter space. Monte Carlo estimates are introduced to overcome the intractability induced in the construction of the transition and the computation of the MH acceptance ratio.

2.5 Causal inference and the Bayesian additive regression tree (BART)

2.5.1 Causal inference

The central task for many scientific questions is to understand causal relationships. It often requires some assumptions to show that two correlated variables have a cause-and-effect relationship. For example, causes always occur before effects; all individuals have a positive probability of receiving the available treatments. We will discuss these assumptions further later on. Causal inference investigates the methods of how, and to what extent, causality can be inferred from the data. For example, causal inference helps to answer the following question: "I have a headache. Will it help if I choose to take an aspirin?". It is worth emphasising that causal inference investigates the effect of causes (causal effects) but does not identify the causes of the effect, which answers the question: "My headache is gone. Why is it gone?".

There are multiple ways to conduct causal inference, such as the potential outcome approach (Rubin, 1974), the graphical approach (Pearl, 2009) and the decision-analytical approach (Dawid, 2000). In this thesis, we use the most widely adopted framework, the potential outcome approach. Statistical methods using this approach are intensively investigated in the literature. Depending on different identification strategies (research designs), various methods can be applied to model the same data. The next section introduces key concepts in the potential outcome approach, the identification strategy "selection on the observables," and two popular associated statistical methods.

The potential outcome approach

The approach is formulated in Rubin (1974) who claims that Jerzy Neyman first used the language in randomised experiments (Rubin, 1990). Randomised experiments are regarded as the gold standard for investigating causal effects, where subjects are randomised into either the treatment or the control group. The differences in outcomes can be attributed to the treatment as the the responses of each group are comparable to those of the other due to the randomisation. Rubin points out that randomised experiments are impractical for many questions. For example, the cost of a randomised experiment may be prohibitively expensive. There may be ethical reasons for not conducting such experiments. The results can be delayed many years for some long term studies. Noting this, Rubin (1974) proposes the potential outcome framework for observational studies when no randomisation is involved in the treatment assignment.

Let $Z_i \in \{0, 1\}$ be a binary treatment indicator with 0 indicating that the ith unit is in the control group and 1 that it is in the treatment group. Denote $\mathbf{X}_i$ and Y_i as the pre-treatment covariates and the response variable respectively. The potential outcomes are defined as $Y_i(0)$ and $Y_i(1)$ to represent the outcome of each group. The effect of cause Z on the ith unit is the difference between $Y_i(1)$ and $Y_i(0)$. Equivalently, it states that the treatment Z_i causes the effect $Y_i(1) - Y_i(0)$. The observed continuous outcome variable Y_i is therefore defined as

$$\begin{aligned} Y_i &= Y_i(1), \quad \text{if } Z_i = 1, \\ Y_i &= Y_i(0), \quad \text{if } Z_i = 0. \end{aligned}$$

The paired definition is often written compactly as

$$Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0).$$

The fundamental problem of causal inference (Holland, 1986) is that at most one of $Y_i(0)$ and $Y_i(1)$ is observed for each unit i. Therefore, it is impossible to obtain the effect of Z on the ith unit directly, which implies that it is implausible to infer individual-level casual effects. However, such effects can be aggregated into the average treatment effect (ATE) for a subset of the population. The definition of ATE in the population as a whole is

$$\tau_{ATE} := E(Y(1) - Y(0)).$$

Likewise, the average treatment effect on the treated units (ATT) is defined as

$$\tau_{ATT} := E(Y(1) - Y(0)|Z = 1),$$

and the conditional treatment effect (CATE) as

$$\tau_{CATE}(\mathbf{x}) := E(Y(1) - Y(0)|\mathbf{X} = \mathbf{x}).$$

“Selection on the observables”

To estimate ATE, ATT or CATE, identification strategies are often required in causal inference: research designs which aim to solve causal inference identification problems. Angrist and Pischke (2008); Morgan and Winship (2015) provide in-depth reviews for identification strategies based on different type of studies. The identification strategy “the selection on observables” (Barnow et al., 1980) is adopted in Chapter 5. Under such a strategy, it is asserted that the treatment assignment is random conditioning on a set of observed covariates, which means that the treatment is conditionally independent of potential outcomes given the observed covariates. Two commonly adopted assumptions under this strategy are the stable unit treatment value assumption (SUTVA) and strong ignorability.

SUTVA (Rubin, 1978) states that the potential outcomes of individuals are unaffected by the changes of treatments that all other individuals receive. Strong ignorability is built on the critical “ignorable” property (Rubin, 1978). If the treatment status is independent of potential outcomes, the treatment assignment mechanism $\Pr(Z|\mathbf{X}, Y(0), Y(1))$ is said to be “ignorable”. The “ignorable” assumption is written as

$$(Y(0), Y(1)) \perp Z.$$

All the randomised experiments share such a property. Strong ignorability (Rosenbaum and Rubin, 1983) is a stronger and theoretically more convenient assumption. The treatment assignment is strongly ignorable given the covariates $\mathbf{X}$ if

$$(Y(0), Y(1)) \perp Z|\mathbf{X}, \quad 0 < \Pr(Z = 1|\mathbf{X}) < 1.$$

The first condition states the conditional independence of the potential outcomes $Y(1)$, $Y(0)$ and the treatment Z given covariates $\mathbf{X}$. The second condition says that every unit has a positive chance of being assigned to the treatment or the control group.

If the SUTVA and strong ignorability are assumptions are satisfied, then

$$E(Y|Z = 1, \mathbf{X} = \mathbf{x}) = E(Y(1)|\mathbf{X} = \mathbf{x}),$$
$$E(Y|Z = 0, \mathbf{X} = \mathbf{x}) = E(Y(0)|\mathbf{X} = \mathbf{x}).$$

Hence, CATE can be estimated from the observed data as a problem of finding the conditional expectation. For ATE and ATT, they can be derived by averaging the estimates of CATE across relevant observations. Alternative approaches include regression and matching methods; see the section below for an introduction.

Regression and matching methods

Many statistical methods are available to estimate causal effects provided the SUTVA and strong ignorability assumptions hold; see Imbens and Rubin (2015); Morgan and Winship (2015) for textbook-level introductions. Here we introduce the two most popular methods: regression and matching. The proposed model in Chapter 5 uses a hybrid of both methods, which is a regression method with adaptive matching involved.

Regression analysis is widely used to adjust for covariates under the identification strategy "selection on observables", which relies on consistent estimates of the response surfaces $E(Y(1)|\mathbf{X} = \mathbf{x})$ and $E(Y(0)|\mathbf{X} = \mathbf{x})$. The models can have restricted functional forms such as linear regressions or logistic regressions. For example, suppose that the treatment effect is constant for all units i. Consider a linear model for estimating the treatment effect,

$$Y_i = \mathbf{X}_i\boldsymbol{\alpha} + \beta Z_i + \epsilon_i,$$

with $\epsilon_i \perp \mathbf{X}_i$. The strong ignorability assumption asserts that Z_i is independent of ϵ_i conditioning on $\mathbf{X}_i$. The coefficient β is then interpreted as an estimate of the treatment effect. However, the strong ignorability assumption does not imply a linear relationship between the response and the covariates. As critiqued in Freedman (2006), functional form misspecifications may bias the estimate when the response surface is nonlinear. Nonparametric regression methods, on the other hand, are more flexible for fitting complex response surfaces. The use of splines and kernel methods is investigated in Hainmueller and Hazlett (2014); Zhou et al. (2019); see Imbens (2004) for a review. Hill (2011) estimates CATE by the Bayesian additive regression tree (BART) (Chipman et al., 2010), a nonparametric Bayesian regression method which is introduced in the next section.

Matching is another popular technique in causal inference. It represents an intuitive idea which makes the objects from the treatment group comparable to those from the control group in terms of the observed covariates or a function of the covariates. A variety of matching methods are developed in the literature on causal inference; see Stuart (2010) for a detailed review. Since conditioning on all the covariates is impractical in case of a high-dimensional $\mathbf{X}$, an important quantity often used in matching is the propensity score $\Pr(Z = 1|\mathbf{X} = \mathbf{x})$ (Rosenbaum and Rubin, 1983), which is defined as the conditional probability of receiving the treatment given the observed covariates $\mathbf{x}$. The propensity score serves as a balancing score, which says that conditioning on the propensity score, the distribution of the observed covariates is similar between treated and untreated subjects (Austin, 2011). Furthermore, Rosenbaum and Rubin (1983) show that if potential outcomes are independent of treatment conditioning on the covariates $\mathbf{X}$, then they are also independent of treatment conditioning on a balancing score, i.e., $(Y(0), Y(1)) \perp Z| \Pr(Z = 1|\mathbf{X})$. As the true propensity score is usually unknown in practice, its estimate is used prior to the matching procedure. Various practical guides for implementing the propensity score matching discussed in Caliendo and Kopeinig (2008); Heinrich et al. (2010). Abadie and Imbens (2016) derive the large sample distribution of propensity score matching estimators, where the propensity score itself is estimated first. The estimated propensity score is also used in the model proposed in Chapter 5.

2.5.2 BART

The Bayesian additive regression tree (BART) is first introduced in Chipman et al. (2007, 2010) as an alternative regression method. Unlike most regression models which make stringent assumptions on the functional form of the conditional expectation, such as $E(Y|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$, BART avoids such parametric assumptions. It combines a machine learning algorithm with a likelihood-based inference framework to flexibly model a highly non-linear response surface. Below, we first introduce the regression tree, followed by the BART model and its prior. We then represent the iterative backfitting MCMC algorithm and recent developments for the BART model.

Regression tree

The regression tree recursively partitions covariate space into subgroups and is the building block of BART. Figure 2.2 provides an example of a binary tree. The left panel displays the splitting rule $x_1 < 0.9$ followed with $x_2 < 0.4$ in the interior nodes (the boxes). The observations are assigned to different leafs (the circles) by passing them along the tree. The leafs are also termed terminal nodes to differentiate from interior nodes. Each leaf is associated with a parameter $\mu_{hi}, i = 1, 2, 3$, representing the predicted value. The right panel of Figure 2.2 shows the corresponding partitioned sample space. If the tree is bushy with more interior nodes, the partition of the space is expected to be finer. Observations with similar covariates are likely to be assigned to the same subgroup. The criterion for finding appropriate splitting rules are extensively studied in the machine learning literature; see Rokach and Maimon (2005) for a review.

The model

The original BART formulation is as a sum-of-trees model,

$$Y = \sum_{j=1}^{m} g(\mathbf{x}; T_j, M_j) + \epsilon, \quad \epsilon \sim N(0, \sigma^2).$$

Each (T_j, M_j) is a single binary tree with T_j being the collection of splitting variables and decision rules and $M_j = \{\mu_{1j}, \ldots, \mu_{b_j j}\}$ being the parameters attached to the terminal

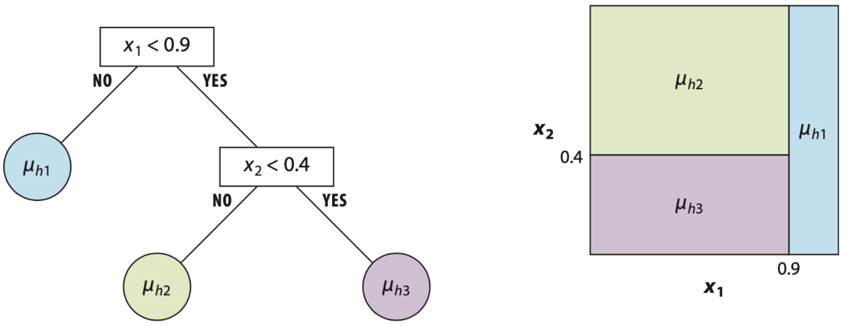


Figure 2.2: An example binary tree with the corresponding partition of the sample space. The figure is taken from Hill et al. (2020).

nodes of tree j. The number of elements in T_j and M_j may change through the MCMC iterates.

The model prediction is the sum over the predictions provided by the m binary trees, which can be regarded as an ensemble learning method (Dietterich et al., 2002). Instead of finding a tree with the best structure to explain the data, the sum-of-trees model makes use of the predictive power of each tree. The model is composed of weak learners (individual regression trees) and combines them to form a stronger learner. The first tree may explain a small amount of variation in the response with part of the remaining variation explained by the next tree. The process is performed m times in total. To avoid overfitting, a regularisation prior is placed on the tree structure. Such prior constrains each tree to be small and each element of M_j is shrunk toward zero.

The BART prior

The default prior (Chipman et al., 2010) is imposed on σ^2 and the pair (T_j, M_j) jointly. Assuming that the trees are independent of each other, the prior can be factorised as,

$$p((T_1, M_1), \cdots, (T_m, M_m), \sigma^2) = \left[\prod_{j=1}^{m} p(T_j, M_j)\right] p(\sigma^2) = \left[\prod_{j=1}^{m} p(M_j|T_j)p(T_j)\right] p(\sigma^2),$$

where $p(M_j|T_j) = \prod_{i=1}^{b_j} p(\mu_{ij}|T_j)$, $\mu_{ij} \in M_j = \{\mu_{1j}, \ldots, \mu_{b_j j}\}$.

The prior is composed of three components, (1) the prior on σ^2, $p(\sigma^2)$ (2) the prior on the tree structure T_j, $p(T_j)$ and (3) the prior on terminal parameters conditioning on the tree structure, $p(M_j|T_j)$. Regularisation on the tree structure is advocated to prevent the effects of a single tree from being too influential. Without the regularisation, a large number of trees tend to overfit the data which leads to poor predictions. It is recommended to use the default specifications in Chipman et al. (2007), which appear to be effective.

The prior on σ^2 is an inverse chi-squared distribution,

$$\sigma^2 \sim \frac{\nu\lambda}{\chi^2_\nu}.$$

The hyperparameters ν and λ are obtained through a rough estimate for σ^2, $\widehat{\sigma}^2$. The degree of freedom ν can be an integer number from 3 to 10. The quantile q is picked as 0.75 (conservative), 0.90 (default) or 0.99 (aggressive) and λ is set to satisfy $\Pr(\sigma^2 < \widehat{\sigma}^2) = q$. The default choice is $(\nu, q) = (3, 0.90)$.

The specification in Chipman et al. (1998) is used for the prior on T_j. Let d be the depth of an interior node. The root node of a tree is at depth 0. The prior on a nonterminal node is

$$\Pr(\text{node is an interior node}) = \alpha(1+d)^{-\beta}, \quad \alpha \in (0,1), \beta \in [0,\infty).$$

The default value for (α, β) is $(0.95, 2)$. Small trees are favoured under such a specification. For example, a tree with ≥ 5 terminal nodes is assigned with a probability 0.03.

The last component is the prior on the parameters at terminal nodes. Let μ_{ij} be the parameter attached to the ith terminal node of the jth tree. Conditioning on the tree structure T_j, a normal distribution is imposed on μ_{ij},

$$\mu_{ij} \sim N(0, \sigma^2_\mu).$$

The observations $\mathbf{y}$ are rescaled, with $y_{\min} = -0.5$ and $y_{\max} = 0.5$. The scalar σ_μ is selected to ensure $k\sqrt{m}\sigma_\mu = 0.5$ with m being the number of trees of the BART model and 0.5 being within $k\sigma_u$ standard deviations of zero. For the choice of k, the values between 1 to 3 yield good results. The number of trees typically ranges from 50 to 200.

The algorithm

The BART model is implemented through an iterative backfitting MCMC algorithm. Let T_{-j} be the set of the structures of all trees except tree j and M_{-j} the set of all parameters excluding those of tree j. The backfitting algorithm can be regarded as a Gibbs sampler which draws (T_j, M_j) conditioning on $T_{-j}, M_{-j}, \sigma^2, \mathbf{y}$. The process is repeated m times by looping through $j = 1, \dots, m$, i.e.,

$$\begin{aligned}
(T_1, M_1)|T_{-1}, M_{-1}, \sigma^2, \mathbf{y}&,\\
(T_2, M_2)|T_{-2}, M_{-2}, \sigma^2, \mathbf{y}&,\\
\vdots&\\
(T_m, M_m)|T_{-m}, M_{-m}, \sigma^2, \mathbf{y}&,\\
\sigma^2|T_1, \dots, T_m, M_1, \dots, M_m, \mathbf{y}&.
\end{aligned}$$

The last step draws samples from the full conditional distribution of σ^2. The effect of T_{-j}, M_{-j} and $\mathbf{y}$ can be replaced by the residuals $R_{-j} = \mathbf{y} - \sum_{i \neq j} g(\mathbf{x}; T_i, M_i)$. The last step is straightforward to implement due to the conjugate prior on σ^2, which results in an inverse chi-squared distribution. For the first m steps, the sampling consists of two stages, drawing from $(T_j|R_{-j}, \sigma^2)$ and $(M_j|T_j, R_{-j}, \sigma^2)$. Updating the terminal parameters M_j conditional on T_j, R_{-j}, σ^2 is also straightforward due to the conjugate normal prior on the elements of M_j.

Sampling T_j conditioning on R_{-j}, σ^2 is completed by the MH algorithm after integrating out M_j. Chipman et al. (1998) propose a new tree based on the current one using one of the four moves: growing, pruning, swapping and changing. Pratola et al. (2014) advise using growing and pruning only as the fits are close to those obtained by the four moves. Kapelner and Bleich (2013) suggest using three alterations, which are growing, pruning and changing. The “swapping” move is excluded because of the complexity of the computational bookkeeping. In implementing the model proposed in Chapter 5, we follow the approach in Kapelner and Bleich (2013), using the derivation of the conditional distributions in Appendix C.1.

Development on BART

The original BART assumes that the response variable is continuous. In Chipman et al. (2007, 2010), the response variable is extended to a binary classifier by the transformation $\Phi(.)$ on fitted values, where $\Phi(.)$ is a standard normal cumulative density function. Murray (2021) extends the BART model to Poisson and binomial count models. For positive observations, Linero et al. (2020) describe probit-based hurdle models and Gamma hurdle models with the logarithm of the response modelled by BART. Another BART extension allows heteroscadastic regression, where the variance of the error depends on covariates (Pratola et al., 2017).

The default BART prior may break down for data with a high-dimensional covariate space. In this setting, Linero (2018) modifies the default prior by introducing a sparsity-inducing Dirichlet prior to filter out most variables in the model. Linero and Yang (2018) propose the soft BART (SBART) model, where a soft threshold is advocated in the splitting rules to introduce smoothness in regression trees. Under SBART, the decision of going to the next leaf is made randomly rather than deterministically.

Ročková and van der Pas (2020) develop theory investigating why and when BART does not overfit. Theoretical results are established on the optimal convergence rate of the BART posterior concentration up to a log factor. Linero and Yang (2018) study such a rate for SBART. Liu et al. (2021) propose ABC Bayesian forests for variable selection problems, with theoretical results obtained for a modified BART prior.

Various software packages are available for efficient BART implementations on real applications, such as the `R` packages `bartMachine` (Kapelner and Bleich, 2013), `dbarts`(Dorie, 2021), and `BART` (Sparapani et al., 2021). Pratola et al. (2014) implement parallel computation on a simplified version of BART to obtain an efficient algorithm.

The development on BART is more substantial than what we have discussed above; see Hill et al. (2020) for a review on BART models with recent developments. In the next section, we focus on an important application of BART, causal inference.

2.5.3 BART for causal inference

Recall the estimation of CATE, which is defined as

$$\tau_{CATE}(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x}, Z = 1) - E(Y|\mathbf{X} = \mathbf{x}, Z = 0),$$

provided that SUTVA and strong ignorability are satisfied. This problem can be reformulated as finding the conditional expectation of Y given $\mathbf{x}$ and z. Hill (2011) first uses BART as a tool to estimate heterogeneous treatment effects for observational studies with continuous outcomes and binary treatments. Green and Kern (2012) adopt a similar approach for survey experiments.

Recall that the BART model is written as (including the covariates $\mathbf{x}$ and treatment indicator z),

$$Y_i = f(\mathbf{x}_i, z_i) + \epsilon_i,$$

where $f(\mathbf{x}_i, z_i)$ is a sum over a set of trees.

The prediction of the BART model conditional on $\mathbf{x}_i, z_i$ is $E(Y_i|\mathbf{x}_i, z_i) = f(\mathbf{x}_i, z_i)$. To obtain the counterfactual, which is $E(Y_i, |\mathbf{x}_i, 1-z_i)$, the original data is imputed by flipping the treatment status from z_i to $1 - z_i$.

Hahn et al. (2020) take a different approach by assuming

$$Y_i = \mu(\mathbf{x}_i) + \tau(\tilde{\mathbf{x}}_i)z_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2),$$

where $\mu(\mathbf{x}_i)$ and $\tau(\tilde{\mathbf{x}}_i)$ are modelled by separate sets of trees. The variables $\tilde{\mathbf{x}}_i$ can be the same as $\mathbf{x}$ or a subset of $\mathbf{x}$. The method is called the Bayesian causal forest (BCF). Unlike the approach in Hill (2011), $\tau_{CATE}(\mathbf{x})$ is modelled explicitly in this method. A similar approach is applied in Zeldow et al. (2019), where the treatment effect $\tau(\tilde{\mathbf{x}}_i)$ is replaced with a linear function of $\mathbf{x}$. As pointed out in Hahn et al. (2020), the choice of treatment level affects posterior inference. For example, if two active treatments are compared, it is inappropriate to code z_i as an element in $\{0, 1\}$ or $\{\pm 0.5\}$ because there is no reference treatment level. Another example is that when there is a big difference in the marginal variance of Y for the treated and control group, inferences based on the CATE estimator are impacted. The BCF method treats the coding of z as a variable, with the model described as:

$$\begin{aligned}
&Y_i = \mu(\mathbf{x}_i) + \tilde{\tau}(\tilde{\mathbf{x}}_i)b_{z_i} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2), \\
&b_0 \sim N(0, 0.5), \quad b_1 \sim N(0, 0.5).
\end{aligned}$$

Then, the CATE can be estimated by

$$\tau(\mathbf{x}_i) = (b_1 - b_0)\tilde{\tau}(\tilde{\mathbf{x}}_i).$$

Chapter 5 proposes our solution of a BART implementation for causal inference and compares it with Hill (2011) and Hahn et al. (2020) in the simulated data sets of the 2016 ACIC challenge (Dorie et al., 2019).

Chapter 3

Analysing symbolic data by pseudo-marginal methods

3.1 Introduction

Large data sets are becoming the norm in science today. The traditional statistical approaches usually involve evaluating the likelihood function for all observations, which may be computationally infeasible. Symbolic data analysis (SDA), on the other hand, aggregates many individual-level observations into "symbols", which provide a summary of the underlying data, giving summary information about the distribution of subsets of the data. Inference is then carried out using the symbols instead of the original data. The procedure has favourable features in terms of computational efficiency for large data sets as SDA uses manageable sized data rather than the full data set, providing a valuable framework for statistical analysis in the big data era.

Symbolic data can be multi-valued, interval-valued or model-valued variables (Billard, 2006), for example, random intervals or random histograms. In this chapter, we use the term symbols indistinguishably with symbolic data. Symbols are similar to summary statistics, but may contain more information depending on the associated construction method. Many popular statistical analyses are extended to SDA, including regression analysis (Billard and Diday, 2000, 2002), time series (Teles and Brito, 2015), clustering (Brito, 2014), etc. However, these techniques are often based on assumptions that the symbols follow uniform distributions. For example, if data $X_1, \ldots, X_n$ are summarised by a random interval $[a, b]$ which contains all the X_i values, then it is often assumed that the distribution of data within $[a, b]$ is uniform. However, the assumption of uniformity is inappropriate in many real applications. It consequently affects statistical inference, as the variability within the interval is specified wrongly.

In contrast to other approaches of treating symbols as observed data, Beranger et al. (2018) propose a likelihood-based framework for constructing a symbolic likelihood function which relaxes the uniformity assumption. The approach incorporates the process of constructing the symbol from the full data set, as part of the analytical process. This allows the fitting of standard statistical models for classical data $\{X_i\}$, only given the symbolic summary information (e.g. $[a, b]$). Evaluating the corresponding likelihood, the so-called symbolic likelihood, requires computing integrals over the space of the unobserved data X_i. Most work on these models has to consider the case where the integrals have a closed form solution. However, this has restricted these techniques from being applied to more general settings, for which the integrals may be intractable. Such intractability prevents the application of popular algorithms to perform Bayesian inference on the model. We propose a pseudo-marginal MCMC framework to carry out Bayesian inference for symbolic data; see Section 3.2.1 for details.

As part of their symbol construction, Beranger et al. (2018) apply their approach to perform inference for the parameters of bivariate normal distributions and skewed normal distributions. There are some limitations with the symbol construction technique they propose, which does not make full use of the available information. As a result, it fails to make reliable inference for the correlation parameter for a large data set with low or moderate correlation. This weakness is noted in the results of a simulation study in Beranger et al. (2018). Another downside of their method is that the applications are limited to low-dimensional problems with tractable likelihood functions only. For modelling histogram-valued random variables, the number of bins to construct the histograms increases exponentially with the number of dimensions (similarly to non-parametric density estimation), which limits direct applications for high-dimensional problems. In contrast, the proposed method in this chapter overcomes the limitation on the data dimension and relaxes the requirement of the tractable likelihood function. We also resolve the problem of correlation underestimation through a novel symbolic construction method.

Various applications are investigated in the literature using the likelihood-based approach in Beranger et al. (2018). Lin et al. (2017) estimate global species richness by a Bayesian hierarchical approach. Whitaker et al. (2020) extend the framework to higher dimensions using composite likelihoods and investigate climate extremes. Rahman et al. (2020) investigate theoretical properties of the likelihood and show the consistency of the estimators in the context of modelling internet network traffic volumes.

In this chapter, we propose a novel symbol construction method for likelihood-based inference in SDA. Following the construction method of Beranger et al. (2018), we extend their framework to solve high-dimensional problems by considering a single quantile-based (or min-max) interval instead of histogram-valued intervals, which is computationally faster. To cope with the intractable likelihood function, we use the pseudo-marginal (PM) framework (Andrieu and Roberts, 2009) to conduct Bayesian analysis, where an unbiased likelihood estimator replaces the intractable likelihood. An exact and an approximate method are proposed to get a (nearly) unbiased estimator of the likelihood function.

The chapter is organised as follows. Section 3.2 introduces the symbolic likelihood-based framework and our extension. The exact and the approximate methods for likelihood estimation are then derived. The signed block pseudo-marginal Metropolis-Hastings (*signed block PMMH*) algorithm completes the chapter. Section 3.3 has three parts. The first part compares the proposed method with Beranger et al. (2018). The second part compares our exact and approximate methods. Third, the approximate method is applied to a factor model and compared with the full data result. Section 3.4 demonstrates the method on an empirical data set using a Bayesian linear regression.

3.2 Methodology

3.2.1 Symbolic data likelihood

Beranger et al. (2018) propose a likelihood-based approach for SDA which incorporates the data generating process of the random variables that underlie the symbols. We first explain how the symbol is constructed in Beranger et al. (2018), then state the redefinition of the symbolic likelihood based on our proposal. Finally, we discuss the differences between the two approaches.

Denote $\mathbf{X} = (X_1, \ldots, X_d)$ as a d-dimensional random vector with density $g_{\mathbf{X}}(\cdot; \boldsymbol{\theta})$ in the domain $\mathcal{X} \subseteq \mathbb{R}^d$. The observed values $\mathbf{x}$ of $\mathbf{X}$ are aggregated into a symbol s according to some known function $f_{S|\mathbf{X}=\mathbf{x}}(s|\mathbf{x}, \phi)$ parameterised by ϕ, which determines how the symbol s is generated conditional on $\mathbf{x}$. The function $f_{S|\mathbf{X}=\mathbf{x}}(s|\mathbf{x}, \phi)$ is a conditional density of s given $\mathbf{x}$ and ϕ. The parameter ϕ is implicitly determined by the construction, and is not modelled. The symbolic likelihood has the form

$$L(s; \boldsymbol{\theta}, \phi) \propto \int_{\mathcal{X}} f_{S|\mathbf{X}=\mathbf{x}}(s|\mathbf{x}, \phi) g_{\mathbf{X}}(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{x}. \tag{3.1}$$

The equation says that $L(s; \boldsymbol{\theta}, \phi)$ is the marginal distribution of s obtained by integrating over the unobserved $\mathbf{x}$, which means all the possible data sets that may have generated s are taken into account.

In this chapter, we write the symbol s as $s = \{\mathbf{x}_b, \mathbf{x}_o, n_b, n_o, n_t\}$, where $\mathbf{x}_b \in \mathbb{R}^d$ refers to the location of those classical data points used to construct a hyper-rectangle, which lie on the boundary of the hyper-rectangle constructed by themselves, and $\mathbf{x}_o \in \mathbb{R}^d$ refers to the observations lying outside the hyper-rectangle. The total number of observations is n_t and the number of boundary ($\mathbf{x}_b$) and external ($\mathbf{x}_o$) data points are n_b and n_o respectively. To construct the symbol, we first select certain points to form the boundary of the hyper-rectangle. After the hyper-rectangle is constructed, we can easily determine the observations outside the boundary of the hyper-rectangle. Those data points within the hyper-rectangle are discarded. We propose two ways of obtaining the boundary points.

The first is to select the points which lie on the marginal minimum or maximum over all the other points in at least one dimension. The second method generalises the first one by choosing those observations containing the q and $1-q$ quantiles of at least one dimension. When $q = 0$, then the first method results in a standard min-max hyper-rectangle with no observations outside, i.e., $\mathbf{x}_o = \emptyset, n_o = 0$. By having q closer to 0.5, we reduce the symbolic likelihood to the classic one as no observations exist inside the hyper-rectangle (unless some points happen to lie exactly on the d-dimensional median).

Figure 3.1 illustrates the two types of symbols constructed by the methods described above. In Figure 3.1, the left panel shows the original data set with $n_t = 1{,}000$ data points. The centre panel constructs a min-max ($q = 0$) hyper-rectangle using the 4 extreme data points as the boundary, and discarding the 996 observations within. The right panel constructs the hyper-rectangle from the marginal $q = 0.005$ quantiles, retaining 17 external data points and discarding 979 points within the hyper-rectangle. Clearly, the symbolic representation of the data significantly reduces the memory storage required. By having a sufficiently small value of q, the quantile-based interval contains a few additional observations over the min-max hyper-rectangle.

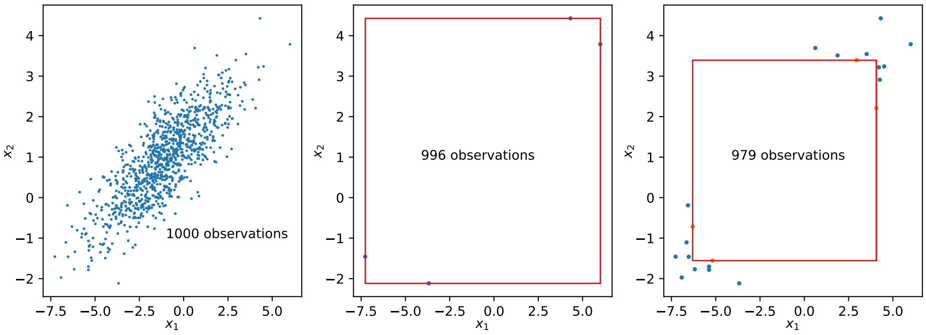


Figure 3.1: Symbol construction. Left panel: Original data set of 1,000 independent observations, which are generated from a bivariate normal distribution with $\boldsymbol{\mu} = (-1, 1), \sigma_1 = 2, \sigma_2 = 1, \rho_{12} = 0.8$. Middle panel: Min-max interval with 4 points on the boundary, discarding the 996 points within the rectangle. Right panel: Quantile-based interval ($q = 0.005$ (0.5th quantile)) with 4 points on the boundary, 17 points outside and the remaining points (979 points) inside the rectangle.

Based on (3.1), and assuming that the underlying data are iid observations from $g_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\theta})$, the symbolic likelihood of the observed symbol s is defined as

$$L_f(s;\boldsymbol{\theta}) \propto \left[\int_S g_{\mathbf{x}}(\mathbf{x};\boldsymbol{\theta})d\mathbf{x}\right]^{n_t-n_b-n_o} \times \mathcal{L}(\mathbf{x}_b;\boldsymbol{\theta}) \times \mathcal{L}(\mathbf{x}_o;\boldsymbol{\theta}), \tag{3.2}$$

where S is the hyper-rectangle defined by the symbolic variable s, and $\mathcal{L}(\cdot;\cdot)$ is the classical likelihood function for the specific individual observations. The symbolic likelihood includes an integral with a large exponent representing the points within the hyper-rectangle. In contrast, evaluating these points in the classical likelihood function is a product over the densities, i.e., $\prod_{i=1}^{n_t-n_b-n_o} g_{\mathbf{x}}(\mathbf{x}_i;\boldsymbol{\theta})$, which can be computationally prohibitive in the absence of low-dimensional sufficient statistics. It is expected that if the integral with a large exponent could be obtained at a low computational cost in the case of large n_t and small n_o, the symbolic approach would gain a potential advantage in reduced computing time compared with evaluating the likelihood function on the full data.

The min-max random hyper-rectangle approach in Beranger et al. (2018) constructs the symbolic likelihood function as

$$L_B(s;\boldsymbol{\theta}) \propto \left[\int_S g_{\mathbf{x}}(\mathbf{x};\boldsymbol{\theta})d\mathbf{x}\right]^{n_t-n_b} \times \mathcal{L}_b(\mathbf{x}_b;\boldsymbol{\theta}), \tag{3.3}$$

where n_t and n_b respectively denote the total number of observations and the number of observations attaining the minimum/maximum in at least one dimension. $\mathcal{L}_b$ is a likelihood term which takes into account the number of classical data points used to construct the random hyper-rectangle and their indicative (but not precise) location, e.g. "two points, one in the bottom left corner, one in the top right". This term does not make use of the precise location of these points $\mathbf{x}_b$. The general expression for the term is complex and case dependent; see the full expression in Beranger et al. (2018, Section 2.3.1).

Comparing (3.2) and (3.3), the differences in the symbol construction lead to different likelihood functions. Our proposal extends the Beranger et al. (2018) approach in two ways. First, we include the full locations of the boundary points instead of just their indicative locations. This enables us to properly capture the correlation between dimensions in the case of a large data set. Second, as the minimal and maximal values are driven by the extreme values of the observations, it is likely that such values will tend to $\pm\infty$ as the sample size increases. Consequently, the integral is close to 1 regardless of the true parameters. To mitigate this problem, Beranger et al. (2018) split the data into multiple subsets and construct a symbol for each subset. When the integral can be evaluated analytically, this approach is computationally feasible. When the integral has no analytical solution, its estimation can be computationally costly, as it is typically done by Monte Carlo integration. In addition, the process has to be repeated for each subset, which makes the whole process even more costly. In contrast, our proposal does not require data splitting. The estimation is based on one single integral, resulting in a moderate computational overhead. Compared with the data splitting approach, the total computational overhead required to evaluate one integral is less costly than multiple integrals.

3.2.2 Likelihood estimators

For $\mathbf{X}$ with more than 2 dimensions ($d > 2$), the analytical evaluation of the integral in (3.2) is often unavailable even for well-known distributions. The form of the likelihood $\mathcal{L}$ is p^n (omitting the known terms), where p is a probability that we can unbiasedly estimate ($\widehat{p}$), e.g. via Monte Carlo integration, with n a known integer. However, using this estimate directly as $\widehat{\mathcal{L}} = \widehat{p}^n$ or $\widehat{\log}\mathcal{L} = n \log \widehat{p}$ results in a biased estimate of the likelihood or log-likelihood function, which is problematic for using within a pseudo-marginal Metropolis-Hastings (PMMH) sampler. Another way of estimating $\mathcal{L}$ is to use n independent estimates of p such that $\prod_{m=1}^{n} \widehat{p}^m$ is an unbiased estimator of $\mathcal{L}$. However, this approach is computationally expensive as n is usually a large number, for example, $n > 10,000$. Inspired by the work of Gelman and Meng (1998) and Papaspiliopoulos (2011), we propose an exact method for unbiased estimation of the symbolic likelihood in Section 3.2.2.1. However, the exact method is slow due to the massive Monte Carlo simulation involved. Section 3.2.2.2 proposes an approximate method to speed up the computation. For observations from a multivariate normal distribution, Section 3.2.2.3 illustrates a more efficient way to estimate the likelihood.

3.2.2.1 An exact method: Path sampling using the Poisson estimator

A two-step procedure constructs an unbiased estimator for the symbolic likelihood function in (3.2). The first step uses path sampling (Gelman and Meng, 1998) to obtain an unbiased estimator of the logarithm of the likelihood. This ensures that multiplication by the scalar $(n_t - n_b - n_o)$ does not bias the estimator of $(n_t - n_b - n_o)\log\left(\int_S g_{\mathbf{x}}(\mathbf{x};\boldsymbol{\theta})d\mathbf{x}\right)$. The second step transforms the unbiased estimator of the log of the symbolic likelihood to an unbiased estimator of the symbolic likelihood by applying the Poisson estimator (Papaspiliopoulos, 2011). We now describe the approach in detail.

The logarithm of the symbolic likelihood function is (with constant omitted)

$$l(s;\boldsymbol{\theta}) = (n_t - n_b - n_o)\log\left[\int_S g_{\mathbf{x}}(\mathbf{x};\boldsymbol{\theta})d\mathbf{x}\right] + \log\mathcal{L}(\mathbf{x}_b;\boldsymbol{\theta}) + \log\mathcal{L}(\mathbf{x}_o;\boldsymbol{\theta}), \tag{3.4}$$

where the last two terms are the classic log-likelihood functions. It is possible to obtain an unbiased estimator for $\int_S g_{\mathbf{x}}(\mathbf{x};\boldsymbol{\theta})d\mathbf{x}$ in the first term by Monte Carlo or other sampling methods. However, unbiasedness is not preserved by the log transformation. Instead, we use the path sampler (Gelman and Meng, 1998) to obtain an unbiased estimator in the logarithmic scale as follows. Let $h_t(\mathbf{z};\boldsymbol{\theta}) = g_{\mathbf{x}}(\mathbf{z};\boldsymbol{\theta})^t$, $t \in [0,1]$ and

$$q_t(\mathbf{z};\boldsymbol{\theta}) = \frac{h_t(\mathbf{z};\boldsymbol{\theta})}{\int_s h_t(\mathbf{z};\boldsymbol{\theta})d\mathbf{z}}.$$

The logarithm of the term (an integral with large exponents) in the likelihood function (3.2) can be expressed as

$$(n_t - n_b - n_o)\log\int_S g_{\mathbf{x}}(\mathbf{x};\boldsymbol{\theta})d\mathbf{x} = (n_t - n_b - n_o)\left(\int_0^1 E_{q_t(\mathbf{z};\boldsymbol{\theta})}\left[\frac{d}{dt}\log h_t(\mathbf{z};\boldsymbol{\theta})\right]dt + \log\int_S 1d\mathbf{z}\right). \tag{3.5}$$

Appendix A.1 proves this result. Based on (3.5), the path sampler offers an elegant way to obtain the desired estimator by integrating over the so-called temperature t. The one-dimensional integral

$$\int_0^1 E_{q_t(\mathbf{z};\boldsymbol{\theta})}\left[\frac{d}{dt}\log h_t(\mathbf{z};\boldsymbol{\theta})\right]dt$$

usually does not have an analytical solution, but it is efficiently approximated using numerical integration methods. This chapter uses the trapezoidal rule.

To select an appropriate so-called "temperature ladder", i.e., a sequence of temperatures $t \in (0, 1]$ [1], we follow the Friel and Pettitt (2008) approach, with $t = (i/T)^5, i = 1, \dots, T$. This geometric series fixes the total number of temperatures T and places more points at the lower temperatures, where the value of $E_{q_t(\mathbf{z};\boldsymbol{\theta})}\left[\frac{d}{dt} \log h_t(\mathbf{z}; \boldsymbol{\theta})\right]$ changes drastically with t. Figure 3.2 demonstrates the results obtained by the path sampler using this temperature ladder. The left panel depicts the result obtained at each temperature. The area of the rectangles edged with blue in the left panel are used for the numerical integration over temperatures. The right panel shows that the path sampler provides an unbiased result (up to the error in the trapezoidal rule) compared with the true value. Algorithm 5 implements the path sampler.

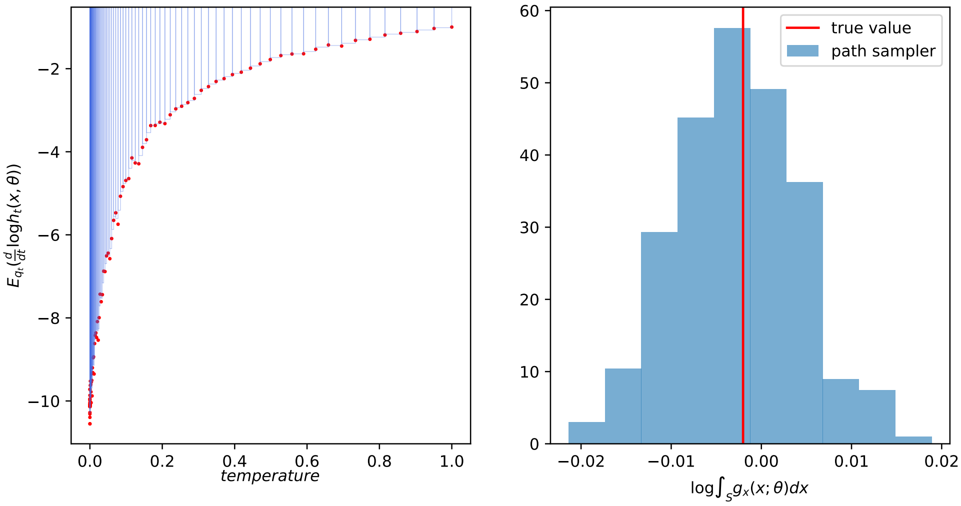


Figure 3.2: Demonstration of path sampler. The data are generated the same way as the middle plot of Figure 3.1. The temperature ladder is set as $t = (i/M)^5, i = 1, \dots M, M = 100$. For each temperature, we sample 2,000 $\mathbf{z}$s from the target distribution at each temperature. The left plot shows the estimates at each temperature. The right plot shows the histogram of 500 independent replications and the theoretical true value.

Recall that path sampling enables us to get an unbiased estimator for the logarithm of the likelihood function. The next step is to transform back to the original scale. Let

[1] The value of t usually cannot be zero as the corresponding density function $q_t(\mathbf{z}; \boldsymbol{\theta}) \propto$ constant, which does not integrate to 1. In the implementation, t usually starts from a small value, e.g., $(1/100)^5$. The approximation error is negligible by using the trapezoidal rule.

Algorithm 5 The path sampling algorithm

```
Input:
    $smin$, $smax$: vectors containing the marginal minimal/maximal values of $S$;
    $\boldsymbol{\theta}$: parameter(s);
    $t$: a vector of length $T$ between 0 and 1;
    $M$: number of samples to draw at temperature $t_i$, $i \in \{1, \dots, T\}$;
Output: an unbiased estimator of $\log \int_{smin}^{smax} g_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{x}$.
for $i = 1 \to T$ do
    $t_i \leftarrow$ the $i$th element of t.
    for $m = 1 \to M$ do
        Sample $\mathbf{z}_m$ from $q_{t_i}(\mathbf{z}; \boldsymbol{\theta})$.
    end for
    $\mathcal{T}_{t_i} \leftarrow \frac{1}{M} \sum_{m=1}^{M} \frac{d}{dt_i} \log h_{t_i}(\mathbf{z}_m; \boldsymbol{\theta})$
                                ▷ $\mathcal{T}_{t_i}$: an unbiased estimator of $E_{qt(\mathbf{z};\boldsymbol{\theta})}[\frac{d}{dt_i} \log h_{t_i}(\mathbf{z}; \boldsymbol{\theta})]$.
end for
Integrate $\mathcal{T}_{t_i}$ from $t_0$ (a number close to 0) to $t_T = 1$ numerically and use (3.5) to
obtain the final result.
```

$A = (n_t - n_b - n_o) \log \int_S g_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{x}$, and denote the corresponding estimator obtained by the path sampler as $\widehat{A_P}$. Due to the non-linearity of the exponential transformation, $E(\exp(\widehat{A_P})) \neq \exp(A)$. The Poisson estimator (Papaspiliopoulos, 2011) $\widehat{\exp}(A_P)$ ensures that $E(\widehat{\exp}(A_P)) = \exp(A_P)$ with

$$\widehat{\exp}(A_P) = \exp(a + \lambda) \prod_{h=1}^{\chi} \frac{(\widehat{A_P}^{(h)} - a)}{\lambda}; \tag{3.6}$$

here $\chi \sim \text{Poisson}(\lambda)$, $\widehat{A_P}^{(h)}$ is the hth ($h = 1, \dots, \chi$) realisation of $\widehat{A_P}$, which is an unbiased estimator of A_P, and a is an arbitrary real number. Note that if a is a lower bound of $\widehat{A_P}$, then the estimator in (3.6) is positive with probability 1. However, it is usually difficult to obtain a tight lower bound in many cases. Here we use the soft lower bound $a = \widehat{A_P} - \lambda$ as proposed in Quiroz et al. (2021), where $\widehat{A_P}$ is a random draw from the realised $\widehat{A_P}^{(h)}$

($h = 1, \ldots, \chi$). The unbiasedness of the estimator is guaranteed based on Property 4 of Appendix A.2. In the implementation, we draw Quiroz et al. (2021) use the block-Poisson estimator in the PM framework, where each block consists of a Poisson estimator. The default value of λ is 1 in the block-Poisson estimator. Here we set λ to a larger integer, e.g. $\lambda = 3$, to avoid a high probability of getting $\chi = 0$ as there is one block. When $\chi = 0$, the Poisson estimator is reduced to $\exp(a + \lambda)$, which can be a poor estimation of $\exp(A)$ if a is far from the tight lower bound of $\widehat{A_P} - \lambda$.

3.2.2.2 The approximate method: Taylor expansion with bias-correction

Even though the exact method generates an unbiased likelihood estimator, it is computationally costly as there are three nested loops in its implementation. To execute the path sampler (Algorithm 5), a loop over T temperatures is required. For each temperature, another loop includes M draws for evaluating the expectation. Furthermore, the Poisson estimator requires λ replications on average based on (3.6). The computational cost associated is more of a concern in the Bayesian context as the nested loop requires re-evaluation for each parameter proposal in each MCMC iteration.

Similarly to the exact method, the approximate method also consists of a two-step procedure. In the first step, the logarithm of the likelihood function is approximated by a quadratic Taylor series. Suppose $E(\widehat{B}^{(m)}) = B = \int_S g_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{x}$, $m = 1, \ldots, M$. It is straightforward to show that the Taylor expansion of $\log \widehat{B}^{(m)}$ in a neighbourhood of B is

$$\log \widehat{B}^{(m)} = \sum_{j=0}^{2} \frac{\log^{(j)}(B)}{j!} (\widehat{B}^{(m)} - B)^j + O((\widehat{B}^{(m)} - B)^3),$$

where $\log^{(j)}(B)$ refers to the jth derivative of $\log(B)$ with respect to B.

Discarding the higher order degree 3 term and taking the expectation on both sides, we have

$$\begin{aligned} E(\log \widehat{B}^{(m)}) &\approx \log B + \frac{1}{B} E(\widehat{B}^{(m)} - B) - \frac{1}{2B^2} E((\widehat{B}^{(m)} - B)^2) \\ &= \log B - \frac{1}{2B^2} \text{Var}(\widehat{B}^{(m)}). \end{aligned}$$

Define $\widehat{\log} B$ as an estimator of $\log B$, the equation above can be re-expressed as

$$\widehat{\log} B = \frac{1}{M} \sum_{m=1}^{M} \log \widehat{B}^{(m)} + \frac{1}{2B^2} \mathrm{Var}(\widehat{B}), \tag{3.7}$$

where $\widehat{B}$ is the random variable from which the $\widehat{B}^{(m)}$s are drawn.

Recall that $A = (n_t - n_b - n_o) \log \int_S g_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{x}$ and use the connection between A and B: $A = (n_t - n_b - n_o) \log B$; then the approximate estimator $\widehat{A_T}$ for A is

$$\widehat{A_T} = (n_t - n_b - n_o) \widehat{\log} B. \tag{3.8}$$

Suppose that $\log \widehat{B}^{(m)}$ is approximately normal distributed, i.e., $\log \widehat{B}^{(m)} \overset{iid}{\sim} N(\mu, \sigma^2)$. It is straightforward to show that $B = \exp(\mu + \sigma^2/2)$. Then, $\widehat{\log} B$ is also likely to be normally distributed as it is a linear combination of $\log \widehat{B}^{(m)}$ plus a constant $\frac{1}{2B^2} \mathrm{Var} \widehat{B}$. Similarly, by (3.8), $\widehat{A_T}$ is also (approximately) normally distributed. It is straightforward to propose the approximate estimator (also called the “approximately bias-corrected estimator”), $\widehat{\exp}(A_T)$ of $\exp(A)$ as :

$$\widehat{\exp}(A_T) := \exp\left(\widehat{A_T} - \frac{1}{2} s(\widehat{A_T})\right),$$

where we use the sample variance $s(\widehat{A_T})$ to replace the unknown quantity $\sigma^2 = \mathrm{Var}(\widehat{A_T})$, given a relatively large M. The Monte Carlo estimates $\widehat{B}^{(m)}$s in (3.7) are used to compute the sample variance of $s(\widehat{A_T})$ based on (3.8).

3.2.2.3 The likelihood estimator for a truncated multivariate normal distribution

The approximate method in the last section assumes the availability of an unbiased estimator $\widehat{B}^{(m)}, m = 1, \ldots, M$, for the integral $B = \int_S g_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{x}$. However, taking the average of the independent samples ($\frac{1}{M} \sum_{m=1}^{M} \widehat{B}^{(m)}$) from the restricted region may result in an estimator with large variability, especially in high dimensions. If the underlying distribution is of a specific form, there may exist a more efficient way of doing the Monte Carlo integration. This chapter focuses on an efficient estimator for multivariate normal distributions.

Assuming $g_{\mathbf{x}}(\mathbf{x};\boldsymbol{\theta})$ is a d-dimensional multivariate normal density function with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ ($\boldsymbol{\theta} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$), the separation of variables (SOV) estimator is widely used to compute probabilities over some region S. SOV, first proposed by Genz (1992), is a numerical computational method to evaluate an integral by decomposing the region S into d one-dimensional areas. Botev (2017) extends the SOV estimator into a so-called minimax-exponentially-tilted (MET) estimator for simulating independent observations from a truncated multivariate normal distribution as well as computing the cumulative distribution function. The MET estimator has lower variance than the SOV estimator based on simulation studies demonstrated in Botev (2017), which is desirable in the PM framework. This chapter implements the MET estimator for computing $\int_S g_{\mathbf{x}}(\mathbf{x};\boldsymbol{\theta})d\mathbf{x}$.

The general idea behind the SOV or the MET estimator is the representation

$$\int_S g_{\mathbf{x}}(\mathbf{x};\boldsymbol{\theta})d\mathbf{x} = \int_{S'} g_{\mathbf{x}}(\mathbf{x};\mathbf{0},\boldsymbol{\Sigma})d\mathbf{x} = (2\pi)^{-d/2}\int_{l'_1}^{u'_1} \exp\left(-\frac{y_1^2}{2}\right) dy_1 \cdots \int_{l'_d}^{u'_d} \exp\left(-\frac{y_d^2}{2}\right) dy_d, \tag{3.9}$$

where $\mathbf{y} = \mathbf{L}^{-1}\mathbf{x}$, $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^T$ ($\mathbf{L}$ is a lower triangular matrix), and $S' = S - \boldsymbol{\mu}$ to ensure $\mathbf{y} = (y_1, \ldots, y_d)^T$ has mean vector 0. Here S is the region defined by the boundary points of the symbol s, e.g., S is a rectangle for a 2-dimensional space and is cuboid for a 3-dimensional space. $S - \boldsymbol{\mu}$ means shifting S along the direction $\boldsymbol{\mu}$. Denote the lower and upper bounds of S' by $\tilde{\mathbf{l}} = (\tilde{l}_1, \ldots, \tilde{l}_d), \tilde{\mathbf{u}} = (\tilde{u}_1, \ldots, \tilde{u}_d)$, and define $\mathbf{l}' = (l'_1, \ldots, l'_d)$, $\mathbf{u}' = (u'_1, \ldots, u'_d)$ as satisfying the following conditions:

$$\begin{aligned} & l'_1 = \tilde{l}_1; \quad u'_1 = \tilde{u}_1; \\ & l'_i = \tilde{l}_i - \sum_{j=1}^{i-1} L_{ij} y_j / L_{jj}; \quad u'_i = \tilde{u}_i - \sum_{j=1}^{i-1} L_{ij} y_j / L_{jj}, \quad i = 2, \ldots, d, \\ & l'_i \leq y_i \leq u'_i \quad i = 1, \ldots, d, \end{aligned}$$

where the subscript i denotes the ith element of the corresponding vector of length d and L_{ij} is the (i, j)th element of $\mathbf{L}$. The SOV estimator first evaluates the integral

$$\int_{l'_1}^{u'_1} \exp\left(-\frac{y_1^2}{2}\right) dy_1,$$

and then samples y_1 between l'_1 and u'_1, which is equivalent to sampling from a truncated normal distribution. The region l'_2, u'_2 is determined based on y_1. Hence, the next integral

$\int_{l_2'}^{u_2'} \exp\left(-\frac{y_2^2}{2}\right) dy_2$ can be evaluated with high precision. All the remaining integrals proceed by similar steps, with the realisation of y_i $(i \leq d)$ computed. The difference between the MET and the SOV estimators is that SOV uses (3.9) directly to evaluate the integral, whereas MET uses a tilting parameter which shifts $\mathbf{l}', \mathbf{u}'$ as well. See Botev (2017) and references therein for further details.

3.2.3 The signed block PMMH algorithm with the Poisson estimator

A popular way of constructing MCMC samplers involving intractable likelihood functions is to use unbiased likelihood estimates in place of the unavailable likelihood function. Andrieu and Roberts (2009) formally investigate the properties of such samplers and call the approach "the pseudo-marginal" (PM) method. The key criterion for an efficient PM method is that the variances of the logarithm of likelihood estimates should be sufficiently small, approximately in the interval $[1, 3]$, to achieve an optimal trade-off between sampling efficiency and computational cost (Doucet et al., 2015; Pitt et al., 2012). If the likelihood is greatly overestimated, then the Markov chain is likely to get stuck for many iterations. Controlling the variance of the logarithm of the likelihood estimator is thus crucial, in particular for the symbolic likelihood setting as the variance increases quadratically with the number of observations, indicating that we need to correspondingly increase the number of random samples for generating the likelihood estimates.

In PM methods, the randomness in the likelihood estimates for $L_f(s; \boldsymbol{\theta})$ is determined by the random numbers $\mathbf{u} = (u_1, \ldots, u_M)$ and therefore the likelihood estimator is denoted as $\widehat{L_f}(s; \boldsymbol{\theta}, \mathbf{u})$. It is now well-known that correlating the logarithm of the likelihood estimators at the current and proposed draws increases the sampling efficiency of PM methods by controlling the variability of the likelihood ratio in the Metropolis-Hastings (MH) acceptance probability (Deligiannidis et al., 2018; Tran et al., 2016). To do so, we adopt the block PM method in Tran et al. (2016). The idea behind the block PM method is to group the random numbers $\mathbf{u}$ into blocks, and update one block at a time, holding the other blocks fixed. The updated blocked random numbers are denoted by $\mathbf{u}'$. The blocking strategy induces the correlation between the logarithm of the likelihood estimates (determined by $\mathbf{u}$ and $\mathbf{u}'$ respectively) in the MH acceptance probability.

For the methods introduced in Section 3.2.2.1 and Section 3.2.2.2, the blocks can be formed flexibly. For the exact (path sampling) method, the block can be generated by grouping at the temperature level or over the particles drawn at each temperature. For the approximate method, there are M particles involved in generating the likelihood estimator. Each particle can form a block on its own. With the update occurring at a few blocks, the correlation between the likelihood estimators at the current and the proposed values is much higher. The key result established by Tran et al. (2016) is that the optimal value for the variance of the logarithm of the likelihood estimator after blocking is $\sigma^2_{opt} \approx 2.16^2/(1-\rho^2)$ where $\rho = 1 - 1/M$ with the update on one out of M blocks. For example, by having $M = 500$, we have the optimal variance for each block is $\sigma^2_{opt}/M \approx 2.33$. The optimal variance on the likelihood estimator is then $1,167$, which is significantly larger than the optimal $\approx [1,3]$ in the standard PM method. Note that allowing a larger variance means that we need less random numbers, which results in a computationally more efficient algorithm. In our implementation, we set a prespecified value for ρ and adjust the number of blocks M to achieve σ^2_{opt}.

Another issue to consider is that in the exact method, the Poisson estimator in (3.6) sometimes generates a negative estimate for the likelihood function. The optimal value of the lower bound a_{opt} is $a_{opt} = A_P - \lambda$ (see Appendix A.2 for the proof). It often impractical to obtain a_{opt} as the quantity A_P itself is intractable. In the application, we set $\widehat{a}_{opt} = \widehat{A_P} - \lambda$. Here $\widehat{A_P}$ is a random draw from all realised $\widehat{A_P}^{(h)}$s in the Poisson estimator. If $\chi = 0$, we generate an estimate $\widehat{A_P}$ from scratch. To cope with the possible negative estimates, rewrite (3.6) by taking its absolute value

$$|\widehat{\exp}(A_P)| = \exp(a + \lambda) \left| \prod_{h=1}^{\chi} \frac{(\widehat{A_P}^{(h)} - a)}{\lambda} \right|. \tag{3.10}$$

The final absolute value of the likelihood estimate for one symbol is $|\widehat{L_f}(s;\boldsymbol{\theta},\mathbf{u})| \propto |\widehat{\exp}(A_P)| \times \mathcal{L}(\mathbf{x}_b;\boldsymbol{\theta}) \times \mathcal{L}(\mathbf{x}_o;\boldsymbol{\theta})$. Obviously, as an estimator of $L_f(s;\boldsymbol{\theta})$, $|\widehat{L_f}(s;\boldsymbol{\theta},\mathbf{u})|$ is no longer unbiased, but we can still carry out inference from the posterior distribution of interest by following the approach in Lyne et al. (2015). Lyne et al. (2015) use a PM method with the biased likelihood estimator in (3.10), where the sign of the likelihood estimates is tracked and subsequently used in an importance sampling step to obtain consistent estimates of expectations with respect to the true posterior. Algorithm 6 illustrates the steps in detail. We refer the reader to Lyne et al. (2015); Quiroz et al. (2021) for more details on the *signed PMMH* algorithm.

The last implementation challenge is finding a good proposal distribution $q(\boldsymbol{\theta}'|\boldsymbol{\theta})$ in Algorithm 6. We use a random walk proposal in this chapter, which is a Gaussian distribution centred on the current value of the Markov chain with a specific covariance matrix. Here, the covariance matrix is formulated by the adaptive MCMC strategy in Haario et al. (2001). We also follow the approach proposed in Garthwaite et al. (2016), where a stochastic search algorithm based on the Robbins-Monro process (Robbins and Monro, 1951) is used to determine the scale of the covariance matrix needed to achieve a specific overall sampler acceptance probability, here $\alpha = 0.234$ in (3.11). The general idea of this approach is that the scale of the covariance matrix gets shrunk in the next iteration of MCMC if the proposed value is rejected, which leads to a potentially smaller proposed move and a larger acceptance probability in the next iteration. If the proposed value is accepted, the next proposal distribution is more likely to result in a larger jump by increasing the scale. The changes in scale are chosen to achieve the target overall sampler acceptance probability. See Garthwaite et al. (2016) and the references therein for further details.

3.3 Simulated examples

3.3.1 Example 1: Correlation in a bivariate normal distribution

We compare the performance of our proposed symbolic likelihood in the case of the full min-max random hyper-rectangle ($n_o = 0$), and the one proposed in Beranger et al.

Algorithm 6 The *signed block PMMH algorithm*

```
Input:
    $S$: symbol information.
    $\mathbf{u}$: random numbers between 0 and 1, grouped in $M$ blocks.
    $\boldsymbol{\theta}_0$: starting value for parameters.
    $iter$: total number of iterations.
Output: An unbiased estimator for $\psi(\boldsymbol{\theta})$ from target distribution.
for $i = 1 \rightarrow iter$ do
    Generate $\mathbf{u}'$ given $\mathbf{u}$ update one block out of $M$ blocks.
    Generate $\boldsymbol{\theta}'$ given $\boldsymbol{\theta}_{i-1}$ by $q(\boldsymbol{\theta}'|\boldsymbol{\theta}_{i-1})$.
    Calculate the acceptance ratio:
        $\text{acceptance ratio } \alpha = \min\left\{1, \frac{|\widehat{L_f}(s;\boldsymbol{\theta}',\mathbf{u}')|\pi(\boldsymbol{\theta}')}{|\widehat{L_f}(s;\boldsymbol{\theta}_{i-1},\mathbf{u})|\pi(\boldsymbol{\theta}_{i-1})} \times \frac{q(\boldsymbol{\theta}_{i-1}|\boldsymbol{\theta}')}{q(\boldsymbol{\theta}'|\boldsymbol{\theta}_{i-1})}\right\}.$    (3.11)
    Generate $a$ from Uniform(0,1).
    if $\alpha > a$ then
        Accept $\boldsymbol{\theta}_i \leftarrow \boldsymbol{\theta}'$.
        Update $\mathbf{u} \leftarrow \mathbf{u}'$.
    else
        Maintain $\boldsymbol{\theta}_i \leftarrow \boldsymbol{\theta}$.
        No update for $\mathbf{u}$.
    end if
    $\text{sign}(\boldsymbol{\theta}_i|s) \leftarrow \text{sign}(\widehat{L_f}(s;\boldsymbol{\theta}_i,\mathbf{u}))$.    ▷ $\text{sign}(x) = 1$ if $x > 0$; $\text{sign}(x) = -1$ if $x < 0$.
end for
$h(\boldsymbol{\theta}) \leftarrow \frac{\sum_{i=1}^{iter} h(\boldsymbol{\theta}_i)\text{sign}(\boldsymbol{\theta}_i|s)}{\sum_{i=1}^{iter} \text{sign}(\boldsymbol{\theta}_i|s)}$.
```

(2018) by replicating the experiment in Beranger et al. (2018, Section 3.2). The symbolic likelihood functions are denoted as L_f (our likelihood) and L_B (Beranger's likelihood) respectively.

We construct $m = 20, 50$ symbols, each of which is obtained from a random sample of size $n_c = 5, 10, 50, 100, 1{,}000, 100{,}000$ from a bivariate normal distribution with $\boldsymbol{\mu} = (2, 5)$, $\sigma_1^2, \sigma_2^2 = 0.5^2$, and the correlations $\rho = 0, 0.3, 0.5, 0.7, 0.9$.

Table 3.1 reports the mean and standard deviation of the estimate $\widehat{\rho}$ under both likelihoods, taken over 100 replicate data sets. For L_f, the estimates are unbiased and close to the true value in all settings, with a smaller standard deviation compared to those using L_B. For L_B with small sample sizes, most results are close to the true values as well, except where $n_c =$ 1,000 and 100,000 with $\rho = 0.3, 0.5, 0.7$. Beranger et al. (2018) explain that when n_c is large, under any fixed correlation, it is increasingly likely that the min-max constructed random rectangle is constructed from 4 unique data points (in 2-dimensional data). As the number of rectangle-defining data points is used by Beranger et al. (2018) to determine the strength of the correlation, (i.e., 2 points imply strong correlation; 4 points imply weaker or no correlation), this means that L_B underestimates the magnitude of ρ for any value of correlation, once n_c is sufficiently large.

Table 3.1 clearly shows this effect, particularly for smaller ρ. This limitation in Beranger et al.'s approach can be overcome by using all the information of the boundary points, including their precise locations, which our proposed method achieves. Hence, the L_f results are unbiased and precise for all ρ and n_c. Accordingly, we have resolved this bias problem with the symbolic likelihood of Beranger et al..

3.3.2 Example 2: Comparing the exact and approximate method

Section 3.2.2.1 and Section 3.2.2.2 explain that both the exact and the approximate methods involve a two-step procedure to estimate the likelihood. We now examine and compare the results of both methods for each step in this simulated example.

We first compare the performance of path sampling (exact method) and the Taylor approximation in estimating the logarithm of the likelihood function. After selecting the method

		m=20					m=50				
ρ	n_c	5	10	100	1,000	100,000	5	10	100	1,000	100,000
0.0	L_B	-0.001	0.015	0.006	-0.003	0.000	-0.009	0.001	-0.001	0.011	0.000
		(0.126)	(0.123)	(0.146)	(0.068)	(0.004)	(0.087)	(0.082)	(0.100)	(0.108)	(0.004)
	L_f	-0.006	0.014	0.000	-0.006	0.000	-0.017	0.001	-0.001	-0.003	0.000
		(0.108)	(0.073)	(0.046)	(0.037)	(0.025)	(0.071)	(0.045)	(0.026)	(0.023)	(0.017)
0.3	L_B	0.304	0.297	0.273	0.168	0.012	0.306	0.303	0.289	0.249	0.045
		(0.112)	(0.129)	(0.160)	(0.217)	(0.087)	(0.067)	(0.066)	(0.100)	(0.152)	(0.143)
	L_f	0.306	0.307	0.299	0.306	0.299	0.308	0.306	0.301	0.305	0.300
		(0.102)	(0.067)	(0.038)	(0.032)	(0.023)	(0.058)	(0.043)	(0.028)	(0.018)	(0.014)
0.5	L_B	0.505	0.499	0.490	0.426	0.212	0.509	0.503	0.494	0.488	0.315
		(0.094)	(0.105)	(0.134)	(0.204)	(0.298)	(0.058)	(0.055)	(0.083)	(0.076)	(0.274)
	L_f	0.504	0.505	0.499	0.505	0.503	0.506	0.505	0.501	0.502	0.503
		(0.084)	(0.059)	(0.035)	(0.029)	(0.021)	(0.048)	(0.036)	(0.023)	(0.018)	(0.014)
0.7	L_B	0.701	0.700	0.696	0.692	0.641	0.706	0.702	0.701	0.701	0.695
		(0.077)	(0.074)	(0.079)	(0.081)	(0.233)	(0.044)	(0.039)	(0.047)	(0.045)	(0.055)
	L_f	0.702	0.705	0.700	0.703	0.703	0.704	0.704	0.700	0.701	0.703
		(0.060)	(0.043)	(0.029)	(0.021)	(0.018)	(0.034)	(0.025)	(0.019)	(0.015)	(0.012)
0.9	L_B	0.901	0.900	0.901	0.901	0.903	0.902	0.901	0.901	0.900	0.902
		(0.030)	(0.026)	(0.025)	(0.028)	(0.023)	(0.017)	(0.014)	(0.016)	(0.016)	(0.015)
	L_f	0.900	0.901	0.901	0.901	0.903	0.901	0.900	0.901	0.900	0.902
		(0.023)	(0.018)	(0.016)	(0.012)	(0.008)	(0.013)	(0.011)	(0.009)	(0.008)	(0.007)

Table 3.1: Mean and standard deviation (in brackets) of the estimated correlation ρ over 100 independent data sets. The estimate is given by maximising the symbolic likelihood L_B (Beranger's method) and L_f (our method), respectively. The table shows the number of symbols (m), the number of data points per symbol (n_c), and the true correlation ρ between the two variables.

with the better performance, the Poisson estimator and the bias-corrected estimator are implemented based on the selected method with the purpose of checking whether they provide similar results.

As an illustrating example, we choose $g_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ to be a d-dimensional multivariate normal distribution with $\boldsymbol{\mu} = \mathbf{0}_d, \boldsymbol{\Sigma} = 0.5\mathbf{I}_d + 0.5\mathbf{1}_d\mathbf{1}_d^\top$, where $\mathbf{0}_d$, $\mathbf{1}_d$ are respectively d-dimensional vectors of 0's and 1's and $\mathbf{I}_d$ is a d-dimensional identity matrix. The integration region between the lower and upper bounds is fixed as $S = [-2 \times \mathbf{1}_d, 2 \times \mathbf{1}_d]$, for $d = 2, \ldots, 10$. The number of observations is $n = 100$, which is large enough to compare the results of the two methods.

Table 3.2 shows the mean and variance of the estimated log-likelihood $\widehat{\log L}$ for $\log L =$

$n \log \int_S g_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\mu}) d\mathbf{x}$, under both path sampling and the Taylor approximate methods taken over 1,000 replicate calculations. The log-likelihood is estimated at the parameter values stated above. Both methods provide similar results for low-dimensional cases ($d \leq 5$). The difference in estimates becomes slightly larger as the number of dimensions increases. The largest difference is 0.25 at $d = 10$ which is around 2% of the mean value. The path sampler gives an unbiased estimator of the log-likelihood function in theory. However, in real applications, the unbiasedness is compromised by the numerical integration in (3.5). Therefore, we cannot evaluate which method provides the result closest to the true value ($n \log \int_S g_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x}$) unless we know the analytical solution of the integral. The estimator obtained by the Taylor expansion has lower variance than the path sampler, and the gap increases with dimension. The computing time of 1,000 replications shows that the Taylor approximation only takes around 1% of the path sampler's time, i.e., it is 100 times faster. We conclude that, compared to the path sampler, the Taylor approximation has good accuracy with lower variance and a significantly lower computing time, particularly for lower dimensions.

dim	mean of $\widehat{\log} L$		var of $\widehat{\log} L$		time(secs)	
	path	Taylor	path	Taylor	path	Taylor
2	-8.645	-8.652	0.113	0.003	298.393	2.959
3	-12.245	-12.215	0.188	0.011	354.071	3.636
4	-15.482	-15.472	0.316	0.027	433.104	4.503
5	-18.449	-18.499	0.382	0.048	510.373	5.318
6	-21.236	-21.314	0.528	0.075	564.963	6.028
7	-23.859	-24.041	0.641	0.100	617.368	6.297
8	-26.333	-26.597	0.772	0.131	656.173	6.751
9	-28.723	-29.141	0.818	0.163	728.830	8.369
10	-30.937	-31.534	0.973	0.201	703.758	8.244

Table 3.2: Mean, variance and execution time of 1,000 independent replications of estimating $\log L = n \log \int_S g_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x}$ with $n = 100$, where $g_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a d-dimensional multivariate normal distribution. Here $\boldsymbol{\mu} = \mathbf{0_d}, \boldsymbol{\Sigma} = 0.5 \times \mathbf{I_d} + 0.5 \times \mathbf{1}_d \mathbf{1}_d^\top, S = [-2\mathbf{1}_d, 2\mathbf{1}_d]$ and $d = 2, \ldots, 10$. For the path sampler (path), the temperature ladder is defined as $(t/T)^5$ with $T = 100$ and $t = 1, \ldots, T$. The number of Monte Carlo draws at each temperature is $M = 2{,}000$. For the Taylor approximation (Taylor), we also set $M = 2{,}000$.

Table 3.3 shows the mean and variance of the logarithm of the absolute likelihood estimator obtained by the Poisson method and the bias-correction method using identical settings

dim	mean of $\log\lvert\widehat{L}\rvert$		var of $\log\lvert\widehat{L}\rvert$		time(secs)	
	pois	bc	pois	bc	pois	bc
2	-8.654	-8.654	0.000	0.003	8.446	2.811
3	-12.217	-12.223	0.003	0.011	17.011	5.803
4	-15.468	-15.473	0.009	0.026	18.591	6.284
5	-18.503	-18.507	0.018	0.043	19.629	6.377
6	-21.341	-21.369	0.032	0.072	20.900	6.849
7	-24.061	-24.089	0.055	0.102	23.680	7.608
8	-26.671	-26.709	0.107	0.131	24.424	7.860
9	-29.194	-29.240	0.170	0.174	27.146	8.668
10	-31.678	-31.680	0.304	0.196	29.563	9.561

Table 3.3: Mean, variance and execution time of 1,000 independent replications of estimating the (absolute) value of likelihood function $L = [\int_S g_{\mathbf{x}}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})d\mathbf{x}]^n$, where the estimator of the log-likelihood function is provided by the approximate method. For the Poisson estimator (pois), $\lambda = 3$, $a = \widehat{A}^{(h)} - \lambda$, with h being a random number from $\{1,\ldots,\chi\}$, $\chi \sim \text{Pois}(\lambda)$. The term "bc" refers to the bias-corrected estimator.

to Table 3.2. We compare the results on the logarithmic scale as the target of interest $[\int_S g_{\mathbf{x}}(\mathbf{x};\boldsymbol{\theta})d\mathbf{x}]^n$ is numerically close to zero for large values of n. Note that this is different from the estimator of the log likelihood ($\log\widehat{L} \neq \widehat{\log L}$). Both estimators provide close results for the mean values across all dimensions. The Poisson estimator has a smaller variance for data of dimension 1 to 9. However, the Poisson estimator has larger variance than the bias-correction method, supported by the result of 10-dimensional data, which is the highest dimension considered in this simulation study. For higher dimensions, the bias-correction method continues to provide the estimator with lower variance compared with the Poisson estimator (results not shown here). The associated computing time is 3 times longer than that of the approximate method, which is a result of setting $\lambda = 3$. For the soft lower bound a, its computation is based on a random draw from the realisations $\widehat{A}^{(h)}$ with $h \in \{1,\ldots,\chi\}$, used in the Poisson estimator. See Appendix A.2 for more details.

We close this example by summarising that the approximate method (the Taylor expansion with bias-correction) offers results close to the exact method, but with significantly less computing time. We thus use the approximate method in the following analyses.

3.3.3 Example 3: Implementation of SDA on a factor model

This simulation study applies our SDA method on a factor model, where observations are assumed from a multivariate normal distribution with the covariance matrix constructed via a low-rank approximation. For d-dimensional observations, the covariance matrix is assumed to be $\mathbf{\Sigma} = \mathbf{B}\mathbf{B}^\top + \mathbf{D}$, where $\mathbf{B}$ is a low triangular matrix of size $d \times k$ $(k \ll d)$ and $\mathbf{D}$ is a $d \times d$ diagonal matrix with positive entries. There are $dk - k(k-1)/2 + d$ elements to be estimated instead of the $d(d+1)/2$ elements for the full covariance matrix. The factor model is $\mathbf{y}_i = \boldsymbol{\mu} + \mathbf{B}\mathbf{f}_i + \boldsymbol{\epsilon}_i$, where $\mathbf{f}_i \sim N(0, \mathbf{I}_d)$, $\boldsymbol{\epsilon}_i \overset{iid}{\sim} N(0, \mathbf{D})$.

One Bayesian approach is to model the latent variables $\mathbf{f}_i$ so that the full conditional distributions of all the parameters can be derived in closed form (Geweke and Zhou, 2015). However, each observation has its own corresponding latent variable(s), indicating that many latent variables are required for a big data set. For example, for a data set with n observations, it is necessary to update $(d \times n) + (d \times k - k \times (k-1)/2) + (2 \times d)$ parameters per MCMC iteration, which requires significant memory and computational resources. The first term $(d \times n)$ is the number of latent variables; the middle term $(d \times k - k \times (k-1)/2)$ refers to the number of parameters in $\mathbf{B}$; and the last term $(2 \times d)$ is the number of the number of elements of $\boldsymbol{\mu}$ and diagonal elements of $\mathbf{D}$. To ensure a fair comparison between the analysis on the full data and SDA, we use the MH algorithm within Gibbs for the full data and symbols, where all elements in $\boldsymbol{\mu}$, $\mathbf{D}$ and $\mathbf{B}$ are updated in a block conditioning on the other parameters.

We consider a data dimension d from 3 to 10 and set $k = 1$ for this simulation study. Each data set includes 50,000, 100,000 or 500,000 independent observations with $\mathbf{y} \sim N(\boldsymbol{\mu}, \mathbf{\Sigma})$. The elements of the d-dimensional vector $\boldsymbol{\mu}$ are equally spaced from -1 to 1. The covariance matrix is constructed by $\mathbf{\Sigma} = \mathbf{B}\mathbf{B}^\top + \mathbf{D}$, with $\log D_{ii} \sim \text{Uniform}(0, 0.25)$, $B_{ij} \sim \text{Uniform}(-0.5, 0.5)$ with $i = 1, \ldots, d$ and $1 \leq j \leq i$. For each data set, only one symbol is constructed by setting $q = 0.005$ (0.5th quantile). In the PM method, we set $\mathbf{u}$ as a collection of 500, 1,000 and 6,000 random numbers respectively to estimate the likelihood function. The number increases with the size of the corresponding data set. To implement the *signed block PMMH* algorithm (Algorithm 6), only one element in $\mathbf{u}$ is updated randomly per MCMC iteration to induce the correlation in the block PM method. For both approaches, we run the MCMC for 10,000 iterations and use the last 5,000 samples to estimate the posterior.

Table 3.4 shows the average value of the root mean squared error (RMSE) for $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and the running time based on 10 independent replications. The calculation of RMSE is for $\boldsymbol{\mu}$ and the lower triangular matrix of $\boldsymbol{\Sigma}$ with the formula:

$$\mathrm{RMSE}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\widehat{\theta}_i - \theta_i)^2},$$

where $\widehat{\boldsymbol{\theta}}$ is the posterior mean vector of length N and $\boldsymbol{\theta}$ is the true value. For symbolic data, as we retain the data outside the q and $1-q$ quantiles across every dimension, the number of observations outside the d-dimensional hyper-rectangle increases roughly linearly with the number of dimensions given a moderate correlation between dimensions. In the worst case, SDA requires $2dq \times 100\%$ of the original data points. The percent of observations required by SDA is in column "obs%" of Table 3.4.

Based on Table 3.4, the mean RMSE ratio between SDA and full data results is between 2 to 3 for the results of $\boldsymbol{\mu}$, and 1 to 6 for the results of $\boldsymbol{\Sigma}$ (lower triangular part). The results show that the the full data analysis is more accurate than SDA, which is unsurprising as there is some loss of information in the symbols compared to the full data set. The computing time ratios for SDA vs full data range between 0.2 and 0.9, showing that SDA requires less time than the full data approach. The computational advantage of SDA is more evident for big data sets (n = 500,000). For smaller data sets (e.g. n = 50,000), given the loss in precision, analysing the full data set is better.

Two implementation details are worth mentioning. First, the symbol construction time is not included here as it is negligible compared with MCMC execution time. For example, when n = 500,000 and $d = 10$, the symbol construction for the quantile min-max interval ($q = 0.005$) costs around 1 second on average. Second, the full data analysis is speeded up by vectorising the computation wherever possible, which saves a large amount of the time. In spite of optimising the code for the full data, SDA is faster than the full data approach across all data sets, with the advantage more pronounced for large data sets.

We close this example by summarising that SDA saves computing time compared to the full data approach, at the cost of less accurate results.

50,000

dim	obs%	RMSE (μ)		ratio	RMSE ($\mathbf{\Sigma}$)		ratio	time(secs)		ratio
		Full	SDA		Full	SDA		Full	SDA	
3	2.980	0.004	0.014	3.832	0.007	0.013	1.992	117.979	111.172	0.942
4	3.946	0.005	0.015	3.130	0.006	0.010	1.788	146.723	131.156	0.894
5	4.912	0.005	0.014	2.610	0.005	0.010	2.068	190.498	137.421	0.721
6	5.860	0.005	0.015	2.788	0.005	0.011	2.225	206.843	155.910	0.754
7	6.806	0.005	0.011	2.434	0.005	0.011	2.249	291.440	175.189	0.601
8	7.716	0.005	0.015	3.248	0.005	0.014	2.872	335.309	173.867	0.519
9	8.642	0.005	0.011	2.434	0.005	0.012	2.641	350.184	196.541	0.561
10	9.562	0.005	0.012	2.452	0.004	0.011	2.597	319.808	227.150	0.710

100,000

dim	obs%	RMSE (μ)		ratio	RMSE ($\mathbf{\Sigma}$)		ratio	time(secs)		ratio
		Full	SDA		Full	SDA		Full	SDA	
3	2.975	0.003	0.009	3.163	0.004	0.010	2.667	176.482	125.282	0.710
4	3.942	0.003	0.012	3.730	0.004	0.009	2.033	325.470	137.355	0.422
5	4.894	0.003	0.007	2.097	0.004	0.009	2.360	285.435	163.247	0.572
6	5.848	0.003	0.009	2.916	0.003	0.008	2.500	409.198	175.770	0.430
7	6.787	0.003	0.008	2.382	0.004	0.009	2.560	493.279	204.891	0.415
8	7.707	0.003	0.010	3.115	0.003	0.010	3.160	446.424	202.067	0.453
9	8.621	0.003	0.010	3.139	0.003	0.013	4.425	480.807	218.753	0.455
10	9.539	0.003	0.008	2.397	0.003	0.010	3.279	546.189	236.209	0.432

500,000

dim	obs %	RMSE (μ)		ratio	RMSE ($\mathbf{\Sigma}$)		ratio	time(secs)		ratio
		Full	SDA		Full	SDA		Full	SDA	
3	2.966	0.001	0.004	2.893	0.002	0.004	2.346	1237.003	258.973	0.209
4	3.938	0.001	0.005	3.279	0.002	0.005	2.505	1458.698	335.016	0.230
5	4.892	0.001	0.005	3.364	0.002	0.006	3.821	1753.631	418.772	0.239
6	5.837	0.001	0.004	3.021	0.002	0.006	4.130	1941.753	500.909	0.258
7	6.770	0.001	0.003	2.610	0.002	0.008	4.950	2234.427	584.210	0.261
8	7.701	0.002	0.004	2.379	0.002	0.007	4.227	2503.626	675.187	0.270
9	8.620	0.002	0.004	2.497	0.002	0.007	4.751	2742.427	787.251	0.287
10	9.525	0.001	0.004	2.651	0.001	0.008	5.900	3099.076	887.121	0.286

Table 3.4: Mean results for 10 independent replications under the data set of size 50,000, 100,000 and 500,000. The columns show the dimension (dim), average percentage points used in SDA (obs%), RMSE of $\boldsymbol{\mu}$, $\mathbf{\Sigma}$ and computing time (time). The ratio of computing times SDA over the full data approach is also included (ratio).

3.4 Empirical study: 2015 U.S. domestic flight delays

The previous simulation studies assume a multivariate normal distribution for the data. However, this is often untrue in real data sets. The relationship between variables is often the key question to explore in many studies. Regression analysis is widely used as a tool for studying the relationship between the response variable and the observed variables. This section conducts SDA in a regression model and compares its performance with the standard approach using the full data.

We analyse flight delay data in the year 2015, which is available from the Kaggle platform (`https://www.kaggle.com/usdot/flight-delays`). The data is originally provided by the U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics. It consists of 5 million records from tracking the on-time performance of domestic flights operated by 14 large air carriers. Removing cancelled flights, there are 5,714,008 observations in total. The aim is to study the relationship between the arrival delay, departure delay and scheduled time (planned time amount). Ordinary least squares analysis shows that the variability in the residuals increases with a longer scheduled time on the logarithmic scale (results not shown here). We therefore model the heteroscedasticity as

$$\begin{aligned} y_{ij} &= \beta_{0,i} + \beta_1 x_{1i,j} + \beta_2 x_{2i,j} + \epsilon_{ij}, \\ \epsilon_{ij} &\sim N(0, \sigma_{ij}^2), \\ \log(\sigma_{ij}^2) &= \alpha_0 + \alpha_1 (x_{2i,j} - \overline{x}_{2i}), \end{aligned}$$

where y is the arrival delay (1 unit = 5 minutes); x_1 is the departure delay (1 unit = 5 minutes); x_2 is the logarithm of scheduled length of the flight (the planned amount of time for the flight; 1 unit = log(5)) [2]. For example, a flight departed 11 minutes early and arrived 22 minutes earlier than the expected arrival time. The scheduled length of the flight was 205 minutes and the actual trip time was $205 - 11 = 194$ minutes. The corresponding

[2] We include the scheduled length of flight to investigate whether the trip time is associated with an arrival delay. The result (not shown here) implies that a longer trip is less likely to have an arrival delay. The logarithm transformation is carried out because the range of length of flight is large compared to x_1 and y, which mitigates the potential influence of extreme points and makes the linear relationship more plausible.

y, x_1 and x_2 are $y = -22/5 = -4.4$, $x_1 = -11/5 = -2.2$, $x_2 = \log(205/5) \approx 3.71$. The air carrier is indexed by i, with $i = 1,\ldots, 14$ and the subscript j denotes the individual observations within each group. The term $\overline{x}_{2i}$ refers to the average value of x_2 in the ith group.

We are interested in the posterior distribution of $\boldsymbol{\theta} := \{\beta_{0,1}, \ldots, \beta_{0,14}, \beta_1, \beta_2, \alpha_0, \alpha_1\}$. The approach for obtaining an unbiased estimator of $\int_S g_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\theta})d\mathbf{x}$, the probability that $\mathbf{x}$ from a multivariate normal distribution falls in the region S, discussed in Section 3.2.2.3 cannot be adapted directly here as the assumption of joint normality of x_1, x_2, y is violated.

As $p(x_1, x_2, y|\boldsymbol{\theta}) = p(y|x_1, x_2, \boldsymbol{\theta})p(x_1, x_2)$, the part of the symbolic likelihood in (3.1) that corresponds to the integration simplifies to

$$\begin{aligned}\int_S p(x_1, x_2, y|\boldsymbol{\theta})d(.) &= \int_S p(y|x_1, x_2, \boldsymbol{\theta})p(x_1, x_2)dx_1dx_2dy \\ &= E_{x_1,x_2\sim p(x_1,x_2)}\left[\int_{s_{y,min}}^{s_{y,max}} p(y|x_1, x_2, \boldsymbol{\theta})dy\right] \\ &= \frac{1}{M}\sum_{m=1}^{M}\left(\Phi(s_{y,max}|x_1^{(m)}, x_2^{(m)}, \boldsymbol{\theta}) - \Phi(s_{y,min}|x_1^{(m)}, x_2^{(m)}, \boldsymbol{\theta})\right), \quad (3.12)\end{aligned}$$

where $x_1^{(m)}, x_2^{(m)}$ are samples from $p(x_1, x_2)$, $m = 1, \ldots, M$, $s_{y,max}, s_{y,min}$ are the upper and lower integration bounds, and $\Phi(\cdot)$ is the standard normal distribution cdf.

In (3.12), the distribution $p(x_1, x_2)$ is required, which is usually unknown. We use a finite mixture of normal distributions to approximate $p(x_1, x_2)$. Fitting a finite mixture normal distribution for multivariate data can be done in many software packages. We use the `scikit-learn` package in `Python`, using the following steps to construct symbols.

For each air carrier i:

1. Fit a K-group mixture normal distribution for $x_{1i,j}, x_{2i,j}$, $j = 1, \ldots, n_i$.
2. Predict the membership of each observation j and extract the observations falling outside $(q, 1-q)$ quantile values of $x_{1i,j}, x_{2i,j}$ based on the predictions of this group.
3. Record the hyper-rectangle S_k constructed by the $q, 1-q$ quantile values of $x_{1i,j}, x_{2i,j}$

and count the observations n_k falling in the hyper-rectangle for each group k, $k = 1, \ldots, K$.

Figure 3.3 illustrates the symbolic data representation for observations from one airline with over 115,000 observations. The hyper-rectangles at the right panel of Figure 3.3 include 62,264, 33,867 and 14,974 observations, respectively. There are 4,088 individual points to be analysed individually by our method, which is only 3.5% of the total number of observations.

A K-group mixture of normal distribution is used to approximate $p(x_1, x_2)$. The number of groups is decided by BIC by changing the number of groups K. It is challenging to do the model selection based on BIC scores as many candidate models are needed to determine which one has the best BIC score. For each air carrier, we find that a model with more than 10 groups has the lowest BIC score. However, in such cases, most groups only contain thousands of observations or even fewer, implying that SDA is unnecessary. The final numbers of groups chosen for each air carrier i are $K = 4, 4, 5, 2, 5, 2, 2, 3, 2, 2, 2, 2, 2, 2$ for $i = 1, \ldots, 14$. Our choice is based on the following 1) the group number is the first point after a sharp change in BIC scores of adjacent groups (the curve starts to become "flat enough" after the point); 2) most groups contain more than 10,000 observations, indicating the potential capability of SDA to reduce computational overhead. Figure 3.3 shows that the marginal distribution of x_2 is inaccurate. However, Table 3.5 shows that the corresponding parameter estimates are close to those obtained using the full data set.

The symbolic likelihood function for one air carrier i is (omitting the subscript i)

$$L(s; \boldsymbol{\theta}) \propto \prod_{k=1}^{K} \left[\left(\int_{S_k} p(y; x_1, x_2, \boldsymbol{\theta}) p(x_1, x_2) dP(x_1, x_2, y) \right)^{n_k} \times \mathcal{L}(\mathbf{z}_{b,k}; \boldsymbol{\theta}) \times \mathcal{L}(\mathbf{z}_{o,k}; \boldsymbol{\theta}) \right],$$

where $\mathbf{z} = \{x_1, x_2, y\}$, $\mathbf{z}_{b,k}, \mathbf{z}_{o,k}$ refer to the points falling on or outside the boundary of hyper-rectangle S_k. The Bayesian analyses on the full data and the symbolic data are performed using 20,000 MCMC iterations. After discarding the first 10,000 iterations, the posterior mean vector $\widehat{\boldsymbol{\theta}}_f$ (full data) and $\widehat{\boldsymbol{\theta}}_s$ (SDA) are compared by considering the mean

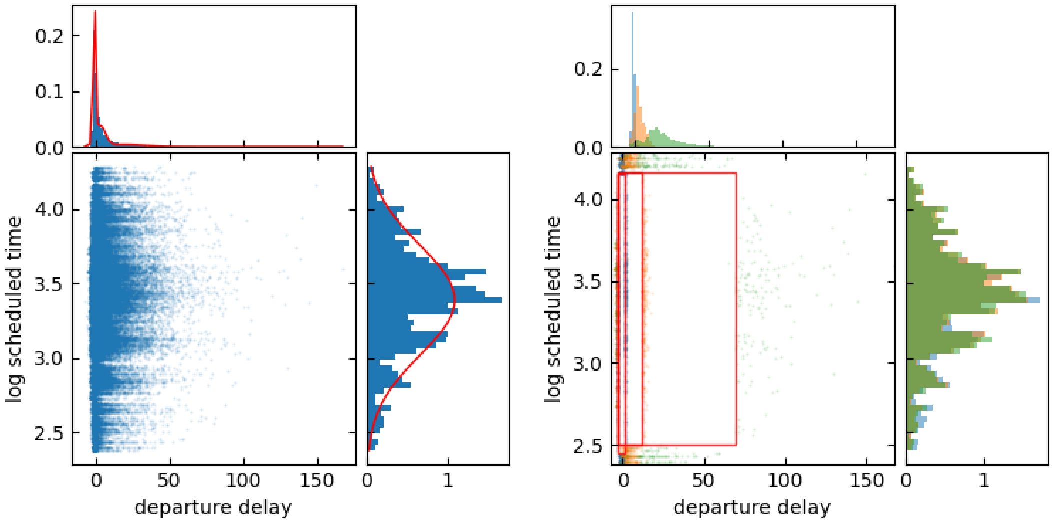


Figure 3.3: Demonstration of the symbol construction for the NK (Spirit Air Lines) airline , which has 115,193 observations. Left panel: The relationship between departure delay (x_1) and log of scheduled time (x_2). The histograms on the top and right present the marginal distribution with a density curve for fitting a 3-component bivariate mixture of normals on x_1 and x_2 jointly. Right panel: The symbol with three rectangles, obtained by setting $q = 0.01$. The top and right plots show the predictive group memberships.

absolute percentage error (MAPE) and RMSE with the definitions:

$$\text{MAPE}(\widehat{\boldsymbol{\theta}}_f, \widehat{\boldsymbol{\theta}}_s) = \frac{1}{N}\sum_{i=1}^{N}\left((|\widehat{\theta}_{f,i} - \widehat{\theta}_{s,i}|)/|\widehat{\theta}_{f,i}|\right),$$

$$\text{RMSE}(\widehat{\boldsymbol{\theta}}_f, \widehat{\boldsymbol{\theta}}_s) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\widehat{\theta}_{f,i} - \widehat{\theta}_{s,i})^2},$$

where $N = 18$ (the number of total parameters).

Table 3.5 shows both metrics as well as the computing time. By increasing the value of q from 0.005 to 0.10, both metrics decrease gradually (the symbolic data become a better approximation of the full data) with a rise in computing time (more individual observations are involved with an increased value of q). A larger value of q increases the similarity of the results to that of the full data approach. It is unclear how to choose an optimal value for q to achieve a principled trade-off between accuracy and computation. According to these results, even an extremely small q ($q = 0.005$) suffices as the MAPE is only 8%, but is about 12 times faster in terms of computing time. In the simulation study (results not shown here), if the underlying distribution of $p(x_1, x_2)$ is known or can be closely approximated by a normal distribution, a small value of q gives satisfactory results, for example, $q \leq 0.025$.

q	obs%	RMSE	MAPE	time(secs)			ratio
				prep	MCMC	total	
full	100.0	0.0	0.0	0.0	11529.74	11529.74	1.0
0.005	2.17	0.17	0.08	31.71	934.05	965.76	11.94
0.01	3.8	0.15	0.07	39.48	1453.83	1493.31	7.72
0.025	9.15	0.15	0.06	39.68	2163.29	2202.98	5.23
0.05	17.64	0.12	0.05	39.82	3378.86	3418.68	3.37
0.1	33.9	0.07	0.03	40.44	5664.21	5704.64	2.02

Table 3.5: Mean, variance and execution time of 20,000 MCMC iterations on the full data and the symbolic data with various quantile cut-offs (q). The “obs%” column shows the percentage points used in SDA. “Prep” includes the time of fitting the mixture of normals and the symbol construction. “MCMC” is the time spent on running 20,000 iterations. The last column is the ratio of running times for the full data over that of the symbolic data.

Evaluating (3.12) can be more computationally costly compared with the direct likelihood evaluation when the number of Monte Carlo draws M is equal to the number of observations. A threshold has to be determined to guarantee that SDA takes less time than analysing the full data. The left panel of Figure 3.4 explores the choice of threshold by comparing computing time for both approaches. The SDA approach requires significantly longer computing time given the same number of observations (Monte Carlo draws). From the plot, the evaluating the likelihood for 10,000 observations takes about the same time as evaluating the integral with 4,000 Monte Carlo samples. The optimal σ^2_{opt} for each block can be satisfied by updating one out of 4,000 blocks, which guarantees an efficient run of the *signed block PMMH* algorithm. Hence, we set the threshold as 10,000 and select the number Monte Carlo estimates to be 4,000.

Another concern is that SDA does not guarantee to give consistent estimators. The middle and right panels of Figure 3.4 show the posterior distributions for $\beta_{0,8}$ (air line NK) and β_1 using different qs. Although the posterior distributions of β_1 are not the same, their difference is negligible. By increasing q to 0.5, the symbolic data is reduced to the full data. However, as mentioned above, a large value of q results in the loss of computational advantage of our method.

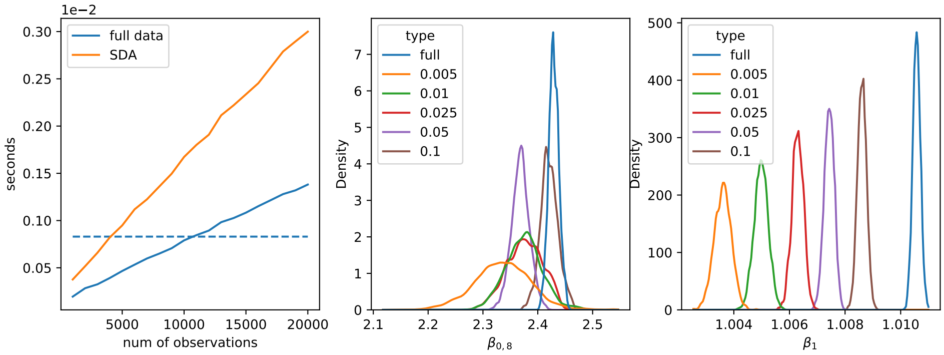


Figure 3.4: Left panel: Time spent on evaluating of the integral part of the symbolic and classical likelihoods. The result is the average time of 100 replications of evaluating the same likelihood function 10 times. Middle panel: The posterior densities of $\beta_{0,8}$ (which is the individualised intercept for air line NK, with the symbol construction demonstrated in Figure 3.3). Right panel: The posterior densities of β_1 under different quantile values of q.

3.5 Conclusions and discussion

We extend the likelihood-based SDA framework in Beranger et al. (2018) by proposing novel symbol construction approaches. This chapter shows that it is possible of applying PM to SDA by proposing exact and approximate computing methods. The approximate method is nearly unbiased and is much faster than the exact approach. Much of the focus is given to controlling the variance of difference in the logarithm of the estimator using the *signed block PMMH* algorithm. By using blocking, estimators that are more variable can be used. The computing time of our proposal is significantly less than that for the full data, with a tolerable difference in the accuracy. The proposed method is also useful when an intractable likelihood function is involved in Bayesian analysis, for example, the doubly intractable problem (Park and Haran, 2018).

The likelihood-based SDA approach provides an elegant framework for inference in large data sets. We believe that we have just scratched the surface of the field. There are several open questions to be investigated in future research. In the symbolic data context, this chapter uses $q, 1-q$ quantiles for a specific symmetric distribution (a multivariate normal distribution). For high-dimensional skewed distributions, constructing symbols that give a better representation remains an open question. For the symbol construction, numerical evidence from simulations and the empirical study indicates that a small value of q provides good results. Although the pragmatic approach works well, understanding how to tune the hyperparameters optimally is important, as their values have a large impact on both the symbol construction and the variance of the logarithm of the likelihood estimator. Regarding the symbol construction method, Beranger et al. (2018) propose alternatives including “marginal only”, “sequential nesting” and “iterative segmentation” in addition to min-max intervals. These methods are worth further investigation. A possible improvement can be the incorporation of precise locations of boundary points in hyper-rectangles into symbols. A more general question for SDA is providing a principled framework for symbol construction which trades off accuracy and computing time. To implement the PM method, we first transform the likelihood function into its logarithm, find a (nearly) unbiased estimator and then transform back to its original scale to circumvent a direct evaluation of the likelihood function. However, the path sampler requires a large amount of computing time. Exploring more efficient methods of getting such an estimator is left for future research.

Chapter 4

A novel pseudo-marginal approach to doubly intractable problems using the block-Poisson estimator

4.1 Introduction

Doubly intractable problems occur in a Bayesian model when the likelihood function contains an intractable normalising constant that depends on the model parameters. Markov chain Monte Carlo (MCMC) methods (see Brooks et al., 2011, for an overview) carry out Bayesian inference for complicated models without computing the marginal likelihood, but they cannot directly be applied in this setting due to the intractability of the likelihood. Many well-known models have such a normalising constant in the likelihood function, for example, the exponential random graph models for social networks (Hunter and Handcock, 2006), the non-Gaussian Markov random field for spatial statistics, including the Ising model and its variants (Hughes et al., 2011; Ising, 1925; Lenz, 1920).

Several MCMC algorithms have recently been proposed for Bayesian inference to tackle the doubly intractable problem. These algorithms are classified into two categories, with some overlap between them. The first approach introduces auxiliary variables to cancel the normalising constant in the Metropolis-Hasting (MH) acceptance ratio (Hastings, 1970; Metropolis et al., 1953). The second approach approximates the likelihood function (including the normalising constant) and substitutes the approximation in place of the exact likelihood in the estimation. The pseudo-marginal (PM) method (Andrieu and Roberts, 2009) is often utilised when an unbiased estimator of the likelihood is available through Monte Carlo approximations. However, in some problems, including doubly intractable models, forming an unbiased estimator that is almost surely positive is prohibitively expensive (Jacob and Thiery, 2015). The so-called Russian roulette (RR) estimator (Lyne et al., 2015) is an example of a method that uses a carefully chosen geometric series to approximate the likelihood function unbiasedly. Section 4.3 further discusses existing methods including RR.

We propose an efficient method for exact inference in doubly intractable problems by utilising the approach in Lyne et al. (2015), where an unbiased, but not necessarily positive, estimator of the likelihood function is used. The algorithm targets a posterior density that uses the absolute value of the likelihood, resulting in iterates from a perturbed target density. The iterates are subsequently reweighted using importance sampling to obtain consistent estimates of the expectation of any function of the parameters with respect to the true posterior density. We refer to such a pseudo-marginal approach as signed pseudo-marginal. The approach is often combined with the MH algorithm, and called the *signed PMMH.*

Our main contribution is to explore the use of the block-Poisson (BP) estimator (Quiroz et al., 2021) in the context of estimating doubly intractable models using the *signed PMMH* approach. Compared to the RR method in Lyne et al. (2015), our method offers the following advantages. First, the BP estimator has a much simpler structure and hence it is more computationally efficient. Second, the block form of our estimator makes it straightforward to correlate the estimators of the doubly intractable posterior at the current and proposed draws in the MH algorithm. Introducing such correlation dramatically improves the efficiency of PM algorithms (Deligiannidis et al., 2018; Tran et al., 2016). Finally, under some simplifying assumptions, the logarithm of the absolute value of our estimator has a closed form expression that can be used to derive heuristic guidelines to optimally tune its hyperparameters. We demonstrate empirically that our method outperforms that proposed in Lyne et al. (2015) when estimating the Ising model. To the best of our knowledge, our method and that of Lyne et al. (2015) are the only alternatives in the PM framework to perform exact inference (in the sense of consistent estimates of posterior expectations) for general doubly intractable problems. Compared with algorithms which use auxiliary variables to avoid evaluating the normalising constant, the *signed PMMH* is more widely applicable and generic as it does not require exact sampling from the likelihood.

The chapter is organised as follows. Section 4.2 formally introduces the problem and Section 4.3 discusses previous research. Section 4.4 consists of three parts. The first part describes the BP estimator. The second part introduces the signed block pseudo-marginal Metropolis-Hastings algorithm with the BP estimator (*signed block PMMH with BP*). The third part establishes guidelines for tuning the hyperparameters in the *signed block PMMH with BP*. Section 4.5 demonstrates the proposed method in three simulation studies: the Ising model, the constrained Gaussian process and the Kent distribution. Section 4.6 analyses four data sets using the Kent distribution. The group of each sample is known. The group is either a collection place or a sample processing procedure. Cross-validation is used to provide the overall prediction accuracy of the group identity for each data set. Section 4.7 concludes.

4.2 Doubly intractable problems

Let $p(\mathbf{y}|\boldsymbol{\theta})$ denote the density of the observation vector $\mathbf{y}$, where $\boldsymbol{\theta}$ is the parameter vector. Suppose $p(\mathbf{y}|\boldsymbol{\theta}) = f(\mathbf{y}|\boldsymbol{\theta})/Z(\boldsymbol{\theta})$, where $f(\mathbf{y}|\boldsymbol{\theta})$ is computable while the normalising constant $Z(\boldsymbol{\theta})$ is not. The reason that $Z(\boldsymbol{\theta})$ is intractable may be that it is computationally expensive or it does not have an analytical form. Two examples are given below to demonstrate this intractability for both discrete and continuous observations $\mathbf{y}$.

Example 1. *The Ising model (Ising, 1925).*

Consider an $L \times L$ lattice with binary observation $y_{ij} \in \{-1, 1\}$ in row i and column j. The likelihood of $\theta \in \mathbb{R}$ is

$$p(\mathbf{y}|\theta) = \frac{1}{Z(\theta)} \exp(\theta S(\mathbf{y})); \quad S(\mathbf{y}) = \sum_{i=1}^{L} \sum_{j=1}^{L-1} y_{i,j} y_{i,j+1} + \sum_{i=1}^{L-1} \sum_{j=1}^{L} y_{i,j} y_{i+1,j}; \tag{4.1}$$

$$\text{with } Z(\theta) = \sum_{\mathbf{y}} \exp(\theta S(\mathbf{y})).$$

The normalising constant $Z(\theta)$ in the Ising model is a sum over 2^{L^2} $S(\mathbf{y})$ terms, making it computationally intractable even for moderate values of L. This example is discussed in Section 4.5.1.

Example 2. *The Kent distribution (Kent, 1982).*

The density of the Kent distribution for $\mathbf{y} \in \mathbb{R}^3, \|\mathbf{y}\| = 1$, is

$$f(\mathbf{y}|\boldsymbol{\gamma_1}, \boldsymbol{\gamma_2}, \boldsymbol{\gamma_3}, \beta, \kappa) = \frac{1}{c(\kappa, \beta)} \exp\left\{\kappa \boldsymbol{\gamma_1}^\top \cdot \mathbf{y} + \beta \left[(\boldsymbol{\gamma_2}^\top \cdot \mathbf{y})^2 - (\boldsymbol{\gamma_3}^\top \cdot \mathbf{y})^2\right]\right\}; \tag{4.2}$$

$$\text{with } c(\kappa, \beta) = 2\pi \sum_{j=0}^{\infty} \frac{\Gamma(j+0.5)}{\Gamma(j+1)} \beta^{2j} (0.5\kappa)^{-2j-0.5} I_{2j+0.5}(\kappa),$$

where $I_\nu(.)$ is the modified Bessel function and $\boldsymbol{\gamma_1}, \boldsymbol{\gamma_2}, \boldsymbol{\gamma_3}$ form a set of 3-dimensional orthonormal vectors. The normalising constant $c(\kappa, \beta)$ is an infinite sum. Section 4.5.3 covers this example.

Let $\pi(\boldsymbol{\theta})$ be the prior for $\boldsymbol{\theta}$. To see why MCMC sampling is difficult for models with an intractable normalising constant, note that the posterior of $\boldsymbol{\theta}$ is

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{Z(\boldsymbol{\theta})p(\mathbf{y})}. \tag{4.3}$$

If we use a MH proposal $q(\boldsymbol{\theta}'|\boldsymbol{\theta})$, then the acceptance probability of $\boldsymbol{\theta}'$ is

$$\alpha(\boldsymbol{\theta}', \boldsymbol{\theta}) = \min\left\{1, \frac{\pi(\boldsymbol{\theta}')f(\mathbf{y}|\boldsymbol{\theta}')/Z(\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta})f(\mathbf{y}|\boldsymbol{\theta}))/Z(\boldsymbol{\theta})} \times \frac{q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{q(\boldsymbol{\theta}'|\boldsymbol{\theta})}\right\}, \tag{4.4}$$

which is computationally intractable because $Z(\boldsymbol{\theta})/Z(\boldsymbol{\theta}')$ is unknown.

4.3 Previous research

Previous research on doubly intractable problems is mainly divided into the auxiliary variable approach and the likelihood approximation approach; see Park and Haran (2018) for an excellent review of both approaches. The auxiliary variable approach cleverly chooses the joint transition kernel of the parameters and the auxiliary variables so that the normalising constant cancels in the MH acceptance ratio. The most well-known algorithms are the exchange algorithm (Murray et al., 2006) and the auxiliary variable method (Møller et al., 2006). Both algorithms are model dependent and rely on the sampling technique to draw observations from the likelihood function. Perfect sampling (Propp and Wilson, 1996) is often used to generate samples from the model without knowing the normalising constant. However, for some complex models, such as the Ising model on a large grid, perfect sampling is prohibitively expensive. To overcome such issues, Liang (2010) and Liang et al. (2016) relax the requirement of exact sampling and propose the double MH sampler and the adaptive exchange algorithm. However, the former generates inexact inference results and the latter suffers from memory issues as many intermediate variables need to be stored within each iteration.

The likelihood approximation approach approximates the likelihood function, often so that the corresponding algorithm leads to a simulation consistent result, which means that the posterior mean of any function of the parameters is a consistent estimator of that function. Atchadé et al. (2013) directly approximate $Z(\boldsymbol{\theta})$ through multiple importance sampling. Their approach also depends on an auxiliary variable, but does not require perfect sampling. The downside is similar to that of the adaptive exchange algorithm; generally, a large memory is required to store the intermediate variables generated in each iteration. An alternative method is to approximate $1/Z(\boldsymbol{\theta})$ directly and use the *signed PMMH* algorithm to replace the likelihood function by an unbiased estimator as proposed in Lyne et al. (2015). To obtain the unbiased estimator, $1/Z(\boldsymbol{\theta})$ is expressed as a geometric series which is truncated using an RR approach. The RR method first appears in the physics literature (Carter and Cashwell, 1975) and is useful for obtaining an unbiased estimator through a finite time stochastic truncation of the infinite series. To implement RR, a tight upper bound for $Z(\boldsymbol{\theta})$ is required, otherwise the convergence of the geometric series is slow and makes the algorithm inefficient. In practice, an upper bound is usually unavailable, which may lead to negative estimates of the likelihood, and thus a *signed PMMH* approach is necessary, although it inflates the asymptotic variance of the MCMC chain (Andrieu and Vihola, 2016)[1] compared to a standard PM approach, especially if the estimator produces a significant proportion of negative estimates (Lyne et al., 2015). It is therefore crucial to quantify the probability of a negative estimate when tuning the hyperparameters of the estimator, which is difficult for the RR estimator. In contrast, our estimator is more tractable and the probability of a positive estimate is analytically derived under simplifying assumptions. Besides the upper bound, a few other hyperparameters of the RR estimator need to be determined and guidelines have not been established due to its intractability. Wei and Murray (2017) combine RR with Markov chain coupling to produce an estimator with lower variance and a larger probability of producing positive estimates. However, their estimator is still too intractable to derive optimal tuning guidelines.

The *signed PMMH* requires an unbiased likelihood estimator, so obtaining an efficient estimator of the normalising constant is important, which is usually hard for complex models. If such an estimator is normally distributed, then a bias-corrected approximation of the likelihood in combination with an auxiliary variable can be constructed to generate simulation consistent results with faster execution time (Ceperley and Dewing, 1999; Quiroz et al., 2019). In the simulation study of Section 4.5.1, we compare our proposed method to the exchange algorithm and the RR method. We also consider the bias-corrected estimator.

[1] The asymptotic variance of the function $\psi(\boldsymbol{\theta})$ with $\boldsymbol{\theta}$ the trajectories from an MCMC chain $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots\}$, with respect to the posterior distribution π, is defined as

$$\mathrm{Var}(\psi(\boldsymbol{\theta})) + 2\sum_{\tau=1}^{\infty} \mathrm{Cov}_{\tau}(\psi(\boldsymbol{\theta})),$$

where $\mathrm{Cov}_{\tau}(\cdot)$ is the covariance of two iterates that are τ steps part in the chain.

4.4 Methodology

4.4.1 The block-Poisson estimator

Before introducing the *signed PMMH* algorithm, we present the BP estimator, proposed by Quiroz et al. (2021) for estimating $\exp(B)$ unbiasedly, assuming that $E(\widehat{B}) = B$, where $\widehat{B}$ is an unbiased estimator of the log-likelihood B in the data-subsampling context. The BP estimator is formed using blocks of Poisson estimators (Papaspiliopoulos, 2011), to allow for correlation between adjoining iterates in the PM algorithm, as described in Section 4.4.2. Similarly to the likelihood approximation approaches discussed above, the BP estimator is implemented in combination with an auxiliary variable ν, and the estimator of the normalising constant. Omitting many details, assume $B(\boldsymbol{\theta}) = -\nu Z(\boldsymbol{\theta})$. Given ν and an unbiased estimator of $Z(\boldsymbol{\theta})$, the BP estimator produces an unbiased estimator of $\exp(-\nu Z(\boldsymbol{\theta}))$. The BP estimator requires a lower bound for $B(\boldsymbol{\theta})$ to guarantee its positiveness. The BP estimator is more likely to be positive than the RR estimator. This section describes the BP estimator $\widehat{L}_B$ in Definition 1. Lemma 1 gives the expectation and variance of $\widehat{L}_B$. Lemmas 2 and 3 establish useful results for hyperparameter tuning of the algorithm (see Section 4.4.3). Appendix B.1 contains the proofs of all the lemmas.

Definition 1. *The block-Poisson estimator (Quiroz et al., 2021) is defined as*

$$\widehat{L}_B(\boldsymbol{\theta}) = \prod_{l=1}^{\lambda} \exp(\xi_l(\boldsymbol{\theta})), \quad \textit{where } \exp(\xi_l(\boldsymbol{\theta})) = \exp(a/\lambda + m) \prod_{h=1}^{\chi_l} \frac{\widehat{B}^{(h,l)}(\boldsymbol{\theta}) - a}{m\lambda}. \tag{4.5}$$

where m must be a positive real number; $\widehat{B}^{(h,l)}(\boldsymbol{\theta})$ are realisations of a random variable $\widehat{B}(\boldsymbol{\theta})$ with $E(\widehat{B}(\boldsymbol{\theta})) = B(\boldsymbol{\theta})$; λ is the number of blocks with $\chi_l \sim \mathrm{Pois}(m)$ and a is an arbitrary real number (it need not be positive).

Lemma 1. *Denote $\mathrm{Var}(\widehat{B}(\boldsymbol{\theta}))$ as σ_B^2, assume $\sigma_B^2 < \infty$ and $E(\widehat{B}(\boldsymbol{\theta})) = B(\boldsymbol{\theta})$. The following properties hold for $\widehat{L}_B(\boldsymbol{\theta})$,*

(i) $E(\widehat{L}_B(\boldsymbol{\theta})) = \exp(B(\boldsymbol{\theta}))$.

(ii) $\mathrm{Var}(\widehat{L}_B(\boldsymbol{\theta})) = \exp\left[\frac{(B(\boldsymbol{\theta}) - a)^2 + \sigma_B^2}{m\lambda} + 2a + m\lambda\right] - \exp(2B(\boldsymbol{\theta}))$.

(iii) $\mathrm{Var}(\widehat{L}_B(\boldsymbol{\theta}))$ *is minimised at* $a = B(\boldsymbol{\theta}) - m\lambda$*, given fixed* m *and* λ*.*

Lemma 1 shows that given an unbiased estimator $\widehat{B}(\boldsymbol{\theta})$ of $B(\boldsymbol{\theta})$, the BP estimator is unbiased for $\exp(B(\boldsymbol{\theta}))$. Part (iii) of Lemma 1 suggests that we can choose the lower bound $a = \widehat{B}(\boldsymbol{\theta}) - m\lambda$, as $B(\boldsymbol{\theta})$ is unknown. Here a uses $\widehat{B}(\boldsymbol{\theta})$, a realised estimate of $B(\boldsymbol{\theta})$, which is estimated independently of $\widehat{B}^{(h,l)}(\boldsymbol{\theta})$. Similarly to the RR estimator, the downside of the BP estimator is that it is not necessarily positive all the time. Fortunately, by having a relatively large $m\lambda$, the sufficient condition for $\widehat{L}_B(\boldsymbol{\theta}) \geq 0$ is likely to be satisfied. Conversely, it is computationally costly as a large $m\lambda$ value implies many products in the BP estimator. Here we follow Quiroz et al. (2021) and advocate the use of a soft lower bound, i.e., one that may lead to negative estimates, but still gives a $\Pr(\widehat{L}_B(\boldsymbol{\theta}) \geq 0)$ close to one. The probability $\Pr(\widehat{L}_B(\boldsymbol{\theta}) \geq 0)$ is analytically tractable, with details provided in Lemma 2. It is crucial to have this probability close to one for the algorithm to be efficient.

Lemma 2.

$$\Pr(\widehat{L}_B(\boldsymbol{\theta}) \geq 0) = \frac{1}{2}\Big(1 + (1 - 2\Psi(a, m, M))^{\lambda}\Big),$$

with $\Psi(a, m, M) = Pr(\xi < 0) = \frac{1}{2}\sum_{j=1}^{\infty}\left(1 - (1 - 2\Pr(A_m \leq 0))^j\right)\Pr(\chi_l = j)$*,* $\chi_l \sim \mathrm{Pois}(m)$ *and* $A_m = [\widehat{B}(\boldsymbol{\theta}) - B(\boldsymbol{\theta})]/(m\lambda) + 1$*.*

Lemma 3. *If* $\widehat{B}^{(h,l)}(\boldsymbol{\theta}) \overset{\mathrm{iid}}{\sim} N(B(\boldsymbol{\theta}), \sigma_B^2)$ *for all* h *and* l*, when* $a = B(\boldsymbol{\theta}) - m\lambda$*, the variance of* $\log|\widehat{L}_B|$ *is*

$$\sigma^2_{\log|\widehat{L}_B|} = m\lambda(\nu_B^2 + \eta_B^2)$$

where

$$\eta_B = \log(\sigma_B/(m\lambda)) + 0.5\left(\log 2 + E_J(\psi^{(0)}(0.5 + J))\right)$$

and

$$\nu_B^2 = 0.25\left(E_J(\psi^{(1)}(0.5 + J)) + \mathrm{Var}_J(\psi^{(0)}(0.5 + J))\right)$$

where $J \sim \mathrm{Pois}((m\lambda)^2/(2\sigma_B^2))$ *and* $\psi^{(q)}$ *is the polygamma function of order* q*.*

Lemma 3 derives the variance of the logarithm of the absolute value for the block-Poisson estimator by assuming a normal distribution for $\widehat{B}^{(h,l)}(\boldsymbol{\theta})$. This result is used for hyperparameter tuning in Section 4.4.3.

4.4.2 The *signed block PMMH with BP* algorithm

This section first incorporates the BP estimator into the PMMH algorithm to cope with the doubly intractable problem. To deal with the negative estimates, a signed correction is placed in the *signed PMMH* algorithm. The *signed block PMMH with BP* algorithm is then presented, and its performance demonstrated in the simulation studies and the empirical data sets in this chapter.

Following Lyne et al., an auxiliary variable $\nu \sim \text{Expon}(Z(\boldsymbol{\theta}))$ is introduced. The joint distribution of $\boldsymbol{\theta}$ and the auxiliary variable ν is

$$\begin{aligned}\pi(\boldsymbol{\theta},\nu|\mathbf{y}) &= Z(\boldsymbol{\theta})\exp(-\nu Z(\boldsymbol{\theta}))\frac{f(\mathbf{y}|\boldsymbol{\theta})}{Z(\boldsymbol{\theta})}\pi(\boldsymbol{\theta})\frac{1}{p(\mathbf{y})}\\ &= \exp(-\nu Z(\boldsymbol{\theta}))f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\frac{1}{p(\mathbf{y})}.\end{aligned}$$

Sampling $\nu \sim \text{Expon}(Z(\boldsymbol{\theta}))$ directly is impractical as we never know the true value of $Z(\boldsymbol{\theta})$. In the implementation, we draw $\nu \sim \text{Expon}(\widehat{Z}(\boldsymbol{\theta}))$, where $\widehat{Z}(\boldsymbol{\theta})$ is an unbiased estimator for $Z(\boldsymbol{\theta})$. Note that ν depends on $\boldsymbol{\theta}$ implicitly, which could be written as $\nu(\boldsymbol{\theta})$. In this chapter, we omit the dependence of ν on $\boldsymbol{\theta}$ and write $\nu(\boldsymbol{\theta})$ as ν for simplicity.

The *signed PMMH* algorithm

Suppose an unbiased, but not necessarily positive, estimator $\widehat{\exp}(-\nu Z(\boldsymbol{\theta}))$ of $\exp(-\nu Z(\boldsymbol{\theta}))$, is available; e.g., the BP estimator in (4.5). We also write the estimator as $\widehat{\exp}(-\nu Z(\boldsymbol{\theta})|\mathbf{u})$, where $\mathbf{u}$ is a set of random numbers with density $p(\mathbf{u})$. The unbiasedness of $\widehat{\exp}(-\nu Z(\boldsymbol{\theta})|\mathbf{u})$ means that

$$\exp(-\nu Z(\boldsymbol{\theta})) = \int_{\mathbf{u}} \widehat{\exp}(-\nu Z(\boldsymbol{\theta})|\mathbf{u})p(\mathbf{u})d\mathbf{u}.$$

The corresponding posterior distribution is

$$\widehat{\pi}(\boldsymbol{\theta}, \mathbf{u}, \nu|\mathbf{y}) = \widehat{\exp}(-\nu Z(\boldsymbol{\theta})|\mathbf{u}) p(\mathbf{u}) f(\mathbf{y}|\boldsymbol{\theta}) \pi(\boldsymbol{\theta}) \frac{1}{p(\mathbf{y})}.$$

However, the distribution above is not a valid target as the estimate $\widehat{\exp}(-\nu Z(\boldsymbol{\theta})|\mathbf{u})$ can be negative. We follow the *signed PMMH* algorithm in Lyne et al. and replace the likelihood estimate with its absolute value in the MH acceptance ratio. Then, the posterior distribution on $\boldsymbol{\theta}$, $\mathbf{u}$ and ν is

$$|\widehat{\pi}(\boldsymbol{\theta}, \mathbf{u}, \nu|\mathbf{y})| = |\widehat{\exp}(-\nu Z(\boldsymbol{\theta})|\mathbf{u})| p(\mathbf{u}) f(\mathbf{y}|\boldsymbol{\theta}) \pi(\boldsymbol{\theta}) \frac{1}{p(\mathbf{y})}.$$

The MCMC iterates are reweighed using importance sampling to obtain a consistent estimate of the expectation of an arbitrary function $\psi(\boldsymbol{\theta})$ with respect to the posterior density $\pi(\boldsymbol{\theta}|\mathbf{y})$. The expectation is computed as follows,

$$\begin{aligned} E(\psi(\boldsymbol{\theta})|\mathbf{y}) &= \int_{\boldsymbol{\theta}} \psi(\boldsymbol{\theta}) \pi(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\theta}} \psi(\boldsymbol{\theta}) \int_{\mathbf{u}} \int_{\nu} \pi(\boldsymbol{\theta}, \mathbf{u}, \nu|\mathbf{y}) d\nu d\mathbf{u} d\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\theta}} \psi(\boldsymbol{\theta}) \int_{\mathbf{u}} \int_{\nu} \widehat{\pi}(\boldsymbol{\theta}, \mathbf{u}, \nu|\mathbf{y}) \mathrm{sign}(\widehat{\pi}(\mathbf{y}|\boldsymbol{\theta}, \mathbf{u}, \nu)) d\nu d\mathbf{u} d\boldsymbol{\theta}, \end{aligned}$$

where $\mathrm{sign}(x) = 1$ if $x > 0$; $\mathrm{sign}(x) = -1$ if $x < 0$.

The function $\psi(\boldsymbol{\theta})$ is independent of ν, storing only $\boldsymbol{\theta}^{(i)}$ and the sign of likelihood estimate evaluated at the accepted $\boldsymbol{\theta}^{(i)}, \nu^{(i)}, \mathbf{u}^{(i)}$ at the ith iterate. The final estimate of the expectation is

$$\widehat{E}_{\pi}(\psi(\boldsymbol{\theta})) = \frac{\sum_{i=1}^{N} \psi(\boldsymbol{\theta}^{(i)}) s^{(i)}}{\sum_{i=1}^{N} s^{(i)}}, \tag{4.6}$$

where $s^{(i)} = \mathrm{sign}(\widehat{\pi}(\mathbf{y}|\boldsymbol{\theta}^{(i)}, \mathbf{u}^{(i)}, \nu^{(i)}))$.

The signed block PMMH with the BP algorithm

Correlating the estimators at the current and proposed draws decreases the variability of the difference of the log likelihoods, which brings a well-documented substantial advantage over the standard PMMH (Deligiannidis et al., 2018; Tran et al., 2016). We follow the approach in Tran et al. (2016), where the correlation is induced by blocking the random numbers and updating one of the blocks in the evaluation of the likelihood at the proposal while keeping the rest of the blocks fixed. In the BP estimator, we use the random number u_l to estimate ξ_l, $l = 1, \dots, \lambda$ and group them as $\mathbf{u} = (u_1, \dots, u_\lambda) = u_{1:\lambda}$. Note that each u_l may include random numbers of different sizes depending on the realised $\chi_l \sim \text{Pois}(m)$.

At successive MCMC iterates, only one block is updated when obtaining the likelihood estimator at the proposed draw. Given the number of blocks λ is sufficiently large, the correlation ρ between the log of the likelihood estimators evaluated at the current and the proposed draw is approximately $1 - 1/\lambda$ (Quiroz et al., 2021). We can adjust the number of blocks to produce a pre-specified correlation between the log of likelihood estimates.

Algorithm 7 the *signed block PMMH with BP* for one iteration

```
1: **Input:** Current values of $\nu, \boldsymbol{\theta}, u_{1:\lambda}$.
2: **Output:** Updated values of $\nu, \boldsymbol{\theta}, u_{1:\lambda}$ and sign($\widehat{\pi}(y|\boldsymbol{\theta}, u_{1:\lambda}, \nu)$.
3: Generate $u'_{1:\lambda} \leftarrow u_{1:\lambda}$ by updating one block of random numbers.
4: Generate $\boldsymbol{\theta}'$ by $q(\boldsymbol{\theta}'|\boldsymbol{\theta})$.
5: Estimate $\widehat{Z}(\boldsymbol{\theta}')$ and generate $\nu'$ from an exponential distribution with mean $\widehat{Z}(\boldsymbol{\theta}')$ :
   $q(\nu'|\boldsymbol{\theta}') = \widehat{Z}(\boldsymbol{\theta}') \exp(-\nu'\widehat{Z}(\boldsymbol{\theta}'))$
6: Set $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}'$, $\nu \leftarrow \nu'$ and $u_{1:\lambda} \leftarrow u'_{1:\lambda}$ with the probability:
   $$\min\left\{1, \frac{|\widehat{\pi}(\boldsymbol{\theta}', \nu'|\mathbf{y}, u'_{1:\lambda})|}{|\widehat{\pi}(\boldsymbol{\theta}, \nu|\mathbf{y}, u_{1:\lambda})|} \frac{q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{q(\boldsymbol{\theta}'|\boldsymbol{\theta})} \frac{\widehat{Z}(\boldsymbol{\theta})}{\widehat{Z}(\boldsymbol{\theta}')} \frac{\exp(-\nu\widehat{Z}(\boldsymbol{\theta}))}{\exp(-\nu'\widehat{Z}(\boldsymbol{\theta}'))}\right\} \tag{4.7}$$
   where $\widehat{\pi}(\boldsymbol{\theta}, \nu|\mathbf{y}, u_{1:\lambda}) = \widehat{\exp}(-\nu Z(\boldsymbol{\theta})|u_{1:\lambda}) f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})p^{-1}(\mathbf{y})$ and $\widehat{\exp}(-\nu Z(\boldsymbol{\theta})|u_{1:\lambda})$ is estimated by the BP estimator.
7: Record $s = \text{sign}(\widehat{\pi}(\mathbf{y}|\boldsymbol{\theta}, u_{1:\lambda}, \nu))$ which is equivalent to the sign of $\widehat{\exp}(-\nu Z(\boldsymbol{\theta})|u_{1:\lambda})$.
       Note: sign($x$) = 1 if $x > 0$; sign($x$) = −1 if $x < 0$.
```

Algorithm 7 outlines the *signed block PMMH with BP*. For Line 5 in Algorithm 7, we recommend estimating $\widehat{Z}$ by reusing the $\widehat{Z}$ obtained in the BP estimator.

To understand this, rewrite (4.7) as

$$\frac{\pi(\boldsymbol{\theta}')f(\mathbf{y}|\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta})f(\mathbf{y}|\boldsymbol{\theta})} \times \frac{q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{q(\boldsymbol{\theta}'|\boldsymbol{\theta})} \times \frac{\widehat{Z}^{-1}(\boldsymbol{\theta}')}{\widehat{Z}^{-1}(\boldsymbol{\theta})} \times \frac{|\widehat{\exp}(-\nu' Z(\boldsymbol{\theta}')|u'_{1:\lambda})|/\exp(-\nu'\widehat{Z}(\boldsymbol{\theta}'))}{|\widehat{\exp}(-\nu Z(\boldsymbol{\theta})|u_{1:\lambda})|/\exp(-\nu\widehat{Z}(\boldsymbol{\theta}))}.$$

The equation above indicates the term

$$\frac{|\widehat{\exp}(-\nu' Z(\boldsymbol{\theta}')|u'_{1:\lambda})|/\exp(-\nu'\widehat{Z}(\boldsymbol{\theta}'))}{|\widehat{\exp}(-\nu Z(\boldsymbol{\theta})|u_{1:\lambda})|/\exp(-\nu\widehat{Z}(\boldsymbol{\theta}))}$$

corrects the induced bias in $\widehat{Z}^{-1}(\boldsymbol{\theta}')/\widehat{Z}^{-1}(\boldsymbol{\theta})$. As $Z^{-1}(\boldsymbol{\theta}')/Z^{-1}(\boldsymbol{\theta})$ is intractable, it is desirable to decrease the variability in $\widehat{Z}(\boldsymbol{\theta})$ and $\widehat{Z}(\boldsymbol{\theta}')$. Conversely, estimating $\widehat{Z}(\boldsymbol{\theta})$ with low variability may cost extra computational resources which in turn increases the computing time. Hence, we recommend constructing $\widehat{Z}(\boldsymbol{\theta})$ by taking the average value of $\widehat{Z}(\boldsymbol{\theta})$s used in the BP estimator. The unbiasedness property of the BP estimator is still preserved under this construction given a provided ν and the availability of unbiased estimate $\widehat{Z}(\boldsymbol{\theta})$s of $Z(\boldsymbol{\theta})$, and the computing cost involved in obtaining the average is negligible.

Equation (4.6) provides the final estimate of $\psi(\boldsymbol{\theta})$. Quiroz et al. (2021) show that a significant proportion of negative likelihood estimates inflate the asymptotic variance. The worst case occurs when half of the estimates are negative; the expectation is then unbounded because of the zero in the denominator.

4.4.3 Tuning the *signed block PMMH with BP* algorithm

Pitt et al. (2012) provide guidelines to tune the number of particles in a PM algorithm for an optimal trade-off between computing time and MCMC efficiency measured by the integrated autocorrelation time (IACT); the IACT is the sum of the autocorrelation functions of the MCMC iterates of $\psi(\boldsymbol{\theta})$ (after convergence) lagged from zero to infinity (Roberts and Rosenthal, 2009). Quiroz et al. (2021) extend these guidelines to cases when the likelihood estimator is not necessarily positive. The derivation of our guidelines follow those in Quiroz et al. (2021), with modifications that account for our (different) estimator.

Following Section 4.3 of Quiroz et al. (2021), the optimal hyperparameters minimise the computational time (CT) of the algorithm, which can be viewed as a trade-off between computing cost and the inefficiency factor (IF). The simplified expression of CT in Quiroz et al. (2021) is

$$CT = m\lambda M \frac{IF_{|\widehat{\pi}|,\psi s}\left(\sigma^2_{\log|\widehat{L}_B|}(m,\lambda,M|\gamma)\right)}{(2\tau(m,\lambda,M)-1)^2}. \tag{4.8}$$

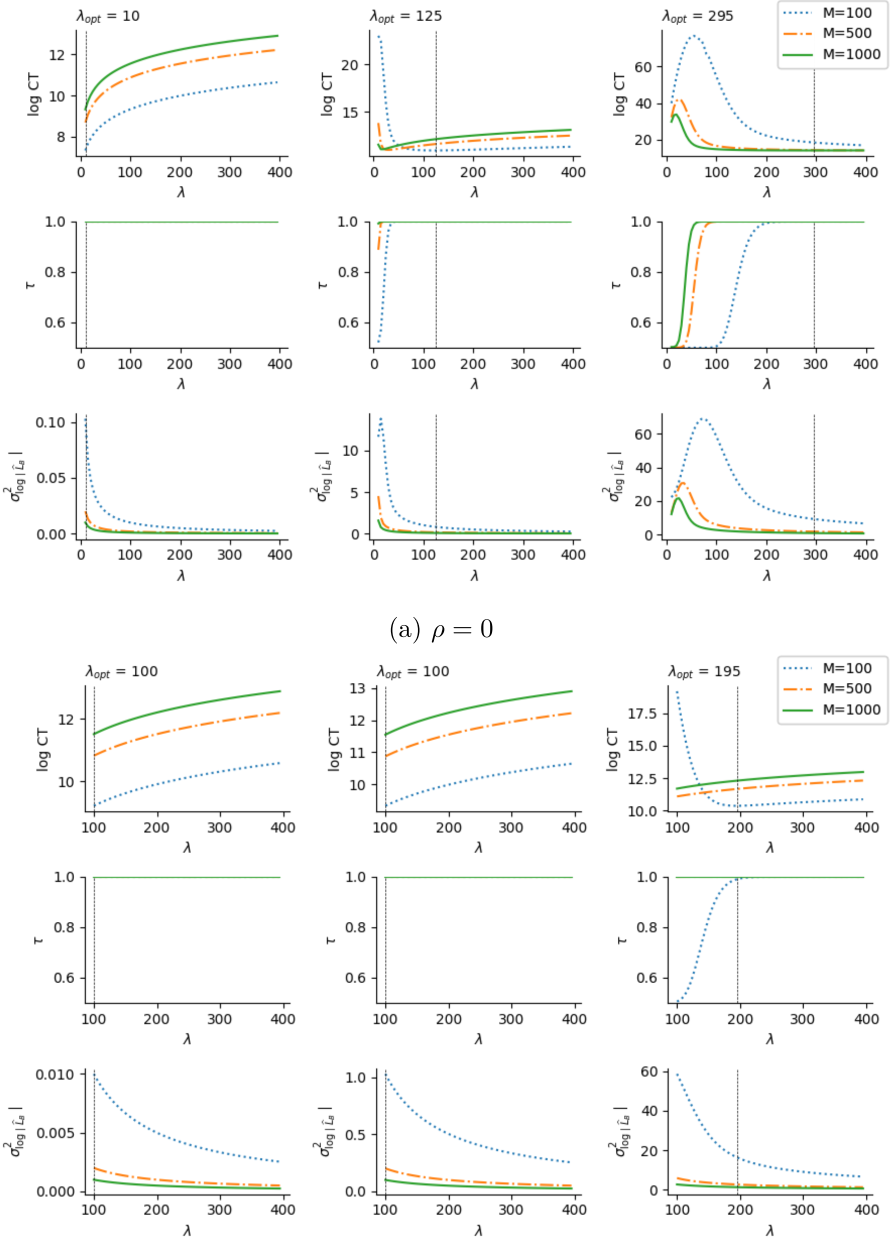
The first term $m\lambda M$ is proportional to the expected cost per iteration since there are λ blocks in total and each block includes m estimates on average with M Monte Carlo samples in each.

The numerator in (4.8) is the inefficiency factor (IF), which measures the MCMC sampling efficiency of drawing ψs from the targeted distribution $|\widehat{\pi}|$. The IF is implicitly determined by the variance of the log of the absolute likelihood estimate $\sigma^2_{\log|\widehat{L}_B|}$, which in turn depends on the hyperparameters m, λ, M. Section S2 of Quiroz et al. (2021) defines and derives IF. The IF evaluation in (4.8) requires $\gamma(\boldsymbol{\theta})$, defined as $\gamma(\boldsymbol{\theta}) = M\text{Var}(-\nu\widehat{Z_M}(\boldsymbol{\theta}))$, provided that the estimator of $Z(\boldsymbol{\theta})$ is obtained by Monte Carlo integration using M particles. Note that $\gamma(\boldsymbol{\theta})$ is the estimator's variance, which does not depend on M. The term $\gamma(\boldsymbol{\theta})$ is decomposed as

$$\gamma(\boldsymbol{\theta}) = M\text{Var}(-\nu\widehat{Z_M}(\boldsymbol{\theta})) = M\text{Var}\left(-\frac{\log(u)}{Z(\boldsymbol{\theta})}\widehat{Z_M}(\boldsymbol{\theta})\right) = M\log(u)^2\frac{\text{Var}(\widehat{Z_M}(\boldsymbol{\theta}))}{Z(\boldsymbol{\theta})^2}. \tag{4.9}$$

The second equality in (4.9) uses $\nu \sim \text{Expon}(Z(\boldsymbol{\theta}))$, or equivalently, $\nu = -\log(u)/Z(\boldsymbol{\theta})$, with $u \sim \text{Uniform}(0,1)$. This decomposition is useful in tuning the hyperparameters. The denominator in (4.8) stands for $\tau(m,\lambda,M) = \Pr(\widehat{L}_B > 0)$. Lemma 2 provides its expression. Equation (4.8) shows that having a large proportion of negative estimates has a detrimental effect on the CT.

Figure 4.1 shows the effects of the number of blocks (λ) and Monte Carlo samples (M) on the logarithm of CT, τ and $\sigma^2_{\log|\widehat{L}_B|}$. We consider the three cases $\gamma = 10^2, 100^2, 500^2$ respectively (left to right panels). The panels from left to right show that the optimal λ (corresponding to minimal CT) varies with different values of M and it increases with γ (top row). The minimum CT is associated with a high probability of a positive estimator (τ) (middle row). The last row indicates that $\sigma^2_{\log\left|\widehat{L}_B\right|}$ decreases as a function of λ. Comparing the top nine panels with the bottom nine, a high correlation $\rho = 0.99$, reduces λ_{opt} from 295 (no correlation, $\rho = 0$) to 195 for $\gamma = 500^2$. On the other hand, $\rho = 0.99$



(a) $\rho = 0$

(b) $\rho = 0.99$

Figure 4.1: The effect of the number of blocks λ on the logarithm of CT, τ and $\sigma^2_{\log|\widehat{L}_B|}$. The Poisson parameter m is fixed as 1 for all the panels. The correlation term $\rho = 0$ (upper panel), 0.99 (bottom panel). The columns from left to right, correspond to three different settings of $\gamma = 10^2$, 100^2, and 500^2. The top, middle and last rows show the CT (4.8), the probability of obtaining a positive estimator (see Lemma 2) and the variance of log of the absolute value of the likelihood estimate.

requires at least 100 blocks. So when the variance γ is small, introducing a high correlation increases the CT as more blocks are required compared to the uncorrelated case. Our implementation follows the approach in Tran et al. (2016) which sets the correlation ρ to a value close to 1. Comparing the first row of the top nine panels in Figure 4.1 with that of the bottom nine, it shows that a high correlation significantly reduces the CT per iteration for large γ.

From above, the optimal tuning depends on γ, which is affected by the intrinsic variability of the estimator $\widehat{Z_M}(\boldsymbol{\theta})$ based on (4.9). In the application, γ is set to a large value γ_{max} by using a grid search over possible $\boldsymbol{\theta}$. The tuning process starts with fixed values of λ and m to find the optimal value for M to minimise (4.8). In Figure 4.2, we fix the values of λ and m, with $\lambda = 50, 100$ (the corresponding ρ are 0.98 and 0.99 respectively), and $m = 1$. A standard optimiser is used to find the optimal value M_{opt} for each of the γ. The scattered dots in the left panel of Figure 4.2 plot various values of $\sqrt{\gamma}$ and the corresponding M_{opt}. The figure shows that M_{opt} increases as a function of the $\sqrt{\gamma}$ and similarly for the logarithm of CT as the right panel of Figure 4.2 shows. To illustrate the relationship between M_{opt} and $\sqrt{\gamma}$, a quadratic polynomial is fitted to the dots in the left panel.

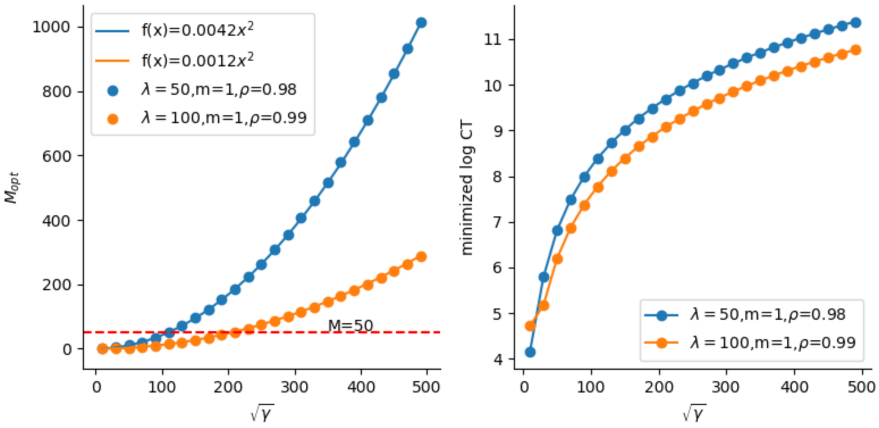


Figure 4.2: Left panel: The optimal value M_{opt} vs $\sqrt{\gamma}$. The lines are quadratics fitted to the scattered dots. Right panel: the minimised CT vs $\sqrt{\gamma}$.

The tuning strategy is based on γ_{max}, leading to a conservative choice of M_{opt}. We recommend the following heuristic approach to choose the values for the hyperparameters.

1 Have a general idea of the posterior distribution of $\boldsymbol{\theta}$. This can be accomplished by conducting an exact method for a few iterations, optimising the posterior distribution by plugging the biased estimator $(1/\widehat{Z}(\boldsymbol{\theta}))$, or adopting an available approximate method.

2 Estimate the corresponding $\mathrm{Var}(\widehat{Z_M}(\boldsymbol{\theta}))/Z^2(\boldsymbol{\theta})$ using a grid search over possible $\boldsymbol{\theta}$ values based on results from Step 1. The estimator $\widehat{Z_M}(\boldsymbol{\theta})$ can be plugged into (4.9) to replace the unknown $Z(\boldsymbol{\theta})$. The variability induced by ν needs to be considered here. A conservative choice is $\gamma(\boldsymbol{\theta}) = 2M\mathrm{Var}(\widehat{Z_M}(\boldsymbol{\theta}))/\widehat{Z_M}^2(\boldsymbol{\theta})$. Appendix B.2 discusses this in more detail.

3 Obtain the maximum value $\gamma_{max}(\boldsymbol{\theta})$ of $\gamma(\boldsymbol{\theta})$ from Step 2. A good starting point is to set $\lambda = 100, m = 1, \rho = 0.99$ and $M_{opt} = \max\{50, 0.0012 \times \gamma_{\max}(\boldsymbol{\theta})\}$.

   When $\gamma_{\max}(\boldsymbol{\theta})$ is small or moderate large, e.g. $\gamma_{\max}(\boldsymbol{\theta}) < 100^2$, having many blocks increases CT. A weaker correlation also produces an efficient algorithm with smaller CT. Another suitable setting is $\lambda = 50, m = 1, \rho = 0.98$ and $M_{opt} = \max\{50, 0.0042\times \gamma_{\max}(\boldsymbol{\theta})\}$.

   For a even smaller $\gamma(\boldsymbol{\theta})$, the correlation can be relaxed further. In the Ising model example, setting $\lambda = 10$ is sufficient when variability is low; see Section 4.5.1 and Appendix B.3.

4.5 Simulation studies

We demonstrate the algorithm on three examples with different targets. The first example is an Ising model, which is usually the benchmark example for doubly intractable problems, as perfect sampling is available for this model. The example serves the purpose of comparing the *signed PMMH* with BP to other methods and showing that the *signed PMMH* with BP generates simulation consistent results with less computing time. In the second example, the intractability is caused by constraint on the Gaussian process $\mathcal{GP}$. We show that not accounting for the constraint leads to erroneous inference. The last example considers the Kent distribution, where the intractable normalising constant is an infinite sum. The last two examples are for models where the non-pseudo methods (the auxiliary variable approaches) cannot be easily applied.

4.5.1 The Ising model

The Ising model (Ising, 1925; Lenz, 1920) has widespread applications such as understanding phase transitions in thermodynamic systems (Fredrickson and Andersen, 1984), interactive image segmentation in vision problems (Kolmogorov and Zabin, 2004) and modelling small-world networks (Herrero, 2002). It is the typical benchmark example in the literature to evaluate different methods for tackling the doubly intractable problem (Atchadé et al., 2013; Lyne et al., 2015; Møller et al., 2006; Park and Haran, 2018). However, most of the existing methods use auxiliary variable approaches, as it is feasible to draw observations from the likelihood function perfectly. The PM methods such as RR and our approach do not require perfect sampling, which makes them applicable to more general problems. In this section, we implement and compare the results given by the BP estimator, the bias-corrected estimator (Quiroz et al., 2019) and the RR method for the Ising model.

Recall Example 4.1 here. Consider an $L \times L$ lattice with binary observations y_{ij} of row i and column j ($y_{ij} \in \{-1, 1\}$). The model is

$$\begin{aligned} p(\mathbf{y}|\theta) &= \frac{1}{Z(\theta)} \exp(\theta S(\mathbf{y})), \\ \text{with } S(\mathbf{y}) &= \sum_{i=1}^{L} \sum_{j=1}^{L-1} y_{i,j} y_{i,j+1} + \sum_{i=1}^{L-1} \sum_{j=1}^{L} y_{i,j} y_{i+1,j} \\ \text{and } Z(\theta) &= \sum_{y_i, y_j} \exp(\theta S(\mathbf{y})), \end{aligned}$$

where the only parameter is θ and the spatial dependence is imposed by $S(\mathbf{y})$. A stronger interaction between observations is associated with a larger θ. Obtaining $Z(\theta)$ is computationally expensive with a sum over 2^{L^2} possible configurations. The simulations are conducted using perfect sampling (Propp and Wilson, 1996), which samples exactly without knowing the normalising constant. Perfect sampling uses coupling to guarantee that the samples are generated from a Markov chain which has already converged to its equilibrium distribution. Following the settings in Park and Haran (2018), two scenarios are considered on a 10×10 grid, with $\theta = 0.2, 0.43$ respectively. See Figure 4.3 for a graphical illustration.

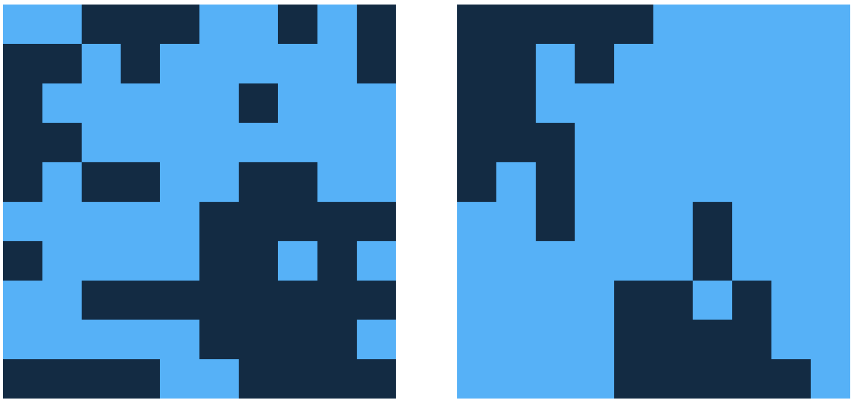


Figure 4.3: Illustrating an Ising model on a 10×10 grid. The samples are drawn using perfect sampling with $\theta = 0.2$ (left) and $\theta = 0.43$ (right). The light and dark blue squares correspond to the values 1 and -1.

For all the algorithms considered, a uniform distribution on $[0, 1]$ is selected as the prior for θ. We adopt a random walk proposal centred at the current θ with a step size 0.07. The PM methods (RR, BP, and the bias-corrected estimator) require an unbiased estimator for $Z(\theta)$. We use annealed importance sampling (AIS) (Neal, 2001) to obtain the estimate of $Z(\theta)$. The method starts by sampling from a tractable distribution (prior) and transfers to an intractable target (posterior) via a sequence of intermediate distributions. The transitions between the distributions are completed via Gibbs updates and the weights associated with the transitions finally constitute the normalising constant of interest; see Neal (2001) for details of AIS in general and Appendix B.3 for its implementation for the Ising model.

As the MH algorithm cannot be applied to the model due to intractability, a “gold” standard is desired to facilitate the evaluation. To obtain such a standard, the approach in Park and Haran (2018) is followed, where an exchange algorithm is conducted with 1,010,000 iterations. The first 10,000 iterations is discarded for burn-in and the remaining iterates are thinned to every 100 iterations. The final samples contain 10,000 iterations.

$\boldsymbol{\theta}_{true} = 0.2$							
Method	Mean	95%HPD	IACT	Time(sec)	ESS/sec	λ	NoImp
Gold	0.205	(0.075, 0.337)	1	-	-	-	-
BP	0.203	(0.066, 0.328)	7.43	676	4.0	10	100
Approx	0.204	(0.077, 0.331)	7.09	62	45.5	-	100
RR	0.202	(0.062, 0.328)	11.65	853	2.0	-	100
$\boldsymbol{\theta}_{true} = 0.43$							
Method	mean	95%HPD	IACT	time(sec)	ESS/sec	λ	NoImp
Gold	0.433	(0.330, 0.533)	1.04	-	-	-	-
BP	0.435	(0.332, 0.545)	6.91	5877	0.5	50	100
Approx	0.441	(0.331, 0.549)	7.78	745	3.5	-	500
RR	0.432	(0.334, 0.549)	10.77	9134	0.2	-	500

Table 4.1: Inference results for the Ising model. All the chains, except for "Gold standard", run for 20,000 iterations using the algorithms described (Gold=exchange algorithm, BP= block-Poisson, Approx = bias-corrected estimator, RR = Russian roulette). The mean estimates are corrected for the negative estimates (BP, RR). The highest posterior density (HPD) is calculated by the `coda` package in `R`. The IACT calculation is based on all the samples as the chains start at the true value. For BP and RR, the calculation of IACT accounts for the negative estimates (see (4.8)). ESS/sec is the effective sample size per second. For BP, λ refers to the number of blocks. NoImp is the number of particles used in the AIS.

Table 4.1 summarises the results of the simulation. When $\theta = 0.2$, the estimates of all the algorithms are close to that of the gold standard. The bias-corrected method has the least computing time and has the best IACT. In the implementation, both the bias-corrected and BP methods exploit the block structure used in the *signed block PMMH* to control the variability in log of likelihood estimates between the current and the proposed value. As suggested in Section 4.4.3, the target correlation is set to no less than 0.98 with at least 50 blocks. We find that when $\theta = 0.2$, the AIS method already gives a sufficiently low value for γ_{max}. So we reduced the number of blocks and set it to 10 ($\lambda = 10$). As a result, the targeted correlation cannot be preserved and the algorithm with a weaker correlation still provides good results. When $\theta = 0.43$, the strong dependence leads to a higher variability in $\widehat{Z}(\theta)$ (see Appendix B.3). We increased the number of blocks to 50 for the BP estimator. To ensure a fair comparison, we also increased the number of particles in the importance samplers of AIS from 100 to 500 for the RR method to bring down the variance. The results of the BP and RR methods match well with that of the gold standard, whereas the bias-corrected method slightly overestimates the parameter. This may be due to the violation of the normality assumption of the bias-corrected estimator when θ is large. The bias-corrected method is 8 times faster than BP and 12 times faster than RR. Comparing the two exact methods with respect to ESS/sec, the BP estimator is around twice as efficient as the RR method.

To sum up, both the BP and the RR methods provide exact inference on the Ising model. We propose a faster bias-corrected estimator; however, the estimator is biased for this method if $\widehat{Z}(\theta)$ is not normally distributed. The normality assumption is unlikely to hold for large θ; see Figure B.2 of Appendix B, which shows a large θ results in a heavily skewed distribution of $\widehat{Z}(\theta)$.

4.5.2 A constrained Gaussian process

A Gaussian process $\mathcal{GP}$ is a stochastic process, i.e., a collection of random variables, such that every finite collection has a multivariate normal distribution. It defines a distribution over functions and is widely used as a non-parametric Bayesian regression method (Williams and Rasmussen, 2006). In this section, we focus on a constrained version of a $\mathcal{GP}$ regression problem, where the constraint stems from the underlying $\mathcal{GP}$ and not from the observations; we can usually transform observations to satisfy constraints. Such process arises naturally in many applications. For example, the prediction value of some chemical concentration data is enforced to lie in an interval of positive regions (Rice et al., 2018). Specifically, we assume that $y = g(\mathbf{x}) + \epsilon$, where $\mathbf{x} \in \mathbb{R}^d$, $g(\mathbf{x}) \geq 0$ and $\epsilon \sim N(0, \sigma^2)$. A constrained $\mathcal{GP}$ prior on $g(\mathbf{x})$ is proposed with the covariance of g chosen as the squared exponential (SE) kernel combined with a diagonal matrix with small positive entries,

$g(\mathbf{x}) \sim \mathcal{GP}(0, K(\mathbf{x}, \mathbf{x}'|\alpha, \rho))\mathbb{1}(g \geq 0)$ with

$$K(\mathbf{x}, \mathbf{x}'|\alpha, \rho) = \alpha^2 \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{2\rho^2}\right) + \tau^2 \mathbb{1}(\mathbf{x} = \mathbf{x}').$$

The constrained $\mathcal{GP}$ prior assumes that the function values behave according to

$$\mathbf{g}|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \sim N(\mathbf{0}, K(\mathbf{x}, \mathbf{x}'|\alpha, \rho))\mathbb{1}(\mathbf{g} \geq \mathbf{0}),$$

where $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$. The nugget effect τ^2 is needed to prevent the determinant of a kernel matrix based on the SE kernel from being close to zero in high dimensions (page 97,98, Williams and Rasmussen, 2006). Setting τ^2 to a small positive number avoids this problem, and we set $\tau^2 = 0.05^2$.

To the best of our knowledge, this type of constraint has not been investigated in the literature; see the survey paper for the constrained $\mathcal{GP}$ in Swiler et al. (2020). The demonstration consists of two parts. The first part focuses on using the $\mathcal{GP}$ process on a small data set (n = 100). The second part considers a scalable $\mathcal{GP}$ on a larger data set (n = 1,000), where it is computationally expensive to conduct exact inference.

4.5.2.1 Prior on the hyperparameters

The $\mathcal{GP}$ regression involves two stages. The first stage carries out inference about the hyperparameters (α, ρ, σ^2); the second stage predicts g for a new location $\mathbf{x}^*$. We use a Bayesian approach and make the predictions based on the iterates to obtain $E(g(\mathbf{x}^*)|\mathbf{y})$.

The hyperparameters are α, ρ, σ. We place an informative prior on the logarithms of these parameters. The priors on $\log\alpha, \log\rho, \log\sigma$ respectively are

$$\log\sigma \sim N(0,1); \quad \log\alpha \sim N(0,1); \quad \log\rho \sim \text{inv-Gamma}(a, b),$$

where a, b are parameters obtained by optimising over an inverse gamma cumulative density function to ensure that the prior can cover a reasonable interval. In the simulation study,

$$\Pr(l, u) = \int_l^u \text{inv-Gamma}(\rho|a, b)d\rho = 0.9,$$

where l, u are the narrowest and the widest gaps between $\mathbf{x}$ and $\mathbf{x}'$. See Betancourt (2020, Section 3.2.3) for more details.

4.5.2.2 $\mathcal{GP}$ on small data sets

The latent variable $g(\mathbf{x})$ has the constrained $\mathcal{GP}$ prior

$$g(\mathbf{x}) \sim \mathcal{GP}(0, K(\mathbf{x}, \mathbf{x}'|\alpha, \rho))\mathbb{1}(g \geq 0).$$

The intractability of $g(\mathbf{x})$ is due to the constraint $g(\mathbf{x}) \geq 0$, resulting in the $\mathcal{GP}$ prior having the normalising constant

$$Z(\alpha, \rho) = \int_{\mathbf{g} \geq 0} p(\mathbf{g}|\alpha, \rho) d\mathbf{g},$$

where $p(\mathbf{g}|\alpha, \rho)$ is a multivariate normal distribution with a mean vector 0, and covariance matrix $K(\cdot)$. Here, $\mathbf{g}$ represents the function values of $g(\mathbf{x})$. Denote $\mathbf{K}_{xx} = K(\mathbf{x}, \mathbf{x}'|\alpha, \rho)$, and the unbiased estimator for the posterior distribution is

$$\widehat{\pi}(\alpha, \rho, \sigma^2, \nu|\mathbf{y}) \propto \widehat{Z^*}(\boldsymbol{\mu}_g^*, \boldsymbol{\Sigma}_g^*) p_y(\mathbf{y}|\alpha, \rho, \sigma^2) \pi(\sigma^2, \alpha, \rho) |\widehat{\exp}(-\nu Z(\alpha, \rho))|, \tag{4.10}$$

where $\boldsymbol{\Sigma}_g^* = (\mathbf{I}_n/\sigma^2 + \mathbf{K}_{xx}^{-1})^{-1}$, $\boldsymbol{\mu}_g^* = \boldsymbol{\Sigma}_g^* \mathbf{y}/\sigma^2$, $\widehat{Z^*}(\boldsymbol{\mu}_g^*, \boldsymbol{\Sigma}_g^*) = \Pr(\mathbf{z} \geq \mathbf{0})$, with $\mathbf{z} \sim N(\boldsymbol{\mu}_g^*, \boldsymbol{\Sigma}_g^*)$ and $p_y(\cdot)$ refers to the unconstrained multivariate normal distribution with mean vector 0 and covariance matrix $\mathbf{K}_{xx} + \sigma^2 \mathbf{I}_n$. Appendix B.4.2 derives the posterior distribution.

Note that $Z(\cdot, \cdot)$ and $Z^*(\cdot, \cdot)$ are both intractable if the number of observations is greater than 2. We adopt the separation of variables (SOV) estimator (Genz, 1992) to estimate $Z(\cdot, \cdot)$ and $Z^*(\cdot, \cdot)$. The estimator is a numerical computational method to evaluate the integral by decomposing the d-dimensional region into d one-dimensional areas which are dependent on each other. Its variability is far smaller than naive Monte Carlo simulation, making it attractive for applications.

4.5.2.3 $\mathcal{GP}$ on big data sets

$\mathcal{GP}$ for big data sets is computationally expensive as the matrix inversion and determinant computations have $O(n^3)$ complexity. There is a vast literature on scalable $\mathcal{GP}$'s (Liu et al., 2020; Quinonero-Candela and Rasmussen, 2005; Williams and Rasmussen, 2006). However, in our case, most methods cannot be used directly due to the intractability caused by the constraint. We consider the popular approximation approach known as fully independent training conditionals (FITC) (Quinonero-Candela and Rasmussen, 2005; Snelson and Ghahramani, 2006). The approach considers a pseudo data set, so-called inducing points, $\overline{\mathbf{x}}_m$ of size $m < n$ and the corresponding values of the function $g(\overline{\mathbf{x}}_m)$, $\overline{\mathbf{g}}_m$, known as the pseudo targets. The matrix operations with regards to $\overline{\mathbf{x}}_m$ are far less costly compared to those with $\mathbf{x}$.

Assume that the likelihood of data $\mathbf{y}$ is

$$p(\mathbf{y}|\mathbf{x}, \overline{\mathbf{x}}_m, \overline{\mathbf{g}}_m) \sim N(\mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\overline{\mathbf{g}}_m, \mathbf{\Lambda} + \sigma^2\mathbf{I}_n),$$

where $\mathbf{K}_{mm} = K(\overline{\mathbf{x}}_m, \overline{\mathbf{x}}'_m|\alpha, \rho)$; $\mathbf{K}_{nn} = K(\mathbf{x}, \mathbf{x}'|\alpha, \rho)$ and $\mathbf{K}_{nm} = K(\mathbf{x}, \overline{\mathbf{x}}'_m|\alpha, \rho)$; and $\mathbf{\Lambda} = \operatorname{diag}(\mathbf{K}_{nn} - \mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\mathbf{K}_{mn})$.

The prior is placed on $\overline{g}_m(\overline{\mathbf{x}}_m)$ instead of $g(\mathbf{x})$, i.e., $\overline{g}_m(\overline{\mathbf{x}}_m) \sim \mathcal{GP}(0, \mathbf{K}_{mm})\mathbb{1}(\overline{\mathbf{g}}_m \geq 0)$. Similarly to the exact inference of the constrained $\mathcal{GP}$, the latent variable $\overline{\mathbf{g}}_m \geq 0$ is integrated out and the posterior conditional on the inducing points $\overline{\mathbf{x}}_m$ is

$$\widehat{\pi}(\alpha, \rho, \sigma^2, \nu|\mathbf{y}, \overline{\mathbf{x}}_m) \propto \widehat{Z^*}(\boldsymbol{\mu}_g^*, \mathbf{\Sigma}_g^*)p_y(\mathbf{y}|\alpha, \rho, \sigma^2)\pi(\sigma^2, \alpha, \rho)|\widehat{\exp}(-\nu Z(\alpha, \rho))|,$$

where $\mathbf{\Sigma}_g^* = \mathbf{K}_{mm}\mathbf{Q}_{mm}^{-1}\mathbf{K}_{mm}$, $\boldsymbol{\mu}_g^* = \mathbf{K}_{mm}\mathbf{Q}_{mm}^{-1}\mathbf{K}_{mn}(\mathbf{\Lambda} + \sigma^2\mathbf{I}_n)^{-1}\mathbf{y}$, $\mathbf{Q}_{mm} = \mathbf{K}_{mm} + \mathbf{K}_{mn}(\mathbf{\Lambda} + \sigma^2\mathbf{I}_n)^{-1}\mathbf{K}_{nm}$. The definitions of $Z^*(\cdot, \cdot), Z(\cdot, \cdot)$ are the same as those in (4.10), with the major difference that the integration is constructed from m-dimensional space instead of n, reducing the complexity from $O(n^3)$ to $O(m^2n)$.

The question of how to construct the pseudo data set $\overline{\mathbf{x}}_m$ arises naturally. Again, due to the intractability, a common approach such as a greedy selection of a subset to maximise the information gain (Seeger et al., 2003) is inapplicable. On the other hand, the optimal pseudo data set does not necessarily come from the observations themselves. Instead, it can be treated as an unknown quantity. Without constraints, the optimal pseudo data set can be obtained by the gradient-optimiser as suggested in Snelson and Ghahramani (2006). A corresponding Bayesian treatment is presented in Rossi et al. (2021), where various priors are put on the inducing points. A heuristic approach is now proposed, where the inducing points are fixed before the MCMC chain starts. We first fit the data using a k-means clustering method and then randomly select one observation out of each cluster. An alternative approach is to use a binary variable to label whether the observation is an inducing point or not. However, this method requires using many latent variables. Hence, we do not use this method. We will show that the heuristic approach leads to satisfactory simulation results.

4.5.2.4 Simulation results

In the simulation, the data set is generated by the following function of a d-dimensional $\mathbf{x} = (x_1, \ldots, x_d)$,

$$\begin{aligned} g(\mathbf{x}) &= \frac{5}{\pi}(1 - 0.9t(\mathbf{x}))\exp(-0.5t(\mathbf{x})) + C, \quad \text{with } t(\mathbf{x}) = 0.25\,\|\mathbf{x}\|_2^2\,, \\ y(\mathbf{x}) &= g(\mathbf{x}) + \epsilon, \epsilon \sim N(0, \sigma^2), \end{aligned}$$

where $C = -\min(g(\mathbf{x})) + 0.01, x_i \in [-5, 5]$, $i = 1, \ldots, d$, to ensure the process is constrained to lie above zero.

For the training data, we generate n observations $\mathbf{x} \in \mathbb{R}^d$ from a multivariate normal distribution with mean vector 0 and a diagonal covariance matrix with all its entries equal to 4. The values of $\mathbf{x}$ are constrained to the hyper-rectangular $[-5, 5]^d$. We generate data sets with each of the combinations from $d = 2, 4$ and $n = 100, 1{,}000$, respectively. For the noise level, we choose $\sigma^2 = 0.5^2$, so that there is a considerable percentage of negative observations ($20\% \sim 30\%$). Table 4.2 documents the settings for generating the test data in all the scenarios.

Figure 4.4 illustrates the 2-dimensional function. The test region is slightly bigger than that covered by the training data. Based on the functional form, the points close to the boundary are likely to have negative observations. The design serves the purpose of testing the prediction accuracy of these points, where there is a limited number of neighbouring observations. As the posterior prediction of a point is largely affected by its neighbouring

	$d = 2$	$d = 4$
n = 100	20 pts/dim, 400	5 pts/dim, 625
n = 1,000	60 pts/dim, 3,600	8 pts/dim, 4,096

Table 4.2: Scheme of how the test data are generated. Several equally spaced points are generated between [-5,5] (inclusive) per dimension. The number after the comma is the total number of test points.

points for a $\mathcal{GP}$ process, the prediction results of the points close to the boundary are likely to be less credible if the possibility of obtaining negative predictions is not ruled out in the model. This phenomenon turns out to be more distinct in high dimensions as the number of points near the boundary increases with the number of dimensions. A large data set will overcome the issue, but with an associated high computational cost.

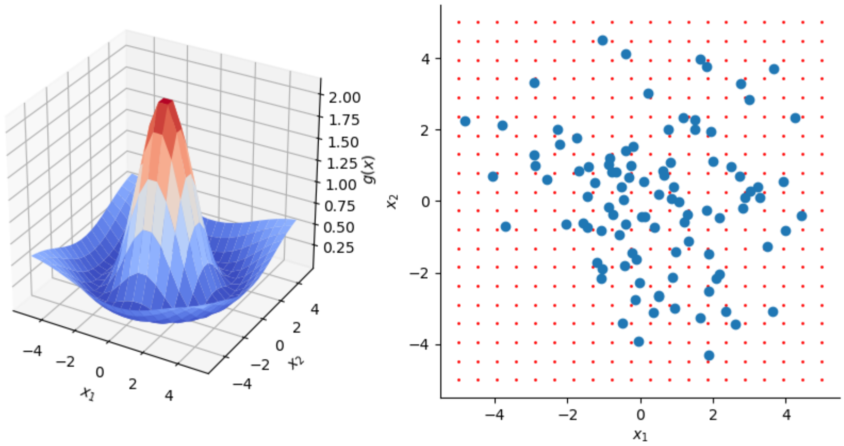


Figure 4.4: Left: Plot of the function ($d = 2$). Right: A realisation of the data generation with $n = 100$ (blue points). The red points are the locations of the testing data.

Table 4.3 presents summary results obtained for 20 independent replications. The inference is based on 10,000 iterations with the first 5,000 iterations discarded as the burn-in period for both the unconstrained (UNCONS) and the constrained (CONS) models. For the $n = 100$ case, both models give fairly good estimates for σ^2 (true $\sigma^2 = 0.25$) with relatively low IACT, indicating the good mixture behaviour of the chain. The RMSE for the training data are close for both models. For the test data, CONS generates better prediction results with lower RMSE. The raw predictions do not take the non-negative constraint into consideration. The correction rounds up the negative predictions to zero for UNCONS. For CONS, the posterior prediction distribution is a constrained truncated normal distribution. We use the posterior median as it is more robust than the posterior mean. Appendix B.4.2 gives more details. The corrected predictions of CONS reduce RMSE by around 25% compared with UNCONS.

For the large data set $n = 1{,}000$, the $\mathcal{GP}$ is placed on the inducing points (50 observations). Similarly to the previous results, the performance of CONS and UNCONS is close in terms of IACT and RMSE for the training data. For the test data predictions, CONS has 30-40% lower RMSE compared to UNCONS. The larger gap can be explained because the predictions are mainly based on the inducing points, not the whole data set, resulting in less support from the neighbouring points. Consequently, the extrapolation based on an inducing point with a negative value is likely to yield a negative prediction. For UNCONS, such extrapolation amplifies the prediction error given the constraint is not incorporated into the model.

4.5.3 The Kent distribution

Directional statistics involves the study of density functions on the collection of unit vectors. The Kent distribution[2] (FB_5) is proposed as an analogue to the bivariate normal distribution to model asymmetrically distributed data on a spherical surface (Kent, 1982). Its distribution is characterised by 5 parameters: $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3, \beta$, and κ, where $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3$ form a 3-dimensional orthonormal basis, representing the mean, major and minor axes; κ is the concentration parameter, and β is a measure of its ovalness, with the constraint $0 \leq \beta < \kappa/2$ to ensure that the distribution is unimodal.

Example 2 is

$$f(\mathbf{y}|\boldsymbol{\gamma_1}, \boldsymbol{\gamma_2}, \boldsymbol{\gamma_3}, \beta, \kappa) = \frac{1}{c(\kappa, \beta)} \exp\left\{\kappa \boldsymbol{\gamma_1}^\top \cdot \mathbf{y} + \beta \left[(\boldsymbol{\gamma_2}^\top \cdot \mathbf{y})^2 - (\boldsymbol{\gamma_3}^\top \cdot \mathbf{y})^2\right]\right\},$$

[2] The Kent distribution is also known as the 5-parameter Fisher Bingham distribution

			$n = 100$			
				RMSE		
Method	d	σ^2	IACT	train	test-raw	test-corrected
UNCONS	2	0.235	11.088	0.180	0.284	0.232
		(0.037)	(1.072)	(0.037)	(0.05)	(0.035)
CONS	2	0.259	11.371	**0.178**	**0.212**	**0.179**
		(0.041)	(0.886)	(0.037)	(0.034)	(0.037)
UNCONS	4	0.229	16.647	0.272	0.495	0.486
		(0.068)	(8.818)	(0.043)	(0.046)	(0.042)
CONS	4	0.308	14.012	**0.256**	**0.488**	**0.371**
		(0.043)	(2.355)	(0.028)	(0.015)	(0.039)

			$n = 1{,}000$			
				RMSE		
Method	d	σ^2	IACT	train	test-raw	test-corrected
UNCONS	2	0.237	10.26	0.076	0.143	0.132
		(0.01)	(1.206)	(0.012)	(0.029)	(0.019)
CONS	2	0.237	10.055	**0.070**	**0.122**	**0.094**
		(0.01)	(1.103)	(0.009)	(0.014)	(0.01)
UNCONS	4	0.204	12.157	**0.203**	0.468	0.463
		(0.023)	(2.075)	(0.02)	(0.007)	(0.007)
CONS	4	0.216	11.848	0.205	**0.461**	**0.275**
		(0.026)	(1.658)	(0.023)	(0.006)	(0.013)

Table 4.3: Results for the $\mathcal{GP}$ prior using observations of sizes 100 and 1,000. The results obtained are the mean value of 20 independent replications with the standard deviation in brackets. For the large data set (n = 1,000), 50 inducing points are used. "CONS" and "UNCONS" stand for the model with/without constraints (no intractable quantity involved). The negative predictions are rounded up to zero for both models.

where $\mathbf{y} \in \mathbb{R}^3$, $\|\mathbf{y}\| = 1$. The normalising constant $c(\kappa, \beta)$ is

$$c(\kappa, \beta) = 2\pi \sum_{j=0}^{\infty} \frac{\Gamma(j+0.5)}{\Gamma(j+1)} \beta^{2j} (0.5\kappa)^{-2j-0.5} I_{2j+0.5}(\kappa),$$

where $I_\nu(\cdot)$ is the modified Bessel function.

The intractable constant involves an infinite sum. Due to the complex form of the density function, Kent (1982) proposes a moment estimator of the parameters, which is consistent. The moment estimation of $\boldsymbol{\gamma}_i, (i = 1, 2, 3)$ is independent of β and κ. Estimation of β and κ requires an approximation that utilises the limiting case when $2\beta/\kappa$ is small or κ is large, provided that the moment estimates of the $\boldsymbol{\gamma}_i$ are available. Alternatively, κ and β can be obtained numerically. Kume and Wood (2005) adopt saddle point techniques to obtain the approximation for the normalising constant directly. Kasarapu (2015) uses the Bayesian framework to model a mixture of FB$_5$ distributions. The $\boldsymbol{\gamma}_i$ basis terms are reparameterised in terms of the three angular parameters ψ, α, η. The infinite sum in $c(\kappa, \beta)$ is truncated in the sense that the successive term to be added is less than a prefixed threshold. However, this approach results in inexact inference. In contrast, the *signed block PMMH with BP* provides exact inference for the parameters. To the best of our knowledge, exact Bayesian inference is not considered for the Kent distribution in the literature.

To obtain an unbiased estimator for $c(\kappa, \beta)$, we use the approach proposed by Papaspiliopoulos (2011). Rewrite $c(\kappa, \beta)$ as $\sum_{j=0}^{\infty} \phi_j(\kappa, \beta)$; then the estimator $\widehat{c}(\kappa, \beta) = \phi_k / q_k$ is unbiased, where k is a non-negative discrete random variable with probability mass function q_k. Either a Poisson or a geometric distribution is suitable, as k is a non-negative integer. It is straightforward to verify that $E(\widehat{c}(\kappa, \beta)) = \sum_k \phi_k / q_k \times q_k = c(\kappa, \beta)$. As $\phi_j(\kappa, \beta)$ is a decreasing function in j, a “hard” truncation is set up for the first K terms to reduce its variability. $c(\kappa, \beta)$ is rewritten as

$$\sum_{j=0}^{K-1} \phi_j(\kappa, \beta) + \sum_{j=K}^{\infty} \phi_j(\kappa, \beta).$$

The first K terms are computed in the implementation and the remaining terms are truncated at a random point.

We adopt the reparameterisation trick (Kasarapu, 2015) in the analysis. The orthonormal basis $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3$ can be reparameterised as $\psi \in [0, \pi], \alpha \in [0, 2\pi], \eta \in [0, \pi]$. An adaptive random walk proposal is used for all the parameters with the optimal covariance matrix proposed in Garthwaite et al. (2016). To accommodate such a proposal, we further transform ψ, α, η into ψ^*, α^*, η^* which take unconstrained values using the following transformations:

$$\psi^* = \log\left(\frac{\psi}{\pi - \psi}\right); \alpha^* = \log\left(\frac{\alpha}{2\pi - \alpha}\right) \quad \text{and} \quad \eta^* = \log\left(\frac{\eta}{\pi - \eta}\right).$$

We also model β and κ in terms of their logarithms to ensure they are unconstrained.

We follow Dowe et al. (1996) and set the prior for κ as $\frac{4\kappa^2}{\pi(1+\kappa^2)^2}$. For a given κ, the prior for β is uniform on $[0, \kappa/2)$. The priors for ψ, α and η follow Kasarapu (2015). The final prior on all the parameters, $\psi, \alpha, \eta, \beta$ and κ is

$$\pi(\psi, \alpha, \eta, \beta, \kappa) = \frac{2\kappa \sin \alpha}{\pi^3(1+\kappa^2)^2} \mathbb{1}(0 \leq 2\beta/\kappa < 1).$$

In the simulation, we generate n observations from FB_5, with different settings for β and κ. The data generation is performed by the R package `Directional`, which implements the acceptance-rejection method in Kent et al. (2013). We set n = 10, 100, 1,000 in combination with β/κ = 0.01, 0.25, 0.49, with κ fixed as 5. The lower and the upper bounds for β/κ are 0 and 0.5 to ensure unimodality of the data (Kent, 1982).

		n=10			n=100			n=1,000		
	Method/RMSE	β	κ	β/κ	β	κ	β/κ	β	κ	β/κ
$\beta/\kappa = 0.01$	Bayesian	**1.33**	**2.40**	0.22	0.45	**0.50**	0.09	0.11	0.17	0.02
	Moment	1.48	4.01	**0.16**	**0.25**	0.55	**0.05**	**0.05**	**0.16**	**0.01**
	MLE	2.29	4.30	0.26	0.47	0.57	0.09	0.11	0.17	0.02
$\beta/\kappa = 0.25$	Bayesian	**0.64**	**2.20**	**0.04**	0.38	0.55	0.07	**0.11**	**0.16**	**0.02**
	Moment	1.01	3.69	0.09	0.63	**0.54**	0.13	0.65	0.17	0.13
	MLE	2.00	4.18	0.16	**0.36**	0.60	**0.06**	0.35	0.24	0.07
$\beta/\kappa = 0.49$	Bayesian	**1.28**	**1.99**	0.22	**0.42**	**0.57**	**0.06**	**0.14**	**0.19**	**0.02**
	Moment	1.38	3.27	0.27	1.48	0.59	0.28	1.50	0.46	0.28
	MLE	1.73	3.79	**0.17**	**0.42**	0.60	**0.06**	0.98	0.59	0.19

Table 4.4: Simulation results for 100 independent replications of a FB_5 distribution. All numbers refer to the RMSE with respect to the true value.

Table 4.4 shows the RMSE with regard to the true values for the three methods based on 100 independent replicates. "Bayesian" refers to the *signed block PMMH with BP* algorithm. The selected hyperparameters are $\lambda = 20$ (number of blocks), $m = 1$ (Poisson mean value of BP), and $K = 10$ (truncated terms). A Poisson distribution with mean value 1 is used for the truncation. "Moment" refers to the moment estimates and "MLE" is based on our modification of function `kent.mle` of the R package `Directional`, where the original version uses the moment estimates of γs. We use an optimiser on the transformed parameters to obtain the MLE of all the parameters in the modified version. For the Bayesian method, the RMSE is calculated using the posterior mean after the sign correction. For a small number of observations ($n = 10$), our method yields the smallest RMSE amongst all three methods.

Comparing the results of different β/κ combinations, the moment estimator gives the best RMSE for $\beta/\kappa = 0.01$, and our method is superior to the other two when the ratio approaches 0.49, where the assumption underlying the moment estimator is almost violated. The MLE method uses the saddle point technique (Kume and Wood, 2005) to approximate the likelihood. Its performance gets closer to the Bayesian method for $\beta/\kappa = 0.01$ with many observations. As β/κ increases, it outperforms the moment estimator, but is inferior to the Bayesian method.

The simulation study shows that the *signed block PMMH with BP* algorithm performs the best when the sample size is small. It also has the lowest RMSE when β/κ approaches the limiting value 0.5.

4.6 An empirical study on spherical data

We now analyse four real world spherical data sets using the Kent distribution and our method. Recall that the non-pseudo marginal approaches cannot be applied to this model. Each data set contains samples from two groups, which are formed naturally from the sample collection process. Figure 4.5 plots the spherical data.

1. **Palaeomagnetic** (Palaeo) (Wood, 1982): Thirty three estimates of previous magnetic pole positions were obtained using palaeomagnetic techniques. Each estimate is associated with a different site in Tasmania. The data is originally in Schmidt (1976) and the author points out that the data is likely to fall mainly into two groups of distinct geographical regions. Following Figueiredo (2009), the first group contains the observation indices 9, 10, 11, 12, 14, 16, 23, 24, 30.

2. **Magnetic** (Fisher et al., 1993, Table B8): Measurements of magnetic remanence from a set of 62 specimens is obtained. The specimens are from Mesozoic Dolerite from Prospect, New South Wales, after successive partial demagnetisation stages (200° and 350°). An experiment was conducted to determine the blocking temperature spectrum of the magnetisation components.

3. **Sandstone** (Fisher et al., 1993, Table B23): Measurements of natural remanent magnetisation in Old Red Sandstone rocks in Pembrokeshire, Wales. The measurements consist of specimens from two sites with the number of observations 35 and 13, respectively.

4. **Stone** (Fisher et al., 1993, Table B25): Measurements of the longest axis (101 observations) and shortest axis (101 observations) orientations of tabular stones on a slope at Windy Hills, Scotland.

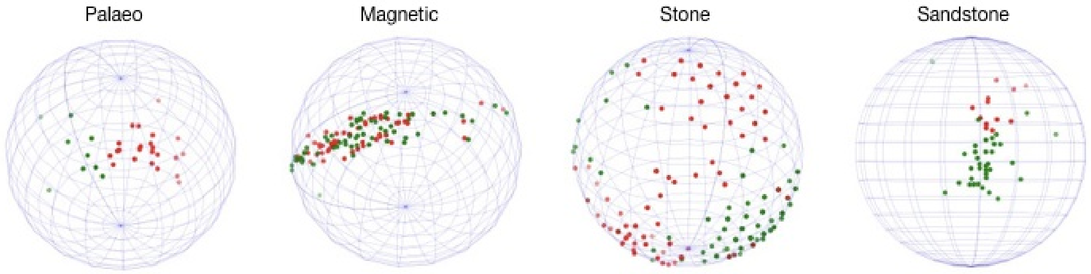


Figure 4.5: Illustration of the data sets. Green points and red points refer to the observations from groups 1 and 2, respectively.

The two groups are modelled separately by assuming a non-hierarchical structure on the prior for all the parameters. The data is modelled in the same way as in Section 4.5.3 using the density function (4.2).

Table 4.5 summarises the results. We first note that the three methods provide different estimates of the same quantity. The gap between the estimates of Bayesian and ML is narrower for the bigger data sets (Magnetic, Stone). The moment estimates are far from the MLE and the Bayesian results, even for the bigger data sets. Since β/κ is close to 0.5 in Magnetic, the moment estimate is unreliable. This result is supported by the simulation results in Section 4.5.3. The second result is that the confidence intervals for the moment estimates and the MLE, especially for small data sets (Palaeo, Sandstone), are wider than the Bayesian intervals. The Bayesian credible interval is constructed using the posterior distribution. For the MLE and moment estimates, we obtain the confidence intervals using the non-parametric bootstrap (Efron, 1992). By construction, the intervals have different interpretations (frequentist vs Bayesian); however, both intervals are expected to be close when the number of observations is sufficiently large. The interval results for the bigger data set, Stone, in Table 4.5 show this, where the MLE and Bayesian results are close to each other. The moment estimates again seem to be less reliable for β. A larger κ indicates the observations are more concentrated. For small data sets, the non-parametric bootstrap is likely to draw the same observation multiple times, resulting in a concentrated data pattern and a correspondingly large estimate of κ.

We use 5-fold cross validation to test the models' performance. The data set is split into 5 folds, with 4 folds being the training set and the fifth the test set. To avoid bias in sampling, the splitting is done for both groups. Denote the training and test sets as $\mathbf{y}_{train,g}, \mathbf{y}_{test,g}$ with g the group membership $g \in \{1, 2\}$.

After fitting the models using $\mathbf{y}_{train,g}$, the prediction for an observation $\mathbf{y}_i$ conditional on samples $\boldsymbol{\theta}_g = \{\beta, \kappa, \psi, \alpha, \eta\}$ from the posterior distribution is

$$p(\mathbf{y}_i|\mathbf{y}_{train,g}) = \int_{\boldsymbol{\theta}} p(\mathbf{y}_i|\boldsymbol{\theta}_g)p(\boldsymbol{\theta}_g|\mathbf{y}_{train,g})d\boldsymbol{\theta}_g.$$

If $p(\mathbf{y}_i|\mathbf{y}_{train,1}) > p(\mathbf{y}_i|\mathbf{y}_{train,2})$, $\mathbf{y}_i$ is classified as being in group 1, and conversely. Appendix B.5 provides more details.

Table 4.6 shows the prediction accuracy on the training and test data sets. There is no notable difference in terms of the accuracy between the methods across the data sets. One possible reason is that the parameters of one group are distinct from that of the other group, so that minor differences in parameter estimates do not appreciably affect the classification.

4.7 Discussion

We propose the *signed block PMMH with BP* algorithm to carry out inference in general doubly intractable problems. This algorithm only relies on the availability of an unbiased

Palaeo	group 1 (n=9)			group 2 (n=24)		
	β	κ	β/κ	β	κ	β/κ
Bayesian	2.78	18.25	0.16	3.86	33.89	0.11
	(0.20,11.64)	(7.92,38.55)	(0.01,0.41)	(0.27,12.76)	(21.33,50.48)	(0.01,0.32)
Moment	4.03	26.54	0.15	5.88	39.84	0.15
	(1.11,98.84)	(18.96,223.76)	(0.05,0.46)	(1.24,30.51)	(28.16,92.60)	(0.04,0.35)
MLE	4.55	26.65	0.17	6.44	40.02	0.16
	(0.78,106.84)	(18.96,223.02)	(0.04,0.50)	(1.06,34.34)	(28.25,96.38)	(0.03,0.38)
Magnetic	group 1 (n=62)			group 2 (n=62)		
	β	κ	β/κ	β	κ	β/κ
Bayesian	7.32	15.23	0.49	15.77	32.04	0.49
	(5.28,10.03)	(11.18,20.34)	(0.43,0.50)	(11.17,21.58)	(22.95,43.29)	(0.46,0.50)
Moment	4.55	12.99	0.35	8.87	23.22	0.38
	(2.58,10.73)	(8.21,26.72)	(0.31,0.40)	(5.28,20.53)	(14.77,49.06)	(0.35,0.42)
MLE	8.24	16.49	0.50	15.57	31.13	0.50
	(4.76,16.95)	(9.57,34.04)	(0.49,0.50)	(9.82,33.12)	(19.65,66.23)	(0.49,0.50)
Sandstone	group 1 (n=36)			group 2 (n=13)		
	β	κ	β/κ	β	κ	β/κ
Bayesian	1.48	20.31	0.07	8.54	47.08	0.19
	(0.08,5.68)	(14.51,28.33)	(0.00,0.23)	(0.70,27.20)	(24.69,89.95)	(0.02,0.39)
Moment	2.07	22.36	0.09	18.45	68.94	0.27
	(0.70,16.37)	(13.38,64.33)	(0.04,0.30)	(8.76,67.07)	(54.64,188.26)	(0.11,0.41)
MLE	2.42	22.44	0.11	20.15	70.18	0.29
	(0.00,17.93)	(13.41,65.55)	(0.00,0.36)	(7.90,76.18)	(55.30,199.75)	(0.09,0.44)
Stone	group 1 (n=101)			group 2 (n=101)		
	β	κ	β/κ	β	κ	β/κ
Bayesian	0.52	4.19	0.13	1.06	2.18	0.49
	(0.05,1.32)	(3.37,5.12)	(0.01,0.30)	(0.79,1.34)	(1.64,2.72)	(0.44,0.50)
Moment	0.23	4.29	0.05	0.41	1.99	0.21
	(0.08,0.54)	(3.33,6.23)	(0.02,0.11)	(0.30,0.52)	(1.79,2.30)	(0.15,0.24)
MLE	0.60	4.32	0.14	1.10	2.19	0.50
	(0.15,1.91)	(3.38,6.31)	(0.03,0.50)	(0.92,1.30)	(1.85,2.61)	(0.50,0.50)

Table 4.5: Results for the Bayesian, moment and MLE approaches for all the data sets. The Bayesian estimate is the posterior mean. The numbers in brackets are the 95% confidence (credible for Bayesian) intervals. For the moment estimates and MLE, the confidence intervals are obtained using the bootstrap (Efron, 1992) with 1,000 repetitions each.

			Train accuracy			Test accuracy		
data	grp1	grp2	Bayesian	Moment	MLE	Bayesian	Moment	MLE
Palaeo	9	24	0.985 (0.019)	0.985 (0.019)	0.985 (0.019)	0.943 (0.070)	0.943 (0.070)	0.943 (0.070)
Magnetic	62	62	0.554 (0.024)	**0.561** (0.031)	0.548 (0.033)	**0.507** (0.014)	0.502 (0.077)	0.501 (0.083)
Sandstone	36	13	0.943 (0.030)	**0.953** (0.031)	**0.953** (0.031)	**0.920** (0.160)	**0.920** (0.117)	0.880 (0.147)
Stone	101	101	**0.892** (0.005)	0.877 (0.006)	0.890 (0.003)	0.886 (0.011)	0.861 (0.025)	**0.896** (0.011)

Table 4.6: Results of 5-fold cross validation on the four data sets. “grp1”,“grp2” are the number of observations for the corresponding group. Accuracy on the training and test data is the average value of the 5 folds, with standard deviation in brackets.

estimator of the normalising constant, which makes it more applicable to a wider range of problems than its competitors, who often require perfect sampling from the model. Our simulation study in Section 4.5.2 shows that it is crucial to take the intractable normalising constant into modelling consideration. Otherwise, inexact inference leads to poor prediction results. Compared with the RR method, the BP estimator controls the variability of the logarithmic difference in the likelihood estimates in the MH acceptance ratio by tuning its hyperparameters. The Ising model example suggests that it is also faster than RR.

In spite of its wide applicability, the *signed PMMH* algorithm suffers from a high computational cost when unbiasedly estimating the normalising constant. The Ising model example shows that AIS is required multiple times during each iteration for both the RR and BP methods. This prevents the computing time of the PM methods being competitive with other methods which do not require computing the normalising estimates explicitly. However, the algorithm gives exact inference in almost all situations regardless of its high computing time. When the normalising constant is an infinite sum as in the Kent distribution, the BP estimator can be obtained relatively cheaply, enabling the *signed block PMMH with BP* algorithm to complete in a reasonable time. Hence, one of area for future work is to obtain an efficient estimator of the normalising constant. With respect to the applications, the proposed algorithm is the first exact Bayesian analysis on the Kent distribution. The application in this chapter only considers 3-dimensional data. Future work could apply the method to higher dimensional problems, provided that an efficient estimator of the likelihood is available.

This chapter establishes guidelines for hyperparameter tuning, where the hyperparameters are fixed before the start of the MCMC. However, it may be desirable to have a dynamic tuning strategy during the MCMC, where the number of particles estimating the normalising constant is reduced when the likelihood estimator has low variability and vice versa.

Chapter 5

Estimating heterogeneous treatment effects by extending the Bayesian additive regression tree (BART) algorithm

5.1 Introduction

Causal inference investigates the causal connection between the occurrence of an event and its effect on an outcome. Unlike traditional statistical models which investigate the pattern behind data and aim for small prediction errors, causal inference focuses on estimating the treatment effect. Based on Rubin's causal framework (Rubin, 1978), the treatment effect is defined by comparing among all outcomes that could have been observed under possible treatment assignments. However, in real experiments, only one assignment is applied and the outcome under that assignment is observed on a single unit. In the absence of randomised experiments, it is challenging to estimate the treatment effect for observational studies as the assignment mechanism is unknown. Researchers often assume ignorability of the treatment assignment conditional on the observed covariates (Rosenbaum and Rubin, 1983), so that the estimation of the treatment effect involves predicting the value(s) for the unassigned treatment(s). In recent years, various attempts were made to adopt machine learning methods for causal inference, as they perform well for prediction on trained models. Early research mainly focuses on implementing existing machine learning techniques to obtain good predictions of either potential outcomes or the propensity score, which is defined as the probability of receiving treatment given the covariates. Much recent effort was made to improve these methods to accommodate the causal inference setting. In this chapter, we are particularly interested in adapting the Bayesian additive regression tree (BART) algorithm (Chipman et al., 2010) for causal inference.

The BART model uses a sum of regression trees to fit complex response surfaces, where the BART prior applies regularisation. Hill (2011) uses BART in estimating heterogeneous treatment effects for causal inference, demonstrating its superior performance in the context of modelling nonlinear response surfaces. In Hill, the treatment indicator is dealt with similarly to the observed covariates. Hill and Su (2013) discuss the weakness of such a naïve implementation. The BART model tends to generate biased results due to extrapolation on the covariate space, where there is no common support from observations of the control and treatment groups. Hahn et al. (2020) develop a new method based on the original BART implementation for causal inference, referred to as the “Bayesian Causal Forest” (BCF). In the BCF, the response surface is separated into the prognostic impact (the conditional mean of the response that is unrelated to the treatment effect) and the treatment effect, which are modelled by separate collections of regression trees. The design provides adequate control over the strength of regularisation over the heterogeneity effect. Hahn et al. (2018) find that regularised models for Bayesian linear regression which were originally designed for prediction can bias causal estimates. The bias can be quantified by the function involving the unknown nuisance parameters and the form of the design matrix. To circumvent the bias induced by regularisation, the BCF includes an estimate of the propensity score as one of the covariates, which substantially improves the treatment effect estimation in the presence of confounding variables (the variables influencing both the treatment and the outcome).

Besides BART and its extensions (Hahn et al., 2020; Murray, 2021; Sparapani et al., 2016), there are non-Bayesian tree-based methods for estimating causal effects. Among them, the generalised random forest (GRF) (Wager and Athey, 2018) is popular for causal inference. The GRF also aims to estimate heterogeneous treatment effects. It is derived from traditional random forests, but with distinctive tree growing/pruning criteria. To accommodate random forests for causal inference, the tree structure is forced to be identical for the observations from the control and the treatment groups, implying that each leaf of the tree must contain at least a few observations from both groups. The bottom layer of the interior nodes of each tree splits at the treatment indicator, implying that the treatment effect can be inferred from the difference in mean values of the leafs. The GRF produces a point-wise consistent estimator of the true treatment effect, which has an asymptotically Gaussian distribution (Wager and Athey, 2018).

We propose extending the BART model (BART-EXT) for estimating heterogeneous treatment effects of an observational study with continuous outcomes and binary treatments. The model combines the idea of the BCF and GRF approaches, where the response surface is split into the baseline (prognostic) effect and the treatment effect, with each part modelled by separate regression trees. We force the trees modelling the treatment effect to have identical structures for observations from the treatment and the control groups.

The proposed model inherits the strength of both the BCF and GRF approaches and overcomes their downsides. For example, the BCF shrinks causal effects to homogeneity, which is not guaranteed in real data sets. In our proposal, the identical tree structure protects the estimates from shrinkage in the presence of strong heterogeneity. Section 5.3.3 discusses the similarities and differences between the methods in more detail. The simulation studies in Section 5.4 show that the BART-EXT method produces substantially better results than GRF and the BCF.

The rest of this chapter is organised as follows. Section 5.2 provides the background knowledge on causal inference, the BART method, and the BCF and GRF approaches. Section

5.3 presents the proposed model with its prior specification, followed with a discussion on the similarities and differences between the BCF and GRF. Section 5.4 compares the BART-EXT method with the other tree-based methods using simulation. Section 5.5 applies the method to an experimental study on fuel intensity. Section 5.6 concludes with a summary and discussion.

5.2 Background

5.2.1 Causal inference

Causal inference on a data set can be treated as a missing data problem, where the missing pattern is not random. Suppose the treatment indicator Z is binary, stating that each individual either receives the treatment ($Z = 1$) or does not ($Z = 0$). The potential outcome is defined as the outcome associated with the paired actions, denoted as $Y(1)$ and $Y(0)$. Based on Rubin's causal framework (Rubin, 1978), the fundamental problem of causal inference, stated by Holland (1986), is that at most one of the potential outcomes is observed. Throughout this chapter, we denote each observation as $(Y_i, \mathbf{X}_i, Z_i)$ for unit $i = 1, \ldots, n$. The scalar Y_i is the observed continuous outcome, $\mathbf{X}_i$ and Z_i respectively denote the observed corresponding covariates and the binary treatment indicator. The observed Y_i can also be represented compactly by $Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$. We are interested in estimating the conditional average treatment effect (CATE), defined as,

$$\tau_{CATE}(\mathbf{x}) := E(Y(1) - Y(0) | \mathbf{X} = \mathbf{x}).$$

The average treatment effect (ATE) and the average treatment effect on the treated units (ATT) are also of interest, defined as,

$$\tau_{ATE} := E(Y(1) - Y(0)), \quad \tau_{ATT} := E(Y(1) - Y(0) | Z = 1).$$

As usually stated in the causal inference literature, we make the stable unit treatment value assumption (SUTVA) (Rubin, 1978) and strong ignorability (Rosenbaum and Rubin, 1983) for the observations. The first is a critical assumption of no interference between different units, where the potential outcomes of one subject are unaffected by the changes of treatments received by all the other subjects. The second assumption has two parts: conditional independence $Z_i \perp Y_i(0), Y_i(1)|\mathbf{X}_i$ and the overlapping condition $0 < \Pr(Z_i = 1|\mathbf{X}_i = \mathbf{x}_i) < 1$, $i = 1, \ldots, n$. Provided that both assumptions hold, CATE can be expressed as,

$$\tau_{CATE}(\mathbf{x}_i) = E(Y_i|\mathbf{X}_i = \mathbf{x}_i, Z_i = 1) - E(Y_i|\mathbf{X}_i = \mathbf{x}_i, Z_i = 0). \tag{5.1}$$

Based on (5.1), the casual inference problem is then transformed into a prediction problem whose target is to predict the response surfaces, $E(Y|\mathbf{X} = \mathbf{x}, Z = 1)$ and $E(Y|\mathbf{X} = \mathbf{x}, Z = 0)$.

There are two approaches to obtain the predictions for $E(Y|\mathbf{X}, Z)$. The first approach is to model the treated and untreated units separately. The final estimate of the CATE is then obtained by combining the predictions from the two models. However, this approach is problematic in the sense that if the treatment and the control groups have different features (self-selection), the prediction step may require extrapolation in one or both of the models, given no support from the covariate space. For example, suppose the issue of interest is the effectiveness of an anti-hypertension drug. Subjects with high blood pressure are more likely to take the pills compared with the healthy ones. If the response model on the treated presumes the contributing factors for high blood pressure as important predictors, whereas those are ignored in the model on the control units, the prediction model for the control group may misinterpret the contributions and hence biases the causal estimates.

The second approach to predicting $E(Y|\mathbf{X}, Z)$ models observations from the treatment and the control groups jointly. The model, including observations from both groups, has components that are no longer independent in the sense that they share common structure to some degree. A typical example is the ordinary least squares (OLS) regression where the response Y is regressed against $\mathbf{X}$ and Z, possibly with interaction terms. The prediction difference obtained by alternating Z and fixing $\mathbf{X}$ gives the CATE. Künzel et al. (2019) discuss both approaches in detail, which are called the “T-learner”(the first approach) and the “S-learner” (the second approach). See Künzel et al. and its references.

5.2.2 BART

Prior to the development of BART, Chipman et al. (1998) introduced Bayesian CART (Bayesian classification and regression tree) which involves one decision tree. BART, introduced by Chipman et al. (2010), assumes a normal distribution of the response Y, with the conditional mean constructed by a sum-of-trees model. It is a Bayesian non-parametric model with a data-driven tree structure. We follow the Chipman et al. (2010) notation throughout this chapter. The BART model is explicitly expressed as

$$Y = \sum_{j=1}^{m} g(\mathbf{x}; T_j, M_j) + \epsilon, \quad \epsilon \sim N(0, \sigma^2).$$

Without loss of generality, in the causal inference setting, input variables $\mathbf{x}$ can include both the covariates and the treatment indicator. The jth binary tree structure T_j includes the variable to split and the associated decision rule for each node. The $M_j = \{\mu_{1j}, \dots, \mu_{b_j j}\}$ represents a set of parameter values associated with the terminal nodes, with b_j the number of terminal nodes (leafs) of T_j. This structure provides great flexibility in modelling the response surface, as it can be recursively partitioned into more refined regions over the covariate space. Each individual tree (T_j, M_j) captures a certain pattern from the underlying data. Analogous to renowned ensemble methods such as boosting, BART uses many trees to increase the model's flexibility. The major difference between BART and boosting is that BART uses an iterative backing fitting strategy under a fixed number of trees (m), where in one MCMC iteration, the structure of each tree is updated with the structure of the other trees fixed. This means that BART cycles through the m trees. In contrast, in boosting, each tree is trained sequentially to compensate the weaknesses of its predecessor.

The critical part of the BART model is the regularisation prior on the parameters. Chipman et al. assume the tree components T_j, M_j, $(j = 1, \dots, m)$ are independent of those in the other trees, so that the distribution of the prior is factorised as

$$p((T_1, M_1), \dots, (T_m, M_m), \sigma^2) = \left[\prod_{j=1}^{m} p(T_j, M_j)\right] p(\sigma^2) = \left[\prod_{j=1}^{m} p(M_j | T_j) p(T_j)\right] p(\sigma^2),$$

where $p(M_j|T_j) = \prod_{i=1}^{b_j} p(\mu_{ij}|T_j), \mu_{ij} \in M_j$.

In addition to the independence in the tree components T_j, M_j, the parameters of the terminal nodes within M_j are also independent of each other, given T_j. The regularisation is embedded in the prior for $p(T_j)$. It is composed of three hierarchical parts: (i) the probability of a node being non-terminal; (ii) the distribution of the variable that selects the splitting variable; and (iii) the distribution on the splitting value conditional on the splitting variable. We now briefly state the default BART prior introduced in Chipman et al. (2010) below.

For (i), suppose the node is at depth (the number of edges back to the root) d ($d = 0, 1, \dots$). Then,

$$\Pr(\text{node is an interior node}) = \alpha(1+d)^{-\beta}, \alpha \in (0,1), \beta \in [0, \infty). \tag{5.2}$$

In (5.2), the probability of being an interior node decreases as the node depth increases. The node at a large depth is more likely to be a terminal node than an interior one. It can be interpreted as the prior favouring shallow trees over bushy trees. For (ii), the default BART prior assumes equal probability of the available variables being chosen as the splitting variable. Linero (2018) suggests using a Dirichlet distribution in place of the prior above when the number of predictors is larger than the number of observations in order to achieve better predictive performance. For (iii), a uniform prior is placed on the discrete set of all available splitting values.

The prior $p(\mu_{ij}|T_j)$ is the conjugate normal distribution $N(\mu_\mu, \sigma_\mu^2)$. In this chapter, we set

$$\mu_{ij} \sim N(0, \sigma_\mu^2), \text{with } \sigma_\mu = \frac{1}{\sqrt{m}}, \tag{5.3}$$

assuming that the response variable Y is standardised. For the prior $p(\sigma^2)$, an inverse chi-squared conjugate prior is chosen for σ^2,

$$\sigma^2 \sim \frac{\nu\lambda}{\chi_\nu^2}. \tag{5.4}$$

Throughout this chapter, the default value for ν is 3, and the value of λ is determined by the data such that $\Pr(\sigma^2 < \widehat{\sigma}^2) = 0.9$, where $\widehat{\sigma}^2$ can be an estimate of σ^2 obtained by regression methods. Chipman et al. (2010) provide guidance for choosing these hyperparameters.

5.2.3 The BCF and GRF approaches

This section introduces the BCF and GRF approaches; both inspire the proposed BART-EXT method. They are non-parametric, and enable the estimation of heterogeneous treatment effects in an observational study.

The BCF approach is Bayesian, and is built on the BART model, whereas the GRF method is frequentist and is based on random forests.

The BCF method models the response variable Y_i by

$$Y_i = \mu(\mathbf{x}_i, \widehat{\pi}_i) + \tau(\mathbf{x}_i) z_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2), \tag{5.5}$$

where $\mu(\cdot)$ and $\tau(\cdot)$ are sums of regression trees. The term $\widehat{\pi}_i$ is the estimated propensity score. The model can be viewed as a linear regression in z_i with covariate-dependent functions for the slope and the intercept. Compared with the original BART model, the CATE estimation is modelled directly in (5.5) by $\tau(\mathbf{x}_i)$. In Hahn et al. (2020), the treatment coding level z_i is also modelled as a variable b_{z_i}, because various coding levels (e.g. $z_i \in \{0, 1\}$ or $z_i \in \{\pm 0.5\}$) affect posterior inference. The final model specification of the BCF method is

$$\begin{aligned} Y_i &= \mu(\mathbf{x}_i, \widehat{\pi}_i) + \tilde{\tau}(\mathbf{x}_i) b_{z_i} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2), \\ b_{z_i} &\sim N(0, 0.5), \quad z_i \in \{0, 1\}. \end{aligned} \tag{5.6}$$

From (5.6), the CATE estimate using the BCF approach is

$$\tau(\mathbf{x}_i) = (b_1 - b_0) \tilde{\tau}(\mathbf{x}_i).$$

The GRF approach uses many regression trees for causal inference. The intuition behind GRF is that if the leaf L is small enough to contain observations from both treatment and control groups, then those observations will share similar covariate values as though they were from a randomised experiment. The treatment effect of $\mathbf{x}_i$ in the leaf L of the jth tree is estimated as

$$\widehat{\tau}_j(\mathbf{x}_i) = \frac{1}{|\{i : z_i = 1, \mathbf{x}_i \in L\}|} \sum_{\{i:z_i=1,\mathbf{x}_i \in L\}} Y_i - \frac{1}{|\{i : z_i = 0, \mathbf{x}_i \in L\}|} \sum_{\{i:z_i=0,\mathbf{x}_i \in L\}} Y_i.$$

CATE estimation by GRF aggregates the predictions of each tree by averaging:

$$\widehat{\tau}(\mathbf{x}_i) = \frac{1}{m} \sum_{j=1}^{m} \widehat{\tau}_j(\mathbf{x}_i). \tag{5.7}$$

Wager and Athey (2018) show that predictions by GRF are asymptotically Gaussian and unbiased. The condition that has to be satisfied to achieve these theoretical results, is that for each training sample, the tree uses its response to estimate the causal effect or the tree uses its covariates to decide the splitting variable and value, but not both.

Section 5.3 introduces the BART-EXT method, which combines the ideas behind BCF and GRF. The similarities and differences BART-EXT and the other two methods are also discussed there.

5.3 Methodology

5.3.1 An extension to the BART model (BART-EXT)

Inspired by BCF (Hahn et al., 2020) and GRF (Wager and Athey, 2018), we propose BART-EXT as an extension to the BART model,

$$Y_i = \begin{cases} \sum_{j=1}^{m_0} g(\mathbf{x}_i; T_j^{'}, M_j^{'}) + \sum_{k=1}^{m_1} g(\mathbf{x}_i; T_k, M_{1k}) + \epsilon_i, & \text{if } z_i = 1, \\ \sum_{j=1}^{m_0} g(\mathbf{x}_i; T_j^{'}, M_j^{'}) + \sum_{k=1}^{m_1} g(\mathbf{x}_i; T_k, M_{0k}) + \epsilon_i, & \text{if } z_i = 0, \end{cases} \qquad \epsilon_i \sim N(0, \sigma^2); \tag{5.8}$$

$i \in \{1, \ldots, n\}$ indexes the observations, T_j' is the structure of the jth tree for capturing the baseline impact, and T_k is the structure of the kth tree of the treatment contribution. Correspondingly, M_j' stands for the node parameters of T_j' with $M_j' = \{\mu_{1j}', \ldots, \mu_{n_j j}'\}$, and M_{1k}, M_{0k} for the node parameters attached to T_k with $M_{zk} = \{\mu_{z,1k}, \ldots, \mu_{z,n_k k}\}$,

$z \in \{0, 1\}$. The number of parameters in M_{0k} and M_{1k} is the same because they share the identical structure T_k.

The conditional mean is composed of the baseline and treatment effects. By omitting the tree structure and node parameters, (5.8) simplifies to

$$Y_i = \begin{cases} \mu(\mathbf{x}_i) + \tau_1(\mathbf{x}_i) + \epsilon_i, & \text{if } z_i = 1, \\ \mu(\mathbf{x}_i) + \tau_0(\mathbf{x}_i) + \epsilon_i, & \text{if } z_i = 0, \end{cases} \tag{5.9}$$

where $\mu(\cdot)$, $\tau_0(\cdot)$ and $\tau_1(\cdot)$ are sums of separate regression trees.

Given the strong ignorability and SUTVA conditions, the CATE is estimated by

$$\widehat{\tau}(\mathbf{x}_i) = \widehat{\tau}_1(\mathbf{x}_i) - \widehat{\tau}_0(\mathbf{x}_i) = \sum_{k=1}^{m_1} \widehat{g}(\mathbf{x}_i; T_k, M_{1k}) - \sum_{k=1}^{m_1} \widehat{g}(\mathbf{x}_i; T_k, M_{0k}).$$

For the trees modelling the baseline impact, (5.8) exploits a specific tree structure with identical node values for the baseline estimate regardless of the treatment status. The trees capturing the treatment effect share the same tree structure T_k, but the node values may vary based on the treatment assignment. This part is analogus to the tree structure in Wager and Athey (2018), where each individual tree is constructed by the same splitting rules for all internal nodes except for the bottom layer, which implies that the last split occurs at the treatment indicator. The treatment effect can then be estimated by accumulating the difference between the values of two leafs from the same terminal node.

In Wager and Athey (2018), the predictions are obtained by averaging the prediction of each tree as in (5.7). Our proposed method (BART-EXT) uses posterior predictions for inference on $\tau(\mathbf{x}_i)$ directly. The identical structure also reduces the need to obtain a good estimate of the propensity score, which is strongly advocated in the BCF approach to reduce the bias. Here, the estimated propensity score is provided as a covariate in the model, which is fixed before the start of the MCMC. In a high-dimensional setting, it is usually difficult to obtain a good estimate of the propensity score. Unlike BCF, BART-EXT is less dependent on the estimate of the propensity score due to the specific structure of trees modelling treatment effects.

Figure 5.1 illustrates BART-EXT and the other methods (vanilla BART/ps-BART, BCF and GRF). The vanilla BART uses $\mathbf{x}$ and z as covariates. The ps-BART includes the estimated propensity score together with $\mathbf{x}$ and z as covariates.

Section 5.3.3 discusses the differences and similarities between these methods.

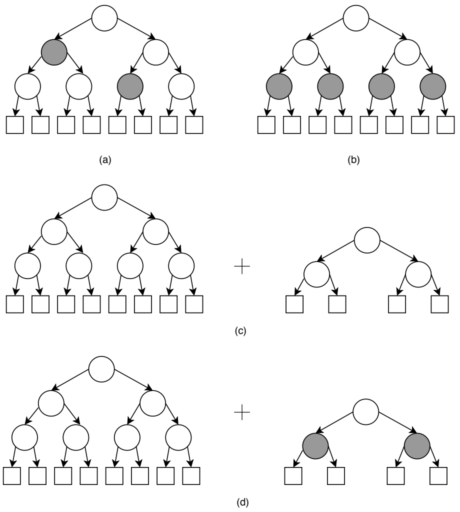


Figure 5.1: A graphical illustration for vanilla BART /BART with propensity score (top left), GRF (top right), BCF (middle) and our proposal (BART-EXT) (bottom). The circles and squares represent interior nodes and terminal nodes (leaves) respectively. The shaded circle is where the tree splits at the binary treatment indicator (Z). The illustration uses one or two trees to describe the tree structure, whereas in real applications, each method involves a group of trees.

5.3.2 Prior specification

We propose different hyperparameters for the prior on T'_j (the tree structure for the baseline) and T_k (the tree structure for treatment effects) in (5.8). For the prior on T'_j, the default suggestion in Chipman et al. is adopted with $\alpha = 0.95$, $\beta = 2$, and $m_0 = 200$. For the prior on T_k, we use $\alpha = 0.25$, $\beta = 3$ and $m_1 = 50$ as proposed by Hahn et al. (2020). With such a setting, the CATE estimate is shrunk towards the homogeneous treatment effect by favouring shallow trees over bushy ones. For the prior on the parameter nodes M'_j, M_{0k}, M_{1k} and the variance σ^2, we follow the priors from Chipman et al. (1998), specified in (5.3) and (5.4).

We also consider a half-Cauchy prior for σ_μ in (5.3), proposed in Hahn et al. (2020), which exhibits heavier tails than a normal distribution. As indicated by results of the simulations and the empirical study in later sections, we do not find much difference between the half-Cauchy prior and the choice adopted in this chapter ($\sigma_u = 1/\sqrt{m}$).

5.3.3 Connection with GRF and BCF

We first state the motivation for the BART-EXT by presenting the weakness of a standard regression tree, which serves as the workhorse for all tree-based methods in causal inference. We then illustrate the "regularisation-induced confounding" (RIC) introduced by Hahn et al. (2018) and explain why vanilla BART suffers from such a phenomenon (Hahn et al., 2020). Finally, we discuss the commonalities and differences between GRF, BCF and BART-EXT.

The problem with a standard regression tree for causal inference

The standard regression tree targets minimising the mean squared error of predictions to determine the splitting rule, defined as

$$\text{SSE} = \sum_{i \in S_1} (y_i - \overline{y}_1)^2 + \sum_{i \in S_2} (y_i - \overline{y}_2)^2,$$

where $\overline{y}_1$ and $\overline{y}_2$ are the mean values for the corresponding observations on the same leaf. Under such a scheme, the treatment indicator has priority to be selected if it results in a smaller SSE compared with the splitting on one of the other covariates. In the following paragraphs, we use an intuitive example to illustrate such splitting behaviour.

Consider the data generation process for the observation i, $(i = 1, \dots, n)$ as

$$y_i = x_{1i} + x_{2i} + \tau_i z_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2),$$

and suppose the propensity score is

$$\Pr(z_i = 1|\mathbf{x}_i) = \begin{cases} 0.8, & \text{if } x_{1i} \geq 0.8, \\ 0.2, & \text{otherwise.} \end{cases}$$

To simplify the computation, we also assume the homogeneous treatment effect is $\tau = 1$, and $x_{1i}, x_{2i} \overset{\text{iid}}{\sim} \text{Uniform}(0, 1)$. In a standard regression tree, the treatment indicator z is treated similarly to the covariates x_1, x_2. Suppose the root node splits on z. Then the SSE can be expressed as

$$\text{SSE}_z = \sum_{\{i:z_i=1\}} (y_i - \overline{y}_{z_i=1})^2 + \sum_{\{i:z_i=0\}} (y_i - \overline{y}_{z_i=0})^2.$$

As evaluation of the equation above needs realised observations, we analyse its expectation instead:

$$\begin{aligned} E(\text{SSE}_z) &= \sum_{\{i:z_i=1\}} E[(y_i - \overline{y}_{z_i=1})^2] + \sum_{\{i:z_i=0\}} E[(y_i - \overline{y}_{z_i=0})^2] \\ &= (n_0 + n_1)(\text{Var}(x_1) + \text{Var}(x_2)) + (n_0 + n_1)\sigma^2, \end{aligned}$$

where n_0, n_1 respectively represent the number of observations in the control and the treatment groups [1].

Alternatively, if the split is on x_1 with the cut-off value c in the root node, the SSE can be expressed as

$$\text{SSE}_{x_1} = \sum_{\{i:x_{1i}\geq c\}} (y_i - \overline{y}_{x_{1i}\geq c})^2 + \sum_{\{i:x_{1i}<c\}} (y_i - \overline{y}_{x_{1i}<c})^2.$$

The expectation of SSE_{x_1} is

$$\begin{aligned} E(\text{SSE}_{x_1}) &= \sum_{\{i:x_{1i}\geq c\}} E[(y_i - \overline{y}_{x_{1i}\geq c})^2] + \sum_{\{i:x_{1i}<c\}} E[(y_i - \overline{y}_{x_{1i}<c})^2] \\ &= n_0'\text{Var}(x_1|x_1 \geq c) + n_1'\text{Var}(x_1|x_1 < c) + (n_0' + n_1')\text{Var}(x_2) \\ &\quad + \tau^2 n_0'\text{Var}(z|x_1 \geq c) + \tau^2 n_1'\text{Var}(z|x_1 < c) + (n_0' + n_1')\sigma^2, \end{aligned}$$

[1] Taking expectations generates the sample variance $s^2(x_1)$ ($\text{Var}(x_1) = E((x - E(x))^2)$, here we use $\overline{x}$ instead of its expectation) instead of $\text{Var}(x_1)$, and similarly for the variance associated with the error term σ^2. Assuming n_0 and n_1 are both relatively large (for this discussion), we cannot distinguish the difference between sample variance and the true variance.

where n'_0, n'_1 are the number of observations with $x_{1i} \geq c$ and $x_{1i} < c$, $n'_0 + n'_1 = n_0 + n_1$. Under the setting where τ is relatively large compared to x_1 and x_2 ($\tau = 1$, $x_1, x_2 \sim$ Uniform$(0, 1)$), and with the uniform assumption on x_1, x_2, we have

$$\begin{aligned} E(\text{SSE}_z) &= \frac{n_0 + n_1}{6} + (n_0 + n_1)\sigma^2, \\ E(\text{SSE}_{x_1}) &= n'_0 \frac{(1-c)^2}{12} + n'_1 \frac{c^2}{12} + \frac{n'_0 + n'_1}{12} + 0.16(n'_0 + n'_1) + (n'_0 + n'_1)\sigma^2. \end{aligned}$$

It is easy to verify that $E(\text{SSE}_{x_1}) > E(\text{SSE}_z)$, which in turn results in the node splitting on z. If the tree does not split further, the expectation for $\hat{\tau}$: $E(\hat{\tau}) = E(\overline{y}_{z=1} - \overline{y}_{z=0}) = \tau + E(x_1|z = 1) - E(x_1|z = 0) = 1.3 > 1$, which overstates the true treatment effect. Such bias can possibly be reduced by further splitting on the nodes. Given that a finer covariate space is formed, the prediction is then obtained by observations with covariate values having greater similarity. However, once the split occurs at the treatment indicator, it is equivalent to the approach that models treatment and control responses separately as discussed in Section 5.2.1. For an arbitrary tree, splitting on the treatment indicator may occur on some branches, which in turns affects the estimates for observations passing along those branches. Unless the counterfactual is independent of the treatment assignment conditioning on previous splitting rules, a reliable estimate cannot be guaranteed.

For the BART model, the prior on selecting the splitting variable assigns equal probability on the available variables, so that the corresponding tree does not necessarily split on z. Meanwhile, the prior on the node parameters is a normal distribution with the standard deviation $1/\sqrt{m}$, which concentrates on 0 given a large m. Such a strong prior helps to avoid treatment effect overestimation, to some extent. However, researchers cannot control the split behaviour on the treatment indicator either in random forests or the standard BART model. Consequently, these approaches bias the treatment effect for at least some observations. On the other hand, the regularisation prior advocates splitting on the treatment indicator z instead of the covariates $\mathbf{x}$ given that z is a good predictor, which can make the problem worse. See below for further discussion.

Regularisation-induced confounding

"Regularisation-induced confounding" (RIC) refers to a phenomenon whereby the usage of a regularisation prior on the coefficients of the covariates leads to a biased treatment effect estimate (Hahn et al., 2018).

Hahn et al. (2020) illustrate the issue in the BART context. In the vanilla BART model, the treatment indicator is treated the same as the covariates. If the splitting on the treatment indicator brings similar prediction power to splitting on the covariates, then the tree is more likely to split on the treatment indicator. This is because the BART prior favours shallow trees. In other words, the treatment indicator and many covariates compete to be the best predictor, and BART tends to select the former because of the regularisation prior. The key observation in Hahn et al. (2018) is that including the estimated propensity score as one of the covariates mitigates RIC in the linear model. In a non-linear setting, including the estimates of propensity score can improve the ATE and CATE estimates in the presence of moderate to strong confounding, as pointed out in Hahn et al. (2020); see (5.6). In implementing the BART-EXT method, we also include the estimated propensity score as one of the covariates.

Similarities and differences in BART-EXT, BCF and GRF

As a non-parametric Bayesian method, the BART-EXT approach distinguishes itself from the GRF, which stems from random forests. For the GRF, all the plausible trees constructed in random forests are searched greedily to model the response surface, whereas BART models the posterior distribution of such trees directly. Another difference is that BART and its extensions are the analogues of boosting methods, which use sums over trees to represent the underlying surface. Random forests, on the other hand, work on the bagging principle where each tree is generated from random samples of the data set. The final result is based on the average value obtained from the predictions of all trees. In Wager and Athey (2018), obtaining the asymptotically unbiased estimator requires an appropriate subsample size to construct individual trees, which scales at a specific rate with the total number of observations. However, such a requirement is not satisfied in most applications. Based on the simulation results in Section 5.4, the GRF does not outperform BART and its extensions in various settings. The results of Dorie et al. (2019);

Hahn et al. (2020); Wendling et al. (2018) also support such findings.

To mitigate the weakness of a standard regression tree in estimating causal effects, BCF adopts different strategies from GRF. It models the prognostic and the treatment effect separately by reparameterising the model as

$$E(y_i|\mathbf{x}_i, z_i) = \mu(\mathbf{x}_i) + \tilde{\tau}(\mathbf{x}_i)b_{z_i},$$

with both $\mu(\mathbf{x}_i)$ and $\tilde{\tau}(\mathbf{x}_i)$ are estimated by sets of trees; see (5.6). The trees for $\tilde{\tau}(\mathbf{x}_i)$ shrink more strongly toward homogeneous treatment effects. BCF includes the estimated propensity score as one of its covariates to reduce RIC. Even though the approach excludes the treatment indicator z_i as one of the variable candidates for tree splitting, which means that the splitting on z_i never occurs in the BCF, it is unclear whether this method of constructing trees avoids RIC. For example, in the area of covariate space dominated by the treated subjects, the corresponding tree structure is heavily influenced by the features of the treatment group. Without the support from observations in the control group, such tree structures may overestimate or underestimate the treatment effect.

The BART-EXT is similar to GRF with respect to the tree structures. In GRF, all the leafs of each tree contain units from the treated and the control groups. The equivalent approach in BART-EXT is that for the trees modelling causal effects, the treated and control units share the same tree structure, with the terminal nodes splitting at the treatment indicator. Such an approach mimics the classic idea of matching, which has not previously been implemented in standard BART and BCF. Matching is an intuitive idea that given one pair of observations with different treatment assignments, but almost the same values of the covariates, the average difference of the outcome can be used as an estimate of the treatment effect. In a high-dimensional setting, it is usually hard to find the matching pairs exactly. Tree-based methods solve the problem by finding the nearest neighbours adaptively. The matching is executed via a “coarsened” covariate value with one splitting process. Finally, the observations falling on the same leaf node are the ones filtered by the common splitting rules of the ancestral nodes, leading to a similarity between covariate values.

Table 5.1 provides a brief comparison between the methods.

Method	$E(Y_i \vert \mathbf{X}_i = \mathbf{x}_i, Z_i = z_i)$	Description
BART-EXT	$\mu(\mathbf{x}_i) + \tau_{z_i}(\mathbf{x}_i)$	An extension to BART with identical tree structures for the treated and untreated subjects in trees estimating $\tau_{z_i}(.)$
BART/ps-BART	$\mu(\mathbf{x}_i, z_i)$	Original BART method: Both $\mathbf{x}$ and z are treated as covariates. ps-BART: Estimates of propensity score are also included as one of covariates.
BCF	$\mu(\mathbf{x}_i) + \tilde{\tau}(\mathbf{x}_i)b_{z_i}$	An extension to BART, modelling treatment effects and treatment level together by $\tilde{\tau}(.)$ and b_{z_i} respectively.
GRF	$\frac{1}{m}\sum_{j=1}^{m} \mu_j(\mathbf{x}_i, z_i)$	An extension to random forests with identical tree structures for treatment and control groups.

Table 5.1: Comparison between BART-EXT, BART (ps-BART), BCF and GRF. For BART related methods (BART-EXT, BART/ps-BART, BCF), $\mu(\cdot)$, $\tau(\cdot)$, $\tilde{\tau}(\cdot)$ refer to the sum over multiple regression trees. For GRF, $\mu_j(\cdot)$ is the prediction given by the jth tree.

5.3.4 Algorithm

We adopt the iterative Bayesian backfitting Markov chain Monte Carlo (MCMC) algorithm in Chipman et al. (2010). Appendix C describes the computational details of the algorithm. The algorithm is implemented in R using the Rcpp package.

5.4 Simulation studies

This section first illustrates the strengths and weaknesses of BART-EXT compared with other popular methods via a toy example in Section 5.4.1. As argued in Dorie et al. (2019), some methods work particularly well for a certain data generating process, whereas they fail for other scenarios. No method dominates the others in all cases. However, it is worthwhile to compare the performance of different methods in a wider setting. Section 5.4.2 implements our model for the 2016 ACIC data challenge (2016 Atlantic causal inference conference competition). Dorie et al. (2019) analyse this challenge using 30 competitors, including BART and its extensions. The challenge attempts to address the limitations of methods in the causal inference literature by using many synthetic data sets.

5.4.1 A low-dimensional but complicated model

We start with a simple toy example to illustrate the effects of RIC on various tree-based methods. The data generating process consists of the following components,

$$\begin{aligned}
\text{response equation:} \quad & \mu_i = -\mathbb{1}(x_{i,2} > x_{i,1}) + \mathbb{1}(x_{i,1} > x_{i,2}) + 0.5x_{i,3}, \\
& Y_i = \mu_i + \tau_i z_i + \epsilon_i, \quad \epsilon_i \sim N(0,1) \\
\text{selection equation:} \quad & \Pr(z_i = 1 | x_{i,1}, x_{i,2}) = \Phi(0.6\mu_i), \\
\text{treatment effect:} \quad & \tau_i = 0.25\mathbb{1}(x_{i,2} > 2) + 0.25x_{i,1} + 0.25,
\end{aligned}$$

where $i = 1, \ldots, n$, with $n = 250, 500$, and $x_{i,j} \overset{\text{iid}}{\sim} N(0, 0.5^2)$ with $j = 1, 2, 3$. The indicator function $\mathbb{1}(A) = 1$ if statement A is true, otherwise $\mathbb{1}(A) = 0$. The function $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

The following features are noted for this data generating process.

1. The dimension of the covariates is relatively small compared with the number of observations. However, both the response surface and the treatment effect are non-linear. There exists an independent variable x_3 affecting the baseline surface only.

2. There appears to be moderate correlation between the propensity score and the response equation. When the response surface (μ_i) has a high value, the unit is more likely to receive treatment as the right panel of Figure 5.2 shows. Simply taking the mean difference overestimates the treatment effect for most observations, as the left panel of Figure 5.2 shows.

3. The treatment effect is heterogeneous, with a relatively larger magnitude compared with the response surface.

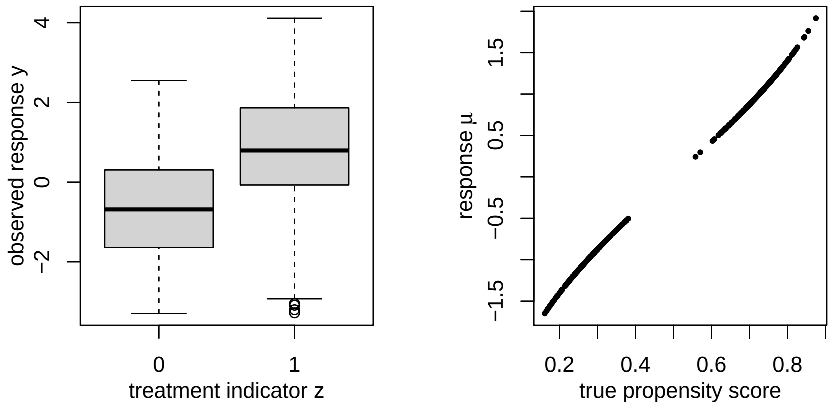


Figure 5.2: Left panel: The boxplot of observed response y against the treatment indicator z. Right panel: The boxplot of the baseline response μ_i against the propensity score $\Pr(z_i = 1|x_{i,1}, x_{i,2})$.

We apply the following methods together with the BART-EXT to the simulated data set. Without extra specifications, all the methods are implemented using the default setting provided in the corresponding packages. For the Bayesian methods, we use 4,000 iterations for the MCMC sampler with the first 2,000 iterations discarded as burn-in. To eliminate the influence by the estimated propensity scores, the true propensity score (which is often unavailable in observational studies) is provided.

1. BART-EXT: The prior is described in Section 5.3.2. We use $m_0 = 200$ and $m_1 = 50$ trees to model the response surface and the treatment effect, respectively. The true propensity score is included in both parts of the tree.

2. Vanilla BART: The default setting in `R` package `BART` (version 2.0) is adopted without the propensity score as one of the covariates.

3. ps-BART: This method is identical to vanilla BART except that the true propensity score is provided as one of the covariates.

4. BCF: The `R` package `BCF` (version 1.3) is used to implement the method proposed by Hahn et al. (2020), with the true propensity score included.

5. GRF: The R package `grf` (version 0.10.4) is applied with 4,000 trees.

6. OLS: The regular ordinary least squares regression method.

Table 5.2 shows the average results of the CATE estimation based on 100 independent replications of the simulated data. The metrics RMSE and MAE are, respectively, defined as:

$$\text{RMSE} = \sqrt{n^{-1}\sum_{i=1}^{n}(\widehat{\theta}_i - \theta)^2}, \quad \text{MAE} = n^{-1}\sum_{i=1}^{n}|\widehat{\theta}_i - \theta|,$$

where $n = 250$, 500, and $\widehat{\theta}_i$ refers to the CATE estimation of the ith unit. For the Bayesian methods, $\widehat{\theta}_i$ is the posterior mean; for the GRF, $\widehat{\theta}_i$ is the prediction averaging over all trees. The RMSE of CATE estimates is also known as PEHE (precision in estimating heterogeneous effects) (Hill, 2011).

The BART-EXT yields the smallest RMSE and MAE for $n = 250$ and $n = 500$. The results of all the methods improve with a larger number of observations, except for OLS. In terms of credible/confidence interval coverage and its length, the tree-based methods do a relatively good job, especially the Bayesian methods, compared with GRF and OLS. The result of BART-EXT is close to that of ps-BART in coverage, but with a narrower interval length (15% to 20% less). The BCF has coverage rate closest to 95% with the interval length even narrower than that of BART-EXT. We note that the true propensity score is provided for all BART-related methods, except for vanilla BART. In a high-dimensional setting, it is often hard to get a precise estimate of the propensity score. Here, BART-EXT has an advantage over ps-BART, because the method does not rely heavily on propensity scores, due to the identical structures imposed on the trees. The next simulation study explores a high-dimensional problem.

It has to be admitted that we may be asking the data for too much on inference for the treatment effect as the response surface is complex with a relatively large noise. Table 5.2 (the cases of $n = 250$ and $n = 500$) shows that a large sample size helps to improve the estimates for most methods. Figure 5.3 shows the CATE estimate (vertical axis) with its true value (horizontal axis) for one simulated data set with $n = 500$. From this figure, none of the methods capture the true treatment effects particularly well. BART-EXT and ps-BART are arguably the best. Vanilla BART and the GRF overestimate the treatment effects for a large portion of the observations, whereas BCF shrinks the estimates towards some mean values without capturing the variability in the causal effects.

n=250	RMSE	MAE	Coverage	Length
BART-EXT	**0.180**	**0.152**	0.986	1.001
	(6.840×10^{-2})	(6.313×10^{-2})	(4.011×10^{-2})	(4.987×10^{-2})
vanilla.BART	0.389	0.354	0.876	1.269
	(1.442×10^{-1})	(1.441×10^{-1})	(1.727×10^{-1})	(7.386×10^{-2})
ps-BART	0.202	0.169	0.993	1.175
	(7.139×10^{-2})	(6.490×10^{-2})	(2.243×10^{-2})	(7.071×10^{-2})
BCF	0.196	0.166	**0.969**	0.933
	(7.088×10^{-2})	(6.434×10^{-2})	(6.979×10^{-2})	(1.549×10^{-1})
GRF	0.341	0.311	0.713	0.838
	(1.340×10^{-1})	(1.361×10^{-1})	(2.612×10^{-1})	(7.075×10^{-2})
OLS	0.400	0.379	0.386	0.626
	(1.498×10^{-1})	(1.557×10^{-1})	(3.120×10^{-1})	(3.329×10^{-2})
n = 500	RMSE	MAE	Coverage	Length
BART-EXT	**0.152**	**0.126**	0.990	0.843
	(4.875×10^{-2})	(4.423×10^{-2})	(2.975×10^{-2})	(4.405×10^{-2})
vanilla BART	0.361	0.324	0.874	1.128
	(9.857×10^{-2})	(9.759×10^{-2})	(1.336×10^{-1})	(5.762×10^{-2})
ps-BART	0.180	0.147	0.993	1.035
	(5.179×10^{-2})	(4.483×10^{-2})	(1.900×10^{-2})	(5.658×10^{-2})
BCF	0.168	0.141	**0.958**	0.742
	(5.276×10^{-2})	(4.824×10^{-2})	(8.436×10^{-2})	(1.314×10^{-1})
GRF	0.288	0.255	0.723	0.712
	(8.744×10^{-2})	(8.764×10^{-2})	(1.944×10^{-1})	(4.788×10^{-2})
OLS	0.420	0.400	0.151	0.442
	(1.136×10^{-1})	(1.185×10^{-1})	(1.769×10^{-1})	(1.501×10^{-2})

Table 5.2: The CATE estimation results from 100 independent replications with $n = 250, 500$ observations. The results are the mean values of RMSE, MAE, coverage and interval length, where coverage and interval length are reported as 95% credible or confidence intervals. Numbers in parentheses are standard deviations.

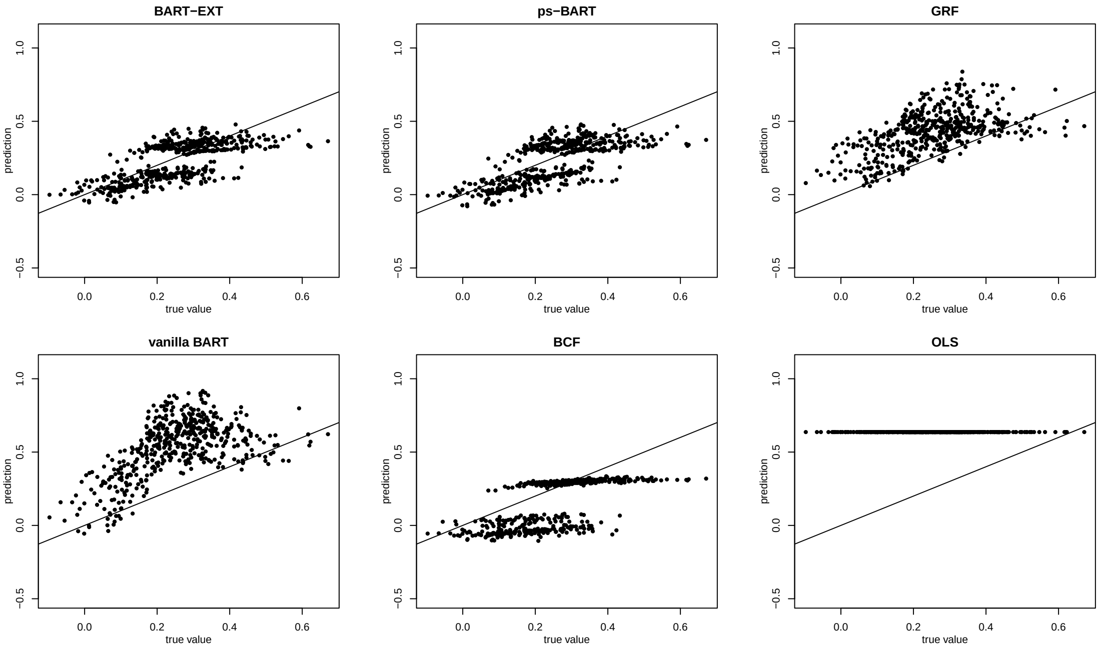


Figure 5.3: Comparisons between the true treatment effect (horizontal axis) and the prediction (vertical axis) obtained by different methods with the top row: BART-EXT, ps-BART, GRF and the bottom row: vanilla BART, BCF, OLS. The result is for a single replication.

5.4.2 2016 ACIC data challenge

A wide range of methodologies are available for causal inference. However, most literature on causal inference advocates their proposed method by comparing it with the existing ones in simulation studies only. Such comparisons are limited in the sense that they only involve a specific data generating process which might favour the advocated method. The 2016 ACIC data challenge is motivated by observing the shortcomings of such comparisons. It was initiated as an attempt to compare the performance of numerous methods on a large synthetic data set, measured by various metrics. The full data sets contain 77 scenarios, each with 100 repetitions of 4,802 observations, available from the website `http://jenniferhill7.wixsite.com/acic-2016/competition`. The simulation framework considers different response models, treatment assignment mechanisms, heterogeneity levels, etc.; see Dorie et al. (2019) for more details.

As BART and its extensions perform better than the other methods in estimating the heterogeneous treatment effects in this challenge, we implement the same methods in Section 5.4.1 on the full data sets. For all the Bayesian methods considered, we run the MCMC sampler for 4,000 iterations, retaining the second half of the iterates for inference. The true propensity scores are replaced by the estimates obtained using BART, where we use the `pbart` function in the `BART` package to fit the binary response (treatment level). The parameters of interest here are the CATE and ATT, which Dorie et al. (2019); Hahn et al. (2020) also target. The performance of the methods is measured against various metrics: RMSE, 95% credible/confidence interval coverage rate, interval length, bias and absolute bias. Note that the simulated data sets only provide the true CATE. The true sample ATT (SATT) is obtained by averaging all the CATE across the treated observations. As the SATT is a scalar, the RMSE results of the Bayesian methods are calculated based on the MCMC samplers, i.e., RMSE $= \sqrt{M^{-1}\sum_{m=1}^{M}(\widehat{\theta}_m - \theta)^2}$, where $\widehat{\theta}_m$ is the estimate of the mth samplers.[2] To ensure that the estimated treatment effects are on a similar scale across different data sets, the final estimates and true values are scaled down by the standard deviation of the observed outcomes.

Table 5.3 presents the average results for all the methods being considered. For the CATE results, BART-EXT has the lowest RMSE, bias and absolute bias among all the methods. The coverage of the confidence interval is 4-5% lower than BART and ps-BART, but with a 50% narrower interval. The BCF has a narrower interval, but with lower coverage. The GRF performs poorly for CATE across all metrics.

For the SATT results, BART-EXT performs comparatively to BCF, which yields the lowest RMSE. ps-BART has the highest coverage, but at the cost of wider intervals (50% wider) compared with that of BART-EXT. All BART-related methods outperform the GRF in the SATT setting as well. However, all methods fail to achieve 95% coverage, which indicates overfitting, although the Bayesian tree-based methods achieve the best relative coverage for SATT (over 80%).

[2] We make this choice to account for variation in the posterior distribution. Otherwise, the RMSE results of SATT are the same as |Bias|.

CATE	RMSE	Coverage	Interval length	Bias	|Bias|
BART-EXT	**0.19**	0.69	0.20	**0.0011**	**0.089**
	(7.76×10^{-2})	(6.23×10^{-2})	(3.87×10^{-2})	(2.56×10^{-3})	(2.84×10^{-2})
BCF	0.21	0.59	0.17	0.0016	0.096
	(8.07×10^{-2})	(5.83×10^{-2})	(3.85×10^{-2})	(3.23×10^{-3})	(2.96×10^{-2})
BART	0.23	0.72	0.40	0.0012	0.12
	(8.62×10^{-2})	(3.86×10^{-2})	(9.80×10^{-2})	(3.08×10^{-3})	(3.66×10^{-2})
ps-BART	0.24	**0.74**	0.40	0.0018	0.124
	(8.71×10^{-2})	(4.97×10^{-2})	(9.79×10^{-2})	(1.16×10^{-2})	(3.83×10^{-2})
GRF	0.34	0.61	0.35	0.016	0.19
	(1.23×10^{-1})	(9.30×10^{-2})	(6.26×10^{-2})	(5.98×10^{-3})	(6.25×10^{-2})
OLS	0.76	0.078	0.087	0.037	0.54
	(1.94×10^{-1})	(7.22×10^{-2})	(8.28×10^{-3})	(2.62×10^{-2})	(1.33×10^{-1})
ATT	RMSE	Coverage	Interval length	Bias	|Bias|
BART-EXT	**0.012**	0.84	0.028	0.0013	0.011
	(2.47×10^{-3})	(7.81×10^{-2})	(2.88×10^{-3})	(1.87×10^{-3})	(2.41×10^{-3})
BCF	**0.012**	0.82	0.027	**0.0011**	**0.010**
	(2.11×10^{-3})	(7.75×10^{-2})	(1.78×10^{-3})	(1.74×10^{-3})	(2.09×10^{-3})
BART	0.017	0.82	0.041	0.0018	0.014
	(3.80×10^{-3})	(8.05×10^{-2})	(6.78×10^{-3})	(2.40×10^{-3})	(3.42×10^{-3})
ps-BART	0.015	**0.86**	0.037	0.0014	0.013
	(3.08×10^{-3})	(7.32×10^{-2})	(5.72×10^{-3})	(2.98×10^{-3})	(2.85×10^{-3})
GRF	0.025	0.62	0.056	0.0184	0.025
	(6.90×10^{-3})	(1.49×10^{-1})	(8.81×10^{-3})	(6.80×10^{-3})	(6.90×10^{-3})
OLS	0.112	0.22	0.087	0.0449	0.112
	(4.82×10^{-2})	(2.53×10^{-1})	(8.28×10^{-3})	(3.46×10^{-2})	(4.82×10^{-2})

Table 5.3: The result of the 2016 ACIC data challenge, which collects the average value of difference metrics across 7,700 cases with standard deviation included in parenthesis. Coverage and interval length refer to 95% credible or confidence intervals.

5.5 An empirical study

This section illustrates the strength of our proposal through an experimental study. Unlike observational studies, experimental studies usually adopt randomisation from which the ATE can be estimated unbiasedly by a simple linear regression. Unbiasness is not affected by the missing confounders or model misspecification (Lei and Ding, 2021). This property facilitates a direct comparison between various methods and the “true value” obtained by an OLS regression.

Allcott and Knittel (2019) study the effect of providing the information about the fuel economy of a vehicle on consumers’ purchasing behaviour. A small-scale dealership experiment and a large-scale online experiment are conducted among participants who intend to buy a vehicle within the next six months. Here, we focus on the online experiment only, as the tree-based methods perform better in a large data setting. In the online experiment, potential participants are screened against a number of criteria for eligibility. After that, they are randomised into either the treatment or the control group. Participants in the control group answer several basic questions, including their first and second choice of car purchase. Those in the treatment group answer the same questions, but with additional information on the fuel economy provided. For example, given the participants' first and second choice, collectively denoted as the consideration set, the highest-MPG (miles per gallon) vehicle in the same class is shown on screen with annual and lifetime fuel costs. A follow-up survey collects the MPG of the vehicles purchased by the participants, if the purchases are made by the time of the follow-up survey. The information from 1,489 participants are analysed. For the analysis, the following estimation equation is used,

$$Y_i = \tau z_i + \mathbf{x}_i \boldsymbol{\beta} + \epsilon_i,$$

where the dependent variable Y_i is the fuel intensity of the vehicle purchased by consumer i, measured in gallons per 100 miles; the treatment indicator is z_i, which identifies if fuel economy information is provided ($z_i = 1$) or not ($z_i = 0$). The details of the covariates $\mathbf{x}$ are given below.

Male: The gender indicator, with 1 for male and 0 for female.

Age: The age of the participant.

Caucasian: The race indicator for Caucasian people.

lnIncome: The natural log of annual income (1 unit = 1,000 dollars)

Miles_per_year: Miles driven per year (1 unit = 1,000 miles).

current_Ford: Whether the consumer's current vehicle is a Ford.

current_SelfCombGP100M: Fuel intensity of the current vehicle (gallons/100 miles).

considerSet_GP100M: The average fuel intensity of the consideration set, which includes consumers' first and second choice (gallons/100 miles).

Allcott and Knittel (2019) analyse the ATE of providing information on the fuel intensity of purchased vehicles by OLS. The heterogeneous treatment effects are also investigated on subgroups formed arbitrarily by user-defined criteria in Allcott and Knittel (2019). We do not need to reanalyse the subgroups as all tree-based methods provide estimates of the CATE.

We now implement the same set of methods examined in Section 5.4.2. As the study is a randomised experiment, RIC is not induced. However, we still put the BCF and ps-BART in the method pool to check their performance. Table 5.4 depicts the estimates of the ATE, together with the standard deviation, confidence or credible interval, and its length. The treatment here is the provision of the fuel information. A negative treatment effect indicates that the consumers buy a less fuel-intensive (more economical) vehicle given the fuel information, which implies positive feedback from the treatment. Given the randomisation is still valid after subject attrition (the loss of subjects in the sample) between the baseline and follow-up surveys, the OLS result is treated as the benchmark here. Except for the BCF, all the other methods generate close estimates to OLS, but with a larger standard deviation and wider interval length. Figure 5.4 gives the histograms of the estimated CATE of all participants. BART-EXT, vanilla BART and ps-BART all

	Mean	Sd	Lower bound	Upper bound	Interval length
BART-EXT	0.034	0.089	−0.125	0.132	0.256
vanilla BART	0.034	0.099	−0.151	0.153	0.304
ps-BART	0.036	0.099	−0.134	0.143	0.278
BCF	0.017	0.016	−0.016	0.038	0.055
GRF	0.027	0.063	−0.084	0.153	0.236
OLS	0.031	0.041	−0.051	0.112	0.163

Table 5.4: The ATE estimates for the fuel intensity of purchased vehicles. The lower and upper bounds are called the boundary points of the 95% confidence interval. For the tree-based methods, the ATE estimates are obtained as the average values of the CATE estimates. The same rule applies for the standard deviations. The lower and upper bounds are obtained as the 2.5% and 97.5% quantiles of the corresponding estimates.

provide close results. This is a reasonable outcome as it is a randomised experiment with no RIC involved. However, it is surprising that opposite causal effects are observed for the female and the male participants. The female participants are more likely to purchase less fuel-intensive (more economical) vehicles after the treatment intervention. In contrast, the treatment seems to result in negative feedback for the male participants. The results of the BCF do not exhibit much variability compared with the other methods, as BCF imposes strong homogeneity assumptions for the causal effects. For the GRF results, there is no distinguishing difference between the results of males and females. Unfortunately, we cannot use OLS as the benchmark for the subgroup constructed by gender as randomisation validity is questionable here[3]. The different patterns exhibited in BART-related methods and the GRF are potentially explained by two reasons. First, GRF selects a random subset of the covariates as candidates to split on. Gender is probably not selected in some trees. Second, GRF adopts a different measure for a potential split to a classic decision tree, which is that the goodness of a split relates to maximisation of heterogeneity across the child nodes. Even though gender is a good predictive variable, GRF does not necessarily choose it to split according to the adopted measure.

Figure 5.5 depicts the relationship between the estimated CATE from the BART-EXT method and the covariates: age, miles (in thousands) driven per year, log of annual income, and race (Caucasian or non-Caucasian). The top left panel of Figure 5.4 shows that the CATE estimates are clustered in two groups based on the variable gender, with the estimated causal effect for females being opposite to that of males. The left panel of Figure 5.5 shows that the treatment effect seems to have a positive relationship with age and the log of annual income. Most of the 95% credible intervals include zero (results not shown here), indicating that the positive relationship may not be statistically significant based on the data. Miles driven per year appears to be unrelated to the estimated CATE (the top right panel in Figure 5.5). The treatment intervention tends to have positive impacts on the non-Caucasian participants, as the estimated CATE is less than that of the Caucasian participants (the bottom right plot in Figure 5.5). Compared with the classic approach

[3] A Pearson's Chi-squared test between gender and treatment indicator yields a p-value of 0.07, which is marginally significant.

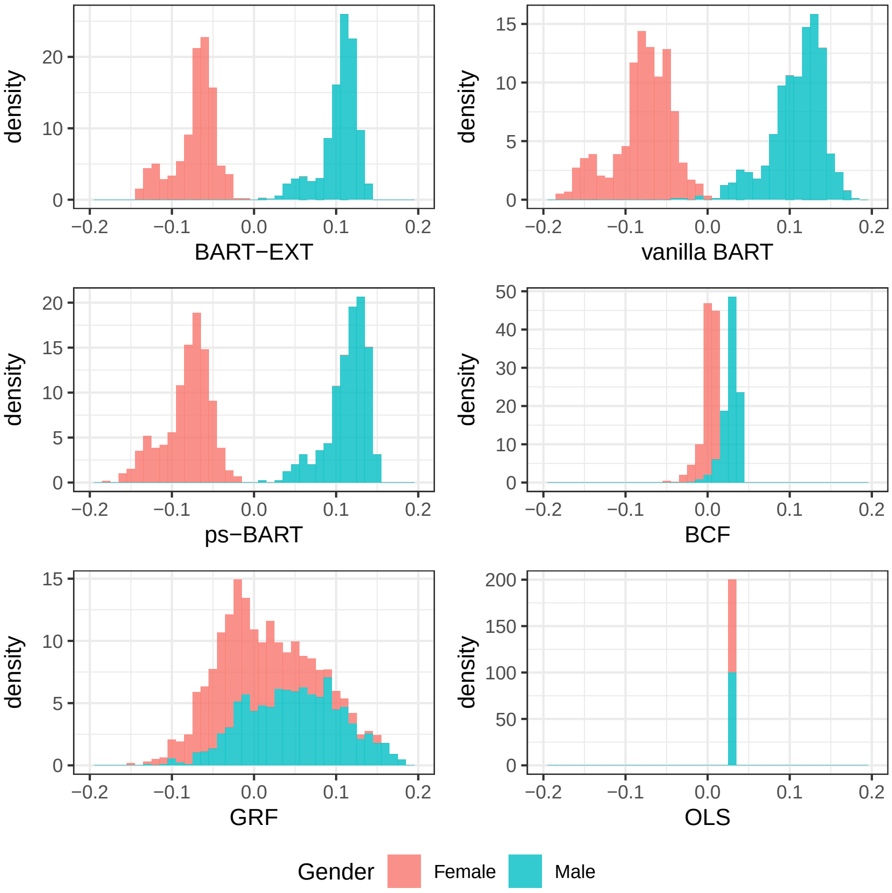


Figure 5.4: Histograms for the CATE estimates on fuel intensity under the different models.

such as OLS, the CATE estimates cast more insights on the treatment intervention on individuals with different characteristics.

5.6 Discussion

This chapter proposes an extension to the original BART model, BART-EXT, for estimating heterogeneous treatment effects with binary treatments and continuous outcomes. The original BART model can bias causal effects in the presence of strong confounding, as it has a regularisation prior to prevent overfitting. Inspired by the GRF (Wager and Athey, 2018) and the BCF (Hahn et al., 2020) approaches, we propose a solution to the

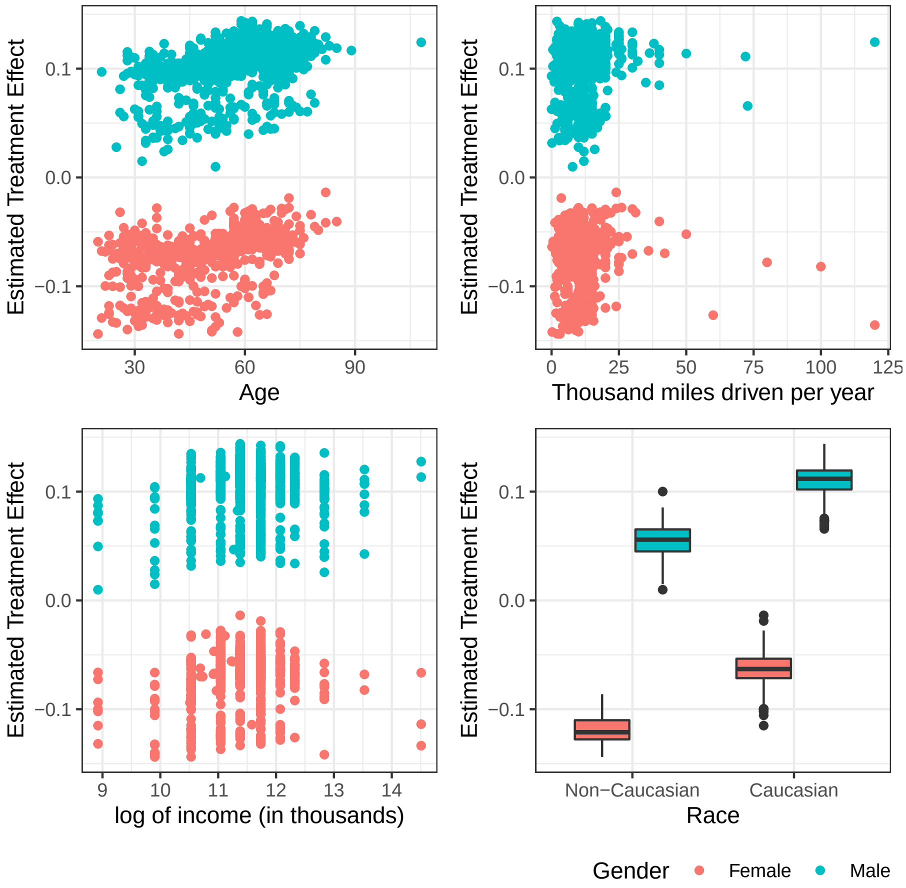


Figure 5.5: Plots for the estimated CATE of the BART-EXT method on the fuel economy in relation to different covariates, including age, thousands miles driven per year (1 unit = 1,000 miles), nature log of annual income (1 unit = 1,000 dollars) and race of being Caucasian or non-Caucasian.

problem by separating the response surface into baseline and treatment parts. For trees capturing the treatment effects, we use identically structured trees for the treated and untreated subjects so that the treatment indicator is only considered in the terminal nodes of each tree. Section 5.4.1 demonstrates the advantage of BART-EXT in a toy example with strong confounding. Section 5.4.2 analyses the data set from the 2016 ACIC data challenge (Dorie et al., 2019), which is a large collection of synthetic data sets generated by different mechanisms. BART-EXT performs well compared to all the tree-related methods considered in this chapter, including the original BART, ps-BART, the BCF and the GRF in respect to different metrics. For the experimental study in Section 5.5, the performance of BART-EXT is close to the benchmark (OLS). The OLS method estimates the ATE only, whereas BART-EXT estimates heterogeneous effects for each unit, which reveals more information about the magnitude of the causal effects with regard to various covariates.

The BART-EXT method fixes hyperparameters before the MCMC begins; for example, the number of trees for the baseline (m_0) and the treatment (m_1) effects. As some covariates are strong predictors of treatment status, the baseline tree might suffer from confounding to some extent. How to determine the tree type stochastically is worth future investigation. With respect to computing time, all the examples in this chapter are implemented using the `R` package `Rcpp`, making the reported computing times comparable to the other discussed methods. One direction of future work is to parallelise computation of the BART-EXT method using the proposal suggested in Pratola et al. (2014). Another future improvement is to develop a more efficient algorithm to implement the proposed model. This chapter adopts the iterative Bayesian backfitting MCMC algorithm proposed by Chipman et al. (2010), where the proposal of a new tree only considers the evolution on a single node at each iteration. As a consequence, the tree structure may stabilise over many iterations, which impedes the algorithm's ability to explore tree structures efficiently. He et al. (2019) propose a grow-from-root strategy, XBART, to accelerate posterior simulation. We can also consider the approach of Lakshminarayanan et al. (2013), which uses a sequential Monte Carlo method for the Bayesian decision tree.

Chapter 6

Conclusion and future directions

This thesis investigates three different problems in Bayesian statistics. The conclusion of each chapter is provided first and followed by a discussion on future directions.

Chapter 3 focuses on SDA, which is a relatively new field in statistics. In SDA, the volume and complexity of the original data is reduced by aggregating the data into symbols of interest. Our proposed approach provides an improved solution to identify information in the original data compared with that of Beranger et al. (2018). The symbols are constructed from a min-max interval or a quantile-based interval, and are modelled by the so-called symbolic likelihood function. Such functions are usually intractable and hence we utilise the signed block PMMH algorithm (Quiroz et al., 2021) to deal with the intractability and the negative likelihood estimates. An exact method is proposed to estimate the likelihood function, which involves path sampling (Gelman and Meng, 1998) and the Poisson estimator (Papaspiliopoulos, 2011). As the method requires large computational resources, we also propose an approximate method to speed up the algorithm, based on a Taylor expansion and the bias-corrected estimator (Ceperley and Dewing, 1999; Quiroz et al., 2019). The approximate method is applied to a factor model and a linear regression involving heteroscedasticity. Compared to the results on the full data, our method requires substantially less running time with a tolerable difference in the parameter estimates.

Chapter 4 investigates the doubly intractable problem, where the likelihood function involves an intractable normalising constant. To overcome such intractability, we adopt a similar algorithm (the signed block PMMH) to Chapter 3, but with a few differences. First, an auxiliary variable is introduced, resulting in a likelihood function on an augmented space being estimated on the exponential scale. Second, the block-Poisson estimator (Quiroz et al., 2021) is used instead of the Poisson estimator (Papaspiliopoulos, 2011). The estimator has a tractable form so that analytical results for the likelihood estimator can be derived under simplifying assumptions, which are used in establishing guidelines for hyperparameter tuning. The proposed method is applied to the Ising model (Ising, 1925), a constrained Gaussian process, and the Kent distribution (Kent, 1982). The first application is one of the classic problems in the literature, whereas for the remaining two applications, the methods for exact Bayesian inference are rarely investigated, because approximate methods usually provide good results as long as the associated assumptions are satisfied. The proposed method can also be applied to a wide range of problems as its only requirement is to have an unbiased estimator for the normalising constant.

Chapter 5 considers causal inference. This chapter extends the original BART model to estimate heterogeneous treatment effects of an observational study with binary treatments and continuous outcomes. The proposed method, BART-EXT, is inspired by the BCF (Hahn et al., 2020) and the GRF (Wager and Athey, 2018) methods. To control the strength of regularisation of the BART prior, the treatment effect is estimated separately from the baseline effect. The estimated propensity score is included to mitigate the bias caused by the regularisation-induced confounding (Hahn et al., 2018). To make the covariates comparable for treated and untreated observations, the trees modelling the treatment effects are forced to have identical structures for observations from the treatment and the control group, which can be regarded as an adaptive matching process. The benefits are illustrated in an extensive simulation study (Dorie et al., 2019), which is often used in the causal inference literature to compare performance between methods under different data generating processes. We demonstrate the BART-EXT method in a reanalysis of an experimental study to investigate the causal effects of providing fuel intensity information on vehicle purchases (Allcott and Knittel, 2019).

Discussion of our contribution and future directions is given in each technical chapter. Here we present a general discussion. All the topics involve a model with a complex likelihood function, whose evaluation is computationally challenging. The likelihood function in Chapter 3 involves an integral with no analytical solution. Chapter 4 studies a likelihood function which is intractable with an unknown normalising constant. In Chapter 5, the likelihood function is composed of multiple trees with a changing structure through the MCMC iterates. To speed up the MCMC methods, one future direction is to find efficient methods for likelihood evaluations. However, the likelihood estimator is model dependent. For complex models, it is usually difficult to develop efficient unbiased estimators with low variability. An alternative direction is to develop elegant algorithms which ease the difficulties in the direct evaluation of the likelihood function. For example, the exchange algorithm presented in Chapter 2.4 avoids estimating the normalising constant by carefully designing an auxiliary variable and the corresponding transition kernel. An iterative backfitting MCMC algorithm is used for inference on the BART model in Chapter 5, where the likelihood of a single tree is investigated with the other trees fixed. The computing cost is much smaller than evaluating the likelihood function for all the trees.

In future research, it is feasible to incorporate other simulation-based Monte Carlo methods in the algorithms developed in this thesis. For example, sequential Monte Carlo (SMC) algorithms (Andrieu et al., 2003) are a collection of methods, which sequentially sample from a series of distributions to approach the target distribution. Del Moral et al. (2006) show that the output of SMC samplers gives a consistent estimate of the posterior expectation of any function of parameters. The computing cost required by SMC methods is often cheaper than MCMC methods. SMC methods have other advantages such as robustness to multi-modality and the availability of parallel computation. As proposed in McGree et al. (2016), it is possible to implement SMC methods under the PM framework. Linking to the topic of Chapter 3, the combination of SMC and PM methods could reduce the computing cost, which can be achieved by gradually increasing the number in the exponent from a small number to the number in the likelihood function using AIS (Neal, 2001). The SMC method also applies for the doubly intractable problems discussed in Chapter 4 provided that it is practical to sample observations from the likelihood function exactly without knowing the normalising constant. This approach is achievable for the Ising model (Ising, 1925) and some graphical models (Naesseth et al., 2014). However, incorporating SMC methods in the algorithm developed in Chapter 4 is not that straightforward and needs further research. For the BART model, Lakshminarayanan et al. (2013) develop a generic particle filtering method for SMC, which explores the tree structure more efficiently than the MCMC algorithm. The method is proposed for vanilla BART and it can be reformulated in our novel BART model proposed in Chapter 5.

In addition to efficient algorithms exploiting simulation-based methods, it may be promising to use generic techniques such as divide-and-conquer approaches (Bardenet et al., 2017; Neiswanger et al., 2013; Scott et al., 2016) and subsampling-based algorithms (Quiroz et al., 2018). Both techniques attempt to accelerate of simulation-based methods using various approaches. The former divides the sample data into mini batches and combines the posterior distribution of each batch cleverly to approximate the posterior distribution of the whole data set. The latter uses a small portion of observations at each MCMC iteration to obtain an unbiased likelihood function estimate based on the sample. With regards to the applications in this thesis, a direct implementation of the techniques above is impractical, especially for the SDA and doubly intractable problems where the independence assumption does not hold for symbolic objects and correlated observations. For the BART model, it is also difficult to use subsampling directly under the backfitting setting. The divide-and-conquer approach, on the other hand, may lead to an inefficient algorithm as the tree structures for each mini-batch are likely to be totally different. Further work is needed to incorporate generic techniques into the topics in this thesis.

Appendix A

Supplementary materials for Chapter 3

A.1 Details on the path sampler

Let $0 \leq t \leq 1$, and denote $h_t(\mathbf{x}; \boldsymbol{\theta}) = g_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\theta})^t, \mathbf{x} \in S$, $S \subseteq \mathbb{R}^d$. The following equation holds

$$\log \int_S g_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{x} = \int_0^1 E_{q_t(\mathbf{z}; \boldsymbol{\theta})} \left[\frac{d}{dt} \log h_t(\mathbf{z}; \boldsymbol{\theta}) \right] dt + \log \int_S 1 dz.$$

Proof. It is straightforward to show that $h_0(\mathbf{x}; \boldsymbol{\theta}) = 1$ and $h_1(\mathbf{x}; \boldsymbol{\theta}) = g_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\theta})$.

Let $\phi_t(\boldsymbol{\theta}) = \int_S h_t(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{x}$, then $\log(\phi_1(\boldsymbol{\theta})) = \log\left(\int_S g_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{x}\right)$, which is our target and

$\phi_0(\boldsymbol{\theta}) = \int_S 1 d\mathbf{x} =$ volume of S.

$$\begin{aligned}
\frac{d}{dt}\log\phi_t(\boldsymbol{\theta}) &= \frac{1}{\phi_t(\boldsymbol{\theta})}\frac{d}{dt}\phi_t(\boldsymbol{\theta}) = \frac{1}{\phi_t(\boldsymbol{\theta})}\frac{d}{dt}\left(\int_S h_t(\mathbf{x};\boldsymbol{\theta})d\mathbf{x}\right) \\
&= \frac{1}{\phi_t(\boldsymbol{\theta})}\int_S \frac{d}{dt}h_t(\mathbf{x};\boldsymbol{\theta})d\mathbf{x} \\
&= \frac{1}{\phi_t(\boldsymbol{\theta})}\int_S h_t(\mathbf{z};\boldsymbol{\theta})\frac{d}{dt}\log h_t(\mathbf{x};\boldsymbol{\theta})d\mathbf{x} \\
&= \int_S \frac{h_t(\mathbf{x};\boldsymbol{\theta})}{\phi_t(\boldsymbol{\theta})}\frac{d}{dt}\log h_t(\mathbf{x};\boldsymbol{\theta})d\mathbf{x} \\
&= \int_S q_t(\mathbf{x};\boldsymbol{\theta})\frac{d}{dt}\log h_t(\mathbf{x};\boldsymbol{\theta})d\mathbf{x} \quad \text{where } q_t(\mathbf{x};\boldsymbol{\theta}) = \frac{h_t(\mathbf{x};\boldsymbol{\theta})}{\phi_t(\boldsymbol{\theta})} \\
&= E_{q_t(\mathbf{x};\boldsymbol{\theta})}\left[\frac{d}{dt}\log h_t(\mathbf{x};\boldsymbol{\theta})\right].
\end{aligned}$$

Taking the integral from 0 to 1, we have

$$\begin{aligned}
[\log\phi_t(\boldsymbol{\theta})]_0^1 &= \log\phi_1(\boldsymbol{\theta}) - \log\phi_0(\boldsymbol{\theta}) \\
&= \int_0^1 E_{q_t(\mathbf{x};\boldsymbol{\theta})}\left[\frac{d}{dt}\log h_t(\mathbf{x};\boldsymbol{\theta})\right] dt.
\end{aligned}$$

□

A.2 Some properties of the Poisson estimator

The proof of the Poisson estimator closely follows that in Quiroz et al. (2019), which focuses on the block-Poisson estimator. The only difference is that the Poisson estimator is a special version of the block-Poisson estimator with only one block. The appendix of Quiroz et al. (2019) proves some properties of the block-Poisson estimator.

Here, we rewrite the proof in the SDA context.

Recall that the Poisson estimator is

$$\widehat{\exp}(A) = \exp(a+\lambda)\prod_{h=1}^{\chi}\frac{(\widehat{A}^{(h)} - a)}{\lambda}.$$

We show the following properties of the Poisson estimator,

1. $E(\widehat{\exp}(A)) = \exp(A)$.

2. $\mathrm{Var}(\widehat{\exp}(A)) = \exp\left(\frac{(A-a)^2}{\lambda} + \lambda + 2a + \frac{\widehat{\sigma}_A^2}{\lambda}\right) - \exp(2A)$.

3. The optimal value of a which minimises the variance of $\widehat{\exp}(A)$ is $a_{opt} = A - \lambda$.

4. The optimal value for a is unavailable as A is unknown. We can choose $a = A^{(h)} - \lambda$, where h is drawn from $\{1, \dots, \chi\}$ randomly, which still gives an unbiased estimator for $\exp(A)$.

We use the following results for the Poisson distribution in the proof. If $\chi \sim \text{Poisson}(\lambda)$ and $A < \infty$,

1. $E_\chi(A^\chi) = \exp[(A-1)\lambda]$.

2. $\mathrm{Var}_\chi(A^\chi) = \exp(-\lambda)[\exp(A^2\lambda) - \exp(2A\lambda - \lambda)]$.

Proof. Property 1

$$\begin{aligned}
E(\widehat{\exp}(A)) &= E_\chi\left[E_{\widehat{A}|\chi}(\exp(a+\lambda)\prod_{h=1}^{\chi}\frac{\widehat{A}^{(h)} - a}{\lambda}\right] \\
&= \exp(a+\lambda)E_\chi\left[\left(\frac{A-a}{\lambda}\right)^\chi\right] \\
&= \exp(a+\lambda) + \exp(\frac{A-a}{\lambda}\lambda - \lambda) \quad \text{(use result 1)} \\
&= \exp(a+\lambda)\exp(A - a - \lambda) \\
&= \exp(A).
\end{aligned}$$

□

Next, to derive the optimal value for the lower bound a, we derive the variance of $\widehat{exp}(A)$.

Proof. Property 2

$$\begin{aligned}
\mathrm{Var}(\widehat{\exp}(A)) &= \mathrm{Var}\left(\exp(a+\lambda)\prod_{h=1}^{\chi}\frac{\widehat{A}^{(h)}-a}{\lambda}\right)\\
&= \exp(2a+2\lambda)\mathrm{Var}\left(\prod_{h=1}^{\chi}\frac{\widehat{A}^{(h)}-a}{\lambda}\right)\\
&= \exp(2a+2\lambda)\left(\mathrm{Var}_{\chi}E_{\widehat{A}|\chi}\prod_{h=1}^{\chi}\frac{\widehat{A}^{(h)}-a}{\lambda}+E_{\chi}\mathrm{Var}_{\widehat{A}|\chi}\prod_{h=1}^{\chi}\frac{\widehat{A}^{(h)}-a}{\lambda}\right)\\
&= \exp(2a+2\lambda)(C+D),\\
\text{with } C &= \mathrm{Var}_{\chi}E_{\widehat{A}|\chi}\prod_{h=1}^{\chi}\frac{\widehat{A}^{(h)}-a}{\lambda}\\
&= \exp(-\lambda)\left[\exp\left(\left((\frac{A-a}{\lambda})^2\lambda\right)-\exp\left(2\frac{A-a}{\lambda}\lambda-\lambda\right)\right)\right]\\
&= \exp(-\lambda)\left[\exp\left(\frac{(A-a)^2}{\lambda}\right)-\exp(2A-2a-\lambda)\right],\\
\text{and } D &= E_{\chi}\mathrm{Var}_{\widehat{A}|\chi}\left[\prod_{h=1}^{\chi}\left(\frac{\widehat{A}^{(h)}-a}{\lambda}\right)\right].
\end{aligned}$$

To derive term for D, first we compute the conditional variance as

$$\begin{aligned}
\mathrm{Var}_{\widehat{A}|\chi}\left[\prod_{h=1}^{\chi}\left(\frac{\widehat{A}^{(h)}-a}{\lambda}\right)\right] &= \prod_{h=1}^{\chi}\left[\mathrm{Var}\left(\frac{\widehat{A}^{(h)}-a}{\lambda}\right)+\left(\frac{\widehat{A}^{(h)}-a}{\lambda}\right)^2\right]-\prod_{h=1}^{\chi}\left(\frac{\widehat{A}^{(h)}-a}{\lambda}\right)^2\\
&= \left[\frac{\widehat{\sigma}_A^2}{\lambda^2}+\left(\frac{\widehat{A}-a}{\lambda}\right)^2\right]^{\chi}-\left[\left(\frac{\widehat{A}-a}{\lambda}\right)^2\right]^{\chi}.
\end{aligned}$$

Plug the term above into D, we have

$$\begin{aligned}
D &= E\left[\left(\frac{\widehat{\sigma}_A^2}{\lambda^2}+\left(\frac{A-a}{\lambda}\right)^2\right)^{\chi}\right]-E\left[\left(\frac{A-a}{\lambda}\right)^{2\chi}\right]\\
&= \exp\left[\left(\frac{\widehat{\sigma}_A^2}{\lambda^2}+\left(\frac{A-a}{\lambda}\right)^2-1\right)\lambda\right]-\exp\left[\left(\left(\frac{A-a}{\lambda}\right)^2-1\right)\lambda\right]\\
&= \exp\left[\frac{(A-a)^2}{\lambda}-\lambda\right]\left[\exp\left(\frac{\widehat{\sigma}_A^2}{\lambda}\right)-1\right].
\end{aligned}$$

Rearrange terms to derive $\mathrm{Var}(\widehat{\exp}(A))$,

$$\begin{aligned}
&\mathrm{Var}(\widehat{\exp}(A)) = \exp(2a+2\lambda)\cdot\\
&\left\{\exp(-\lambda)\left[\exp\left(\frac{(A-a)^2}{\lambda}\right)-\exp(2A-2a-\lambda)\right]-\exp\left(\frac{(A-a)^2}{\lambda}-\lambda\right)\left[\exp\left(\frac{\widehat{\sigma}_A^2}{\lambda}\right)-1\right]\right\}\\
&= \exp\left(\frac{(A-a)^2+\widehat{\sigma}_A^2}{\lambda}+\lambda+2a\right)-\exp(2A).
\end{aligned}$$

□

Proof. Property 3

To get the optimal value for a, take the first derivative with regards to a,

$$\frac{\partial}{\partial a} \log\left(\mathrm{Var}(\widehat{\exp}(A)\right) = -2\frac{A-a}{\lambda} + 2 = 0,$$

which gives $a = A - \lambda$. The second derivative is $\dfrac{2}{\lambda^2} > 0$, confirming the minimum is achieved. □

Even though the Poisson estimator is an unbiased estimator, regardless of the value of a, it is ideal to set $a = A - \lambda$, which gives a closer result to $\exp(A)$ even if $\chi = 0$, which is likely to happen for a small λ. However, we do not know the true value of A, otherwise it is pointless to use the Poisson estimator. We estimate the value of a as

$$a = f(\widehat{A}_1, ...\widehat{A}_\chi) - \lambda,$$

where $f(\cdot)$ produces a random draw from the inputs.

Proof. Property 4

Assume $j \sim \mathrm{Uniform}(1, \ldots, \chi)$, and let $a = \sum_{h=1}^{\chi} w_h \widehat{A}^{(h)} - \lambda$ with $w_j = 1$ and $w_i = 0$ for $i \neq j$ (we have a random draw from all the candidates).

To show the unbiasedness of the Poisson estimator, starting with the intermediate result of the proof for property 1:

$$\begin{aligned} & E_{\widehat{A}|\chi}\left(\exp(a+\lambda)\prod_{h=1}^{\chi}\frac{\widehat{A}^{(h)}-a}{\lambda}\right) \\ & = E_{\widehat{A}|\chi}\exp(\widehat{A}^{j})\prod_{h=1,h\neq j}^{\chi}\left(\frac{\widehat{A}^{(h)}-\widehat{A}^{(j)}+\lambda}{\lambda}\right) \\ & = E_{\widehat{A}^{(j)}|\chi}\exp(\widehat{A}^{(j)})\left(\frac{A-\widehat{A}^{(j)}+\lambda}{\lambda}\right)^{\chi-1}. \end{aligned}$$

The last step uses the iterative expectation on $\widehat{A_h}|\widehat{A_j}, h \neq j$.

Exchange the order of χ and $\widehat{A}$,

$$\begin{aligned}
E(\widehat{\exp}(A)) &= E_{\widehat{A}}\left[\exp(\widehat{A}^{(j)})\left(\frac{A-\widehat{A}^{(j)}+\lambda}{\lambda}\right)^{-1} E_{\chi|\widehat{A}}\left(\frac{A-\widehat{A}^{(j)}+\lambda}{\lambda}\right)^{\chi}\right] \\
&= E_{\widehat{A}}\left[\exp(\widehat{A}^{(j)})\left(\frac{A-\widehat{A}^{(j)}+\lambda}{\lambda}\right)^{-1} \exp(A-\widehat{A}^{(j)})\right] \\
&= \exp(A)E_{\widehat{A}}\left(\frac{A-\widehat{A}^{(j)}+\lambda}{\lambda}\right)^{-1}
\end{aligned}$$

Assume that $\widehat{A} \xrightarrow{P} A$, then $\left(\frac{A-\widehat{A}+\lambda}{\lambda}\right)^{-1} \xrightarrow{P} 1$, where $\xrightarrow{P}$ refers to convergence in probability. □

Appendix B

Supplementary material for Chapter 4

B.1 Properties of the block-Poisson estimator

Proof of Lemma 1[1]:

Proof. Recall that the block-Poisson estimator is expressed as $\widehat{L}_B(\boldsymbol{\theta}) = \prod_{l=1}^{\lambda} \exp(\xi_l(\boldsymbol{\theta}))$ with

$$\exp(\xi_l(\boldsymbol{\theta})) = \exp(a/\lambda + m) \prod_{h=1}^{\chi_l} \frac{\widehat{B}^{(h,l)}(\boldsymbol{\theta}) - a}{m\lambda},$$

where λ is the number of blocks with $\chi_l \sim \text{Pois}(m)$ and a is an arbitrary constant. For notational convenience, dependence on $\boldsymbol{\theta}$ is omitted for $\widehat{L}_B$, $\widehat{B}$ and ξ.

The following proofs closely follow the proofs in Quiroz et al. (2021, Section S8). In the paper, they assume $m = 1$, whereas here m can be any non-negative integer. The two properties below are useful for the proof. Suppose that $X \sim \text{Pois}(m)$ and $A < \infty$. Then,

[1] The proof of Lemma 1 is almost the same as in Appendix A.1, with the only difference that the number of blocks m is also included.

(i) $E_X(A^X) = \exp((A-1)m)$.

(ii) $\mathrm{Var}_X(A^X) = \exp(-m)[\exp(A^2m) - \exp(2Am - m)]$.

Proof of unbiasedness

$$\begin{aligned}
E(\exp(\xi_l)) &= \exp(a/\lambda + m)E\left[\prod_{h=1}^{\chi_l} \frac{\widehat{B}^{(h,l)} - a}{m\lambda}\right] \\
&= \exp(a/\lambda + m)E_\chi E_{\widehat{B}|\chi}\left[\prod_{h=1}^{\chi_l} \frac{\widehat{B}^{(h,l)} - a}{m\lambda}\right] \\
&= \exp(a/\lambda + m)E_\chi \left[\frac{B - a}{m\lambda}\right]^\chi \\
&= \exp(a/\lambda + m)\exp((B-a)/\lambda - m) \\
&= \exp(B/\lambda).
\end{aligned}$$

As $\xi_1, \ldots, \xi_\lambda$ are independent, $E(\widehat{L}_B) = \exp(B)$.

In the implementation, we use $a = \widehat{B} - m\lambda$, where $\widehat{B}$ is an estimate of B and it is independent of $\widehat{B}^{(h,l)}$. Such choice of a preserves the unbiasedness of the block Poisson estimator with the explanation given below.

Treating a as a random variable, the expectation of block-Poisson estimator can be expressed as

$$E(\widehat{L}_B(\boldsymbol{\theta})) = E_a E_{\xi_1,\ldots,\xi_\lambda|a} \prod_{l=1}^{\lambda} \exp(\xi_l(\boldsymbol{\theta})) = E_a \left(\prod_{l=1}^{\lambda} E_{\xi_l|a} \exp(\xi_l(\boldsymbol{\theta}))\right).$$

The conditional expectation of $E_{\xi_l|a}\exp(\xi_l)$ is (omitting the dependence on $\boldsymbol{\theta}$)

$$E_{\xi_l|a}\exp(\xi_l) = E_{\xi_1|a}\exp(a/\lambda + m)\prod_{h=1}^{\chi_l} \frac{\widehat{B}^{(h,l)} - a}{m\lambda}.$$

The conditional expectation is the same as the derived $E(\exp(\xi_l))$, which is independent of a as it is cancelled out in the process. Hence, treating a as a random variable still guarantees the unbiasedness of the block-Poisson estimator.

Derivation of the variance

From the definition of $\widehat{L}_B$,

$$\mathrm{Var}(\widehat{L}_B) = \mathrm{Var}\left(\prod_{l=1}^{\lambda} \exp(\xi_l)\right).$$

For a collection of independent random variables $\exp(\xi_1), \ldots, \exp(\xi_\lambda)$,

$$\mathrm{Var}\left(\prod_{l=1}^{\lambda} \exp(\xi_l)\right) = \prod_{l=1}^{\lambda} \left(\mathrm{Var}(\exp(\xi_l)) + E(\exp(\xi_l))^2\right) - \prod_{l=1}^{\lambda} E(\exp(\xi_l))^2$$

with

$$\mathrm{Var}(\exp(\xi_l)) = \exp(a/\lambda + m) \left[E_\chi \mathrm{Var}_{\widehat{B}|\chi} \left(\prod_{h=1}^{\chi} \frac{\widehat{B}^{(h,l)} - a}{m\lambda}\right) + \mathrm{Var}_\chi E_{\widehat{B}|\chi} \left(\prod_{h=1}^{\chi} \frac{\widehat{B}^{(h,l)} - a}{m\lambda}\right)\right].$$

For the first term in the brackets, making the use of independence of $\widehat{B}^{(h,l)}, h = 1, \ldots, \chi_l$, $\mathrm{Var}_{\widehat{B}|\chi}\left(\prod_{h=1}^{\chi} \frac{\widehat{B}^{(h,l)} - a}{m\lambda}\right)$ can be simplified as

$$\begin{aligned}\mathrm{Var}_{\widehat{B}|\chi}\left(\prod_{h=1}^{\chi} \frac{\widehat{B}^{(h,l)} - a}{m\lambda}\right) &= \prod_{h=1}^{\chi}\left[\mathrm{Var}\left(\frac{\widehat{B}^{(h,l)} - a}{m\lambda}\right) + E\left(\frac{\widehat{B}^{(h,l)} - a}{m\lambda}\right)^2\right] - \prod_{h=1}^{\chi} E\left(\frac{\widehat{B}^{(h,l)} - a}{m\lambda}\right)^2 \\ &= \prod_{h=1}^{\chi}\left(\frac{\sigma_B^2 + (B-a)^2}{(m\lambda)^2}\right) - \left(\frac{B-a}{m\lambda}\right)^{2\chi}.\end{aligned}$$

Taking the expectation with regard to χ and use property 1 twice for the two terms, we have

$$\begin{aligned}E_\chi \mathrm{Var}_{\widehat{B}|\chi}\left(\prod_{h=1}^{\chi} \frac{\widehat{B}^{(h,l)} - a}{m\lambda}\right) &= \exp\left[\left(\frac{\sigma_B^2 + (B-a)^2}{(m\lambda)^2} - 1\right)m\right] - \exp\left[\left(\frac{(B-a)^2}{(m\lambda)^2} - 1\right)m\right] \\ &= \exp\left[\left(\frac{(B-a)^2}{(m\lambda)^2} - 1\right)m\right]\left[\exp\left(\frac{\sigma_B^2}{m\lambda^2}\right) - 1\right].\end{aligned}$$

The second term can be derived similarly,

$$\begin{aligned}\mathrm{Var}_\chi E_{\widehat{B}|\chi}\left(\prod_{h=1}^{\chi} \frac{\widehat{B}^{(h,l)} - a}{m\lambda}\right) &= \mathrm{Var}_\chi\left(\frac{B-a}{m\lambda}\right)^\chi \\ &= \exp\left(-m + \frac{(B-a)^2}{m\lambda^2}\right) - \exp\left(2(B-a)/\lambda - 2m\right).\end{aligned}$$

Combining the two terms, we have

$$\mathrm{Var}(\exp(\xi_l)) = \exp\left[\frac{(B-a)^2 + \sigma_B^2}{m\lambda^2} - m\right] - \exp(2(B-a)/\lambda - 2m).$$

Deriving $E(\exp(\xi))^2$ is straight forward,

$$\begin{aligned} E(\exp(\xi_l))^2 &= \left[E_\chi E_{B|\chi} \prod_{h=1}^{\chi} \left(\frac{\widehat{B}^{(h,l)} - a}{m\lambda} \right) \right]^2 \\ &= \exp(2(B-a)/\lambda - 2m). \end{aligned}$$

Combining all the terms, after some algebra, the variance of the block-Poisson estimator is

$$\mathrm{Var}(\widehat{L}_B) = \exp\left[\frac{(B-a)^2 + \sigma_B^2}{m\lambda} + 2a + m\lambda \right] - \exp(2B).$$

Choice of the constant a

The optimal value a minimising $\mathrm{Var}(\widehat{L}_B)$ is $a = B - m\lambda$. This is obtained by solving the equation $\partial \widehat{L}_B / \partial a$ equal to 0. □

Proof of Lemma 2:

Proof. The proof is exactly the same as Lemma 3 in Quiroz et al. (2021). □

Proof of Lemma 3:

Proof. The variance of the log of the likelihood estimator is

$$\begin{aligned} \mathrm{Var}(\log|\widehat{L}_B|) &= \mathrm{Var}\bigg(\sum_{l=1}^{\lambda} \sum_{h=1}^{\chi_l} \log\left| \frac{\widehat{B}^{(h,l)} - a}{m\lambda} \right| \bigg) \\ &= E_{\chi_{1,\dots,\lambda}} V_{B|\chi_{1,\dots\lambda}} \log\left| \frac{\widehat{B}^{(h,l)} - a}{m\lambda} \right| + \mathrm{Var}_{\chi_{1,\dots,\lambda}} E_{B|\chi_{1,\dots\lambda}} \log\left| \frac{\widehat{B}^{(h,l)} - a}{m\lambda} \right|. \end{aligned}$$

Suppose $\widehat{B}^{(h,l)} \sim N(B, \sigma_B^2)$ and $a = B - m\lambda$, then

$$\begin{aligned} \log\left| \frac{\widehat{B}^{(h,l)} - a}{m\lambda} \right| &= \log\left| \frac{\sigma_B Z}{m\lambda} + 1 \right| \\ &= \log(\sigma_B/(m\lambda)) + \log|Z + m\lambda/\sigma_B| \\ &= \log(\sigma_B/(m\lambda)) + \frac{1}{2}\log((Z + m\lambda/\sigma_B)^2) \\ &= \log(\sigma_B/(m\lambda)) + \frac{1}{2}\log W^{(h,l)}, \quad W^{(h,l)} \sim \chi^2(1, (m\lambda/\sigma_B)^2), \end{aligned}$$

where $\chi^2(k, \lambda)$ denotes the non-central χ^2 distribution with k degrees of freedom and non-centrality parameter λ. Lemma S12 in Quiroz et al. (2021) provides the moments of $\log W$.

Let η_B and ν_B^2 be the expectation and the variance of $\log\left|\frac{\widehat{B}^{(h,l)}-a}{m\lambda}\right|$ respectively. We have

$$\begin{aligned}\eta_B &= E\bigg(\log\bigg|\frac{\widehat{B}^{(h,l)}-a}{m\lambda}\bigg|\bigg) = \log(\sigma_B/(m\lambda)) + \frac{1}{2}log(2 + E_J[\psi^{(0)}(0.5+J)],\\ \nu_B^2 &= \mathrm{Var}\bigg(\log\bigg|\frac{\widehat{B}^{(h,l)}-a}{m\lambda}\bigg|\bigg) = \frac{1}{4}\left[E_J[\psi^{(1)}(0.5+J)] + \mathrm{Var}_J[\psi^{(0)}(0.5+J)]\right],\end{aligned}$$

where $J \sim \mathrm{Pois}((m\lambda)^2/2\sigma_B^2)$ and $\psi^{(q)}$ is the polygamma function of order q.

Finally,

$$\begin{aligned}\mathrm{Var}(\log|\widehat{L}_B|) &= E_{\chi_{1,\ldots,\lambda}}\bigg(\sum_{l=1}^{\lambda}\chi_l\bigg)\nu_B^2 + \mathrm{Var}_{\chi_{1,\ldots,\lambda}}\bigg(\sum_{l=1}^{\lambda}\chi_l\bigg)\eta_B\\ &= m\lambda(\nu_B^2 + \eta_B^2).\end{aligned}$$

Furthermore, $\mathrm{Var}(\log|\widehat{L}_B|) < \infty$. Lemma 7 in Quiroz et al. (2021) derives the result. □

B.2 Implementation details of the *signed block PMMH with the BP* algorithm

The implementation details of Algorithm 7 are covered in this section; it covers the construction of the BP estimator and the choice of the soft lower bound. In Section 4.4.3, the variance of $\gamma(\boldsymbol{\theta}) = M\mathrm{Var}(-\nu\widehat{Z_M}(\boldsymbol{\theta}))$ is treated as a known value for hyperparameter tuning. The decomposition of $\gamma(\boldsymbol{\theta})$ is discussed below, which helps to understand the effect of the randomness in ν.

Construction of the BP estimator

To implement the BP estimator, we first fix the hyperparameters λ, m and a. For each of the blocks h, $h = 1, \ldots, \lambda,$, we sample $\chi_h \sim \mathrm{Pois}(m)$. Depending on the value of χ_h, we need to have the same number of $-\nu\widehat{Z}$ estimates. The whole process can be done in parallel. We can draw χ_h for all the possible h values at one time, and the total replications of $-\nu\widehat{Z}$ required are $\sum_{h=1}^{\lambda}\chi_h$. Parallel computation can also be implemented within the individual estimation process for $\widehat{Z}$ locally, where the calculation of M particles is executed simultaneously.

Choosing the lower bound in the BP estimator

The lower bound $a_{opt} = -\nu Z - m\lambda$ for the BP estimator minimises the variance of the likelihood estimator. In the implementation, Z is replaced by its estimate $\widehat{Z}$. Its computation is exactly the same as that of the $\widehat{Z}$s' used in the estimator. We emphasise that it is necessary to estimate $\widehat{Z}$ independently to ensure that the estimator is unbiased.

Decomposition of γ

Recall the decomposition of $\gamma(\boldsymbol{\theta})$ in (4.9),

$$\gamma(\boldsymbol{\theta}) = M\nu^2\text{Var}(\widehat{Z_M}(\boldsymbol{\theta})) = M\frac{\log(u)^2}{Z^2}\frac{\sigma_Z^2}{M} = \log(u)^2\frac{\sigma_Z^2}{Z^2}.$$

The dependence of Z, σ_Z^2 and γ on $\boldsymbol{\theta}$ is omitted for notational simplicity. The equation above shows that γ is determined by a constant $\log(u)^2$ and the ratio between standard deviation of the estimator $\widehat{Z}$ and mean value: σ_Z/Z. The unconditional variance can also be derived by using the law of total variance, giving a similar conclusion as discussed below.

Effect of $\log(u)^2$*:* It is concerning that $\log(u)^2$ is unbounded as u approaches to 0. As Figure B.1 shows, there is less than a 5% chance of $\log(u)^2 > 9$. Instead, as the equation above shows, the introduction of u reduces the variance by a factor of around 2 with 50% probability as $\Pr(\log(u)^2 < 0.48) = 0.5$. Furthermore, $\Pr(\log(u)^2 < 1) = 0.63$, indicating that the variance does not increase with a probability greater than 0.6. The expectation of $\log(u)^2$ is around 2, i.e., on average, the effect of u doubles the variance.

Effect of σ_Z/Z*:* This ratio is called the coefficient of variation. It describes the magnitude of the variation relative to the mean. However, it is difficult to estimate as both σ_Z and Z

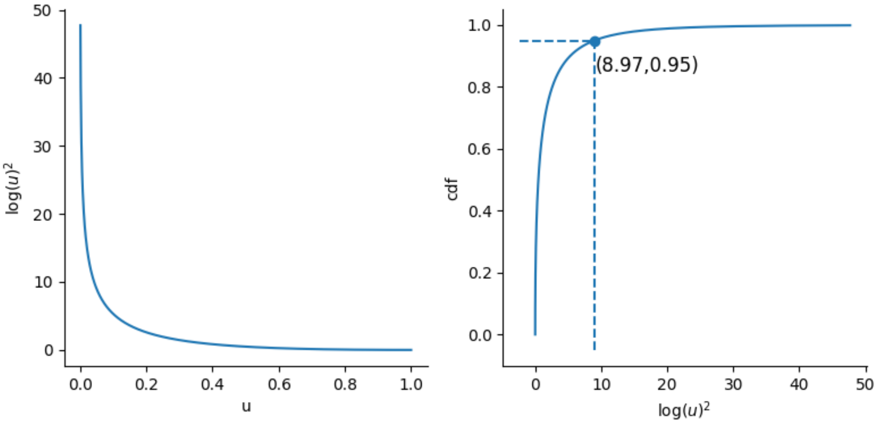


Figure B.1: Left panel: function of $\log(u)^2$ on $u \in (0, 1]$. Right panel: the cdf of $\log(u)^2$ with $u \sim \text{Uniform}(0, 1]$.

are unknown. The term $\frac{\sigma_Z^2}{MZ^2}$ is often estimated by calculating the sample variance of $\widehat{Z}$ over M samples by Monte Carlo integration. The value of σ_Z/Z is implicitly determined by the properties of $\widehat{Z}$, regardless of M.

Conclusion: It is hard to obtain an analytical expression of γ. However, we could estimate it by Monte Carlo integration. There is also uncertainty associated with ν. Setting $\gamma \approx 2\sigma_Z^2/Z^2$ is a conservative choice to account for the effect of $\log(u)^2$.

B.3 Details of the Ising model

B.3.1 An unbiased estimator for the normalising constant

This section supplements the material on AIS sampling in Section 4.5.1. The likelihood function is $p(\mathbf{y}|\theta) = \frac{f(\mathbf{y}|\theta)}{Z(\theta)}$, with $f(\mathbf{y}|\theta) = \exp(\theta S(\mathbf{y}))$.

Consider the following intermediate kernel of the likelihood function

$$f_n(\mathbf{y}|\theta) = f(\mathbf{y}|\theta)^{\beta_n} p(\mathbf{y})^{1-\beta_n},$$

where $0 = \beta_0 < \beta_1 < \cdots < \beta_{n-1} < \beta_n = 1$ and $p(\mathbf{y}) = 0.5^{L \times L}$. In the case of β_0, sampling from the prior density $f_n(\mathbf{y}|\theta)$ of β_0 is straightforward. By gradually increasing β, the samples will be drawn from the likelihood function at the nth step without knowing the normalising constant. The algorithm starts by sampling M particles from $f_0(\mathbf{y}|\theta)$, and proceeds with a certain transition probability to a new configuration for β_i $(i = 1, \ldots, n-1)$ and terminates when $\beta_n = 1$ is reached.

The transition to a new configuration $\mathbf{y}_{i+1,m}$ $(m = 1, \ldots, M)$ from the current configuration $\mathbf{y}_{i,m}$ is completed by the following Gibbs update.

1 Select one random location i, j out of an $L \times L$ grid.

2 Change the corresponding value of $y_{i,j}$ with probability

$$p(y_{i,j} = 1) = \frac{1}{1 + \exp(-\beta_{i+1}\theta \sum y_{neighbour})},$$

where $y_{neighbour}$ refers to the points to the left, right, up and down of $y_{i,j}$.

3 Set the new configuration as $y_{i+1,m}$.

The final weight associated with particle m (omitting θ) is

$$w^{(m)} = \frac{f_1(\mathbf{y}_{0,m})}{f_0(\mathbf{y}_{1,m})} \frac{f_2(\mathbf{y}_{1,m})}{f_1(\mathbf{y}_{1,m})} \cdots \frac{f_n(\mathbf{y}_{n-1,m})}{f_{n-1}(\mathbf{y}_{n-1,m})}.$$

The average of the importance weights $\sum_{m=1}^{M} w^{(m)}/N$ converges to the ratio of $Z_1(\theta)/Z_0(\theta)$, where $Z_i(\theta)$ corresponds the normalising constant of $f_i(y|\theta)$.

The computation of $w^{(m)}$ is in logs to avoid overflow problems. Each particle is independent of the others, which makes parallel computation possible. Our description follows the supplementary code in Park and Haran (2018) which uses OpenMP to implement the parallel computation. The re-evaluation of $S(\mathbf{y})$ from scratch can be computationally costly as it involves $O(L^2)$ operations for each combination of β_i and particle m. We modify the evaluation process by adding or subtracting the local updates of the selected location only. Such changes reduces the complexity to $O(1)$ and consequently decrease the computational time substantially.

B.3.2 The bias-corrected estimator

By introducing the auxiliary variable ν, we transform the problem of unbiasedly estimating the reciprocal $Z(\theta)$ into the problem of unbiasedly estimating $\exp(-\nu Z(\theta))$. Ceperley and Dewing (1999) discus a method for debiasing $\exp(\cdot)$; Quiroz et al. later extend their estimator to subsampling. They call their estimator the approximately bias-corrected likelihood estimator. The core idea of the method is based on the normality assumption of $\widehat{Z}(\theta)$. It is well known that if $x \sim \text{log}-\text{normal}(\mu, \sigma^2)$, then $E(x) = \exp(\mu + 0.5\sigma^2)$. Using this property of the log-normal distribution, the bias-corrected estimator is

$$\exp\left(-\nu\widehat{Z_M}(\theta) - \frac{\text{Var}(-\nu\widehat{Z_i}(\theta))}{2M}\right),$$

where $\widehat{Z_M}(\theta) = \frac{1}{M}\sum_{i=1}^{M}\widehat{Z_i}(\theta)$.

B.3.3 The variability of the normalising constant

The estimate $\widehat{Z}$ and its variability are crucial in hyperparameter tuning for the proposed algorithm. As the Ising model involves one parameter, it is feasible to study the variability of $\widehat{Z}$ by simulation. Figure B.2 shows the estimates of the scaled $\widehat{Z}$ under different θ values, where $\widehat{Z}$ is rescaled by dividing the sample mean of the replications. Each histogram is generated by 1,000 independent replications, each of which uses 100 particles in AIS with 4,000 intermediate transitions equally spaced between 0 and 1. The horizontal axis refers to the scaled $\widehat{Z}(\theta)$. As θ increases, the distribution of the scaled $\widehat{Z}$ is heavily skewed and the normality assumption appears to be invalid for $\theta > 0.4$. Such violation explains the overestimation by the bias-corrected estimator for $\theta = 0.43$ in the example in Section 4.5.1. Examining the range of the horizontal axis, the magnitude also increases sharply with θ. It states that a larger θ is associated with more variability in $\widehat{Z}$. Hence, more particles are required to estimate $\widehat{Z}$ as θ increases.

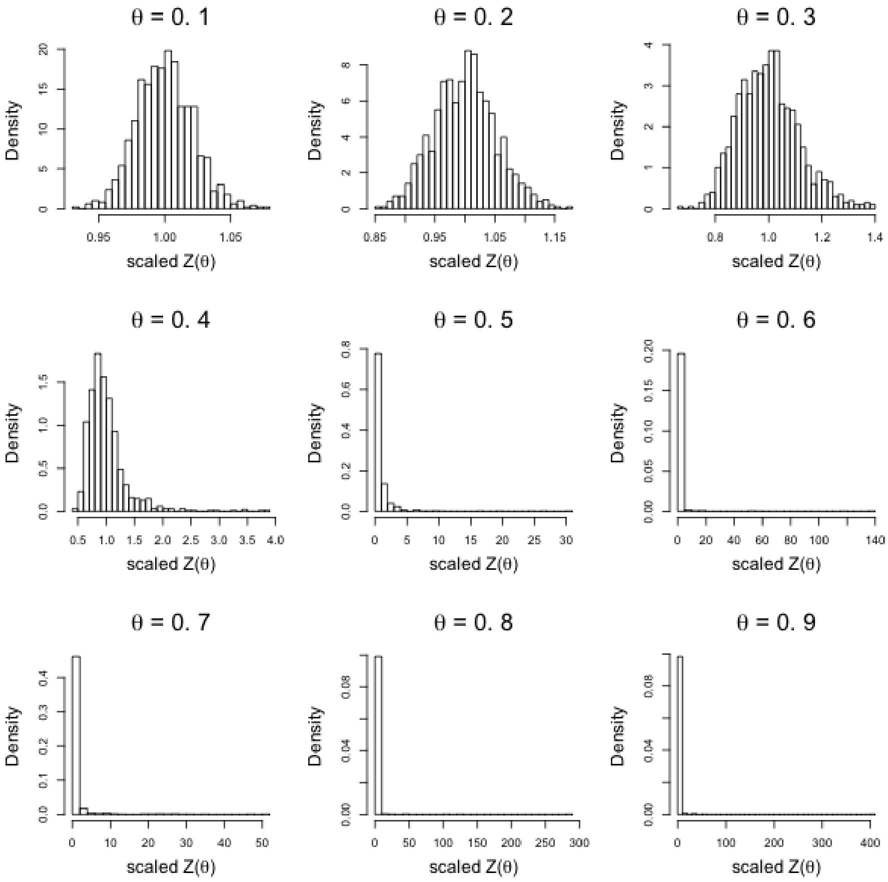


Figure B.2: Histograms of scaled $Z(\theta)$ estimates on a 10×10 2D Ising model.

B.4 Details of the GP example

B.4.1 Derivation of the posterior distribution

Recall the model: $y = g(\mathbf{x}) + \epsilon, \quad \text{with } g \geq 0, \epsilon \sim N(0, \sigma^2)$.

The posterior is:

$$\begin{aligned}
\pi(\alpha, \rho, \sigma^2 | \mathbf{y}) &\propto \pi(\alpha, \rho, \sigma^2) \int_{\mathbf{g} \geq 0} p(\mathbf{y} | \mathbf{g}, \sigma^2) \frac{p(\mathbf{g} | \alpha, \rho)}{Z(\alpha, \rho)} d\mathbf{g} \\
&\propto \frac{\pi(\alpha, \rho, \sigma^2)}{Z(\alpha, \rho)} \int_{\mathbf{g} \geq 0} p(\mathbf{y} | \mathbf{g}, \sigma^2) p(\mathbf{g} | \alpha, \rho) dg \\
&= \frac{\pi(\alpha, \rho, \sigma^2)}{Z(\alpha, \rho)} Z^*(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*) p_y(\mathbf{y} | \alpha, \rho, \sigma^2),
\end{aligned}$$

where $p(\mathbf{g}|\alpha,\rho)$ is a multivariate normal distribution with mean vector 0, covariance matrix $K(\mathbf{x},\mathbf{x}'|\alpha,\rho)$ and $Z(\alpha,\rho)=\int_{\mathbf{g}\geq 0}p(\mathbf{g}|\alpha,\rho)d\mathbf{g}$ is the intractable term. Here we use $p_y(.)$ refers to a unconstrained multivariate normal distribution with mean vector 0 and covariance matrix $\mathbf{K}_{xx}+\sigma^2\mathrm{I}_n$, where $\mathbf{K}_{xx}=K(\mathbf{x},\mathbf{x}'|\alpha,\rho)$. The integration in the second last line is done analytically as it is a convolution of two normal distributions. The posterior for the scalable $\mathcal{GP}$ is done similarly except that the covariance matrix is placed on the inducing points. Sections 4.5.2.2 and 4.5.2.3 provide the results.

The posterior, after introducing the auxiliary variable $\nu\sim\mathrm{Expon}(Z(\alpha,\rho))$, is

$$\pi(\alpha,\rho,\sigma^2,\nu|\mathbf{y})\propto\pi(\alpha,\rho,\sigma^2)\exp(-\nu Z(\alpha,\rho))Z^*(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*)p_y(\mathbf{y}|\alpha,\rho,\sigma^2)$$

In the *signed block PMMH with BP* algorithm, the posterior is estimated as:

$$\widehat{\pi}(\alpha,\rho,\sigma^2,\nu|\mathbf{y})\propto\pi(\alpha,\rho,\sigma^2)|\widehat{\exp}(-\nu Z(\alpha,\rho))|\widehat{Z^*}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*)p_y(\mathbf{y}|\alpha,\rho,\sigma^2),$$

where $\widehat{\exp}(\cdot)$ is the block-Poisson estimator and $\widehat{Z}(\cdot,\cdot)$ and $\widehat{Z^*}(\cdot,\cdot)$ are estimated by the SOV estimator (Genz, 1992).

B.4.2 Prediction based on the constrained GP

Obtaining "raw" predictions follows the same procedure as an ordinary $\mathcal{GP}$ without constraints. The results are established in Williams and Rasmussen (2006) and Snelson and Ghahramani (2006) for scalable $\mathcal{GP}$. We illustrate the posterior prediction under the constraint below.

Prediction at x: Denote $g(\mathbf{x})$ as $\mathbf{g}$ and denote all the hyperparameters (α,ρ,σ) by $\boldsymbol{\theta}$. To get the prediction for $\mathbf{g}$ for the small data case, consider the expectation with respect to the posterior distribution,

$$\begin{aligned}E(\mathbf{g}|\mathbf{y})&=\int_{\mathbf{g}}\mathbf{g}p(\mathbf{g}|\mathbf{y})d\mathbf{g}\\&=\int_{\mathbf{g}}\mathbf{g}\left(\int_{\boldsymbol{\theta}}p(\mathbf{g}|\boldsymbol{\theta},\mathbf{y})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}\right)d\mathbf{g}\\&=\int_{\boldsymbol{\theta}}\int_{\mathbf{g}}\mathbf{g}p(\mathbf{g}|\boldsymbol{\theta},\mathbf{y})p(\boldsymbol{\theta}|\mathbf{y})d\mathbf{g}d\boldsymbol{\theta},\end{aligned}$$

where $p(\mathbf{g}|\boldsymbol{\theta}, \mathbf{y})$ is a truncated multivariate normal distribution with mean $\boldsymbol{\mu}_g^*$, covariance $\boldsymbol{\Sigma}_g^*$, with the lower bound 0. The expressions for $\boldsymbol{\mu}_g^*, \boldsymbol{\Sigma}_g^*$ follow the notation used in Sections 4.5.2.2 and 4.5.2.3. Fortunately, samples of $\mathbf{g}$ are obtained in the process of evaluating $\widehat{Z^*}(\boldsymbol{\theta})$ in the MCMC iterations, so there is no extra computing cost.

For the scalable $\mathcal{GP}$, $\boldsymbol{\mu}_g^*$ and $\boldsymbol{\Sigma}_g^*$ are replaced with or $\boldsymbol{\mu}_{\overline{g}}^*$ and $\boldsymbol{\Sigma}_{\overline{g}}^*$. The samples of $\overline{\mathbf{g}}$ need to be further projected to the location $\mathbf{x}$ by $\overline{g}_m(\mathbf{x}) = \mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\overline{\mathbf{g}}_m$.

Prediction at $\mathbf{x}^*$: Starting with the small sample case first, to predict $g^*(\mathbf{x}^*)$ at a new location $\mathbf{x}^*$, we have:

$$\begin{aligned}
E(\mathbf{g}^*|\mathbf{y}) &= \int \mathbf{g}^* p(\mathbf{g}^*|\mathbf{y}) d\mathbf{g}^* \\
&= \int_{\mathbf{g}^*} \mathbf{g}^* \int_{\mathbf{g}} \int_{\boldsymbol{\theta}} p(\mathbf{g}^*|\mathbf{g}, \mathbf{y}, \boldsymbol{\theta}) p(\mathbf{g}|\boldsymbol{\theta}, \mathbf{y}) p(\boldsymbol{\theta}|\mathbf{y}) d\mathbf{g} d\boldsymbol{\theta} d\mathbf{g}^* \\
&= \int_{\mathbf{g}^*} \int_{\mathbf{g}} \int_{\boldsymbol{\theta}} \mathbf{g}^* p(\mathbf{g}^*|\mathbf{g}, \boldsymbol{\theta}) p(\mathbf{g}|\boldsymbol{\theta}, \mathbf{y}) p(\boldsymbol{\theta}|\mathbf{y}) d\mathbf{g} d\boldsymbol{\theta} d\mathbf{g}^*,
\end{aligned}$$

where $p(\mathbf{g}^*|\mathbf{g}, \boldsymbol{\theta}) = \text{trunc-normal}(\mathbf{g}^*|\boldsymbol{\mu}_{g^*|g}, \boldsymbol{\Sigma}_{g^*|g}; \mathbf{0}, \boldsymbol{\infty})$ with $\boldsymbol{\mu}_{g^*|g} = \mathbf{K}_{x^*x}\mathbf{K}_{xx}^{-1}\mathbf{g}$ and $\boldsymbol{\Sigma}_{g^*|g} = \mathbf{K}_{x^*x} - \mathbf{K}_{x^*x}\mathbf{K}_{xx}^{-1}\mathbf{K}_{xx^*}$. It is expensive to sample a vector $\mathbf{g}^*$ from $p(\mathbf{g}^*|\mathbf{g}, \boldsymbol{\theta})$ as it requires $O(n_{pred}^3)$ evaluations to get the Cholesky decomposition of the matrix $\boldsymbol{\Sigma}_{g^*|g}$. Instead, consider a one-dimensional $\mathbf{g}^*$ so that $p(\mathbf{g}^*|\mathbf{g}, \boldsymbol{\theta})$ is reduced to a one-dimensional truncated normal distribution with a known normalising constant. The analytical forms of quantities such as the mean and median of one-dimensional truncated normal distributions are available. In the simulation study, the median is selected as it is more robust than the mean.

For the scalable $\mathcal{GP}$, $p(\mathbf{g}^*|\mathbf{g}, \boldsymbol{\theta})$ is a truncated normal distribution with mean vector $\boldsymbol{\mu}_{g^*|\overline{g}_m} = \mathbf{K}_{x^*m}\mathbf{K}_{mm}^{-1}\overline{\mathbf{g}}_m$ and the variance $\mathbf{K}_{x^*x^*}(\mathbf{K}_{mm}^{-1} - \mathbf{Q}_{mm}^{-1})\mathbf{K}_{xx^*}$.

B.5 Details of the Kent distribution

Recall the density function of the Kent distribution, $\mathbf{y} = \{y_1, y_2, y_3\}$ with $\sum_{i=1}^{3} y_i^2 = 1$ and $\boldsymbol{\theta} = \{\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3, \beta, \kappa\}$,

$$p(\mathbf{y}|\boldsymbol{\theta}) = \frac{1}{c(\beta,\kappa)} \exp\left\{\kappa\boldsymbol{\gamma}_1^\top \cdot \mathbf{y} + \beta\left[(\boldsymbol{\gamma}_2^\top \cdot \mathbf{y})^2 - (\boldsymbol{\gamma}_3^\top \cdot \mathbf{y})^2\right]\right\} = \frac{f(\mathbf{y}|\boldsymbol{\theta})}{c(\beta,\kappa)},$$

where $\kappa > 0, 0 \leq \beta < \kappa/2$. The parameters $\{\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3\}$ form a 3-dim orthonormal matrix with $\boldsymbol{\gamma}_i, i = 1, 2, 3$ is a 3×1 vector.

For a 3-dimensional FB$_5$ distribution, the mathematical expression for the normalising constant is

$$c(\beta,\kappa) = 2\pi \sum_{j=0}^{\infty} \frac{\Gamma(j+0.5)}{\Gamma(j+1)} \beta^{2j} (0.5\kappa)^{-2j-0.5} I_{2j+0.5}(\kappa)$$

.

Assume n independent observations from an FB$_5$ distribution, together with the auxiliary variables $\nu_i \sim \text{Expon}(c(\beta,\kappa)), i = 1, \cdots, n$. The posterior distribution

$$\begin{aligned}\pi(\boldsymbol{\theta}, \nu_{1:n}|\mathbf{y}_{1:n}) &\propto \pi(\boldsymbol{\theta}) \prod_{i=1}^{n} f(\mathbf{y}_i|\boldsymbol{\theta}) \exp\left(-\nu_i c(\beta,\kappa)\right) \\ &= \pi(\boldsymbol{\theta}) \exp\left(-\sum_{i=1}^{n} \nu_i c(\beta,\kappa)\right) \prod_{i=1}^{n} f(\mathbf{y}_i|\boldsymbol{\theta}).\end{aligned}$$

Calculating the normalising constant $c(\beta,\kappa)$ does not increase with the number of observations and again, we use the BP method for unbiasedly estimating $\exp(\cdot)$.

Classification prediction: In the empirical study, we assume n independent observations are from a mixture of two groups of the Kent distribution with unknown parameters $\boldsymbol{\theta}_g = \{\boldsymbol{\gamma}_{1,g}, \boldsymbol{\gamma}_{2,g}, \boldsymbol{\gamma}_{3,g}, \beta_g, \kappa_g\}, g = 1, 2$. Given the underlying group membership is provided and there is no hierarchical structure for the prior on $\boldsymbol{\theta}_g$, the posterior distribution for the parameters and the auxiliary variables is

$$\pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \nu_{1:n,1:2}|\mathbf{y}_{1:n_1,1}, \mathbf{y}_{1:n_2,2}) \propto \prod_{g=1}^{2} \left[\pi(\boldsymbol{\theta}_g) \exp\left(-\sum_{i=1}^{n_g} \nu_{i,g} c(\beta_g,\kappa_g)\right) \prod_{i=1}^{n_g} f(\mathbf{y}_{i,g}|\boldsymbol{\theta}_g)\right];$$

n_g is the number of observations belonging to group g and the auxiliary variable $\nu_{i,g} \sim \text{Expon}(c(\beta_g,\kappa_g))$.

For prediction, assign $\mathbf{y}_i$ to group 1 if $p(\mathbf{y}_i|\mathbf{y}_{train,1}) > p(\mathbf{y}_i|\mathbf{y}_{train,2})$ or to group 2, otherwise. The density $p(\mathbf{y}_i|\mathbf{y}_{train,g})$ is evaluated as

$$
\begin{aligned}
p(y_i|\mathbf{y}_{train,g}) &= \int_{\boldsymbol{\theta}_g} p(\mathbf{y}_i|\boldsymbol{\theta}_g)p(\boldsymbol{\theta}_g|\mathbf{y}_{train,g})d\boldsymbol{\theta}_g \\
&= \int_{\boldsymbol{\theta}_g}\int_{\nu_i} p(\mathbf{y}_i,\nu_i|\boldsymbol{\theta}_g)p(\boldsymbol{\theta}_g|\mathbf{y}_{train,g})d\nu_i d\boldsymbol{\theta}_g \\
&= \int_{\boldsymbol{\theta}_g}\int_{\nu_i} f(\mathbf{y}_i|\boldsymbol{\theta}_g)\exp(-\nu_i c(\beta_g,\kappa_g))p(\boldsymbol{\theta}_g|\mathbf{y}_{train,g})d\nu_i d\boldsymbol{\theta}_g \\
&= \int_{\boldsymbol{\theta}_g}\int_{\nu_i} f(\mathbf{y}_i|\boldsymbol{\theta}_g)E(\widehat{\exp}(-\nu_i c(\beta_g,\kappa_g)))p(\boldsymbol{\theta}_g|\mathbf{y}_{train,g})d\nu_i d\boldsymbol{\theta}_g.
\end{aligned}
$$

The last equation can be evaluated by importance sampling using the proposal $\nu_i \sim \text{Expon}(\widehat{c}(\beta_g,\kappa_g))$. The inner integral can be estimated by

$$
f(\mathbf{y}_i|\boldsymbol{\theta}_g)\frac{1}{\widehat{c}(\beta_g,\kappa_g)}\frac{1}{M}\sum_{i=1}^{M}\frac{\widehat{\exp}(-\nu_i c(\beta_g,\kappa_g))}{\exp(-\nu_i\widehat{c}(\beta_g,\kappa_g))}.
$$

The outer integral is computed by taking the average of the $\boldsymbol{\theta}_g$ iterates.

Appendix C

Supplementary material for Chapter 5

C.1 The BART MCMC algorithm

We follow the notation in Kapelner and Bleich (2013). Let $\mathbf{x}_i, i = 1, \dots, n$ be p-dimensional covariates. The binary treatment indicator is z_i with the observed response y_i. Denote T'_j as the structure for the jth tree which captures the prognostic (baseline) effect with $j = 1, \dots, m_0$. The corresponding parameters for the terminal nodes are denoted as $M' = \{\mu_{1j}, \dots, \mu_{b_j j}\}$. For the trees modelling the causal effects, let T_k be the structure for the kth tree with $k = 1, \dots, m_1$. As the terminal node splits with z, the corresponding node parameters are M_{0k} and M_{1k} respectively. It is worth noting that M_{0k}, M_{1k} have the same number of elements as they share an identical structure.

C.1.1 Data model

Recall the sum-of-trees model in (5.8). Observations are assumed to be independent of each other, with a normal distribution for residuals with a constant variance.

$$Y_i = \begin{cases} \sum_{j=1}^{m_0} g(\mathbf{x}_i; T_j^{'}, M_j^{'}) + \sum_{k=1}^{m_1} g(\mathbf{x}_i; T_k, M_{1k}) + \epsilon_i & \text{if } z_i = 1, \\ \sum_{j=1}^{m_0} g(\mathbf{x}_i; T_j^{'}, M_j^{'}) + \sum_{k=1}^{m_1} g(\mathbf{x}_i; T_k, M_{0k}) + \epsilon_i & \text{if } z_i = 0, \end{cases} \qquad \epsilon_i \sim N(0, \sigma^2)$$

for all $i = 1, \ldots, n$.

C.1.2 Prior

The prior is defined as,

$$\begin{aligned} &p((T_1', M_1'), \ldots, (T_{m_0}', M_{m_0}'), (T_1, M_{01}, M_{11}), \ldots, (T_{m_1}, M_{0m_1}, M_{1m_1}), \sigma^2) \\ &= \prod_{j=1}^{m_0} \left[p(M_j^{'}|T_j^{'}) p(T_j^{'}) \right] \prod_{k=1}^{m_1} \left[p(M_{0k}|T_k) p(M_{1k}|T_k) p(T_k) \right] p(\sigma^2). \end{aligned}$$

Section 5.3.2 describes the prior on T, T' In the implementation of Chapter 5, we set $\alpha, \beta = 0.95, 2$ for the prior on T' and $\alpha, \beta = 0.25, 3$ for the prior on T.

For $p(M_j^{'}|T_j^{'})$, the conditional distribution of each element in $M_j^{'}$ is set as a normal distribution with mean 0 and variance $1/m_0$. The same rule applies for $p(M_{0k}|T_k)$ and $p(M_{1k}|T_k)$ with the variance changing to $1/m_1$.

The prior on σ^2 is chosen as $\sigma^2 \sim \text{InvGamma}(\nu/2, \nu\lambda/2)$, such that λ is chosen to satisfy $\Pr(\sigma^2 < \widehat{\sigma}^2) = 0.9$ (Chipman et al., 2010). The estimate $\widehat{\sigma}^2$ is obtained by an OLS regression. The degrees of freedom parameter ν is set as 3 as suggested in Chipman et al. (2010).

C.1.3 Posterior sampling

The likelihood cannot be simply described by the parameters specified in the model due to its complexity. It is also difficult to express the posterior distribution explicitly. Chipman et al. (2010) use the Gibbs sampler within an iterative Bayesian backfitting MCMC algorithm. Without loss of generality, define the "residual" term of the tree index j or k

for each observation i, as the difference between y_i and the fitted value for all the other trees. We denote the residuals as R'_{-j} (for the baseline tree) or R_{-k} (for the treatment tree), i.e.,

$$R'_{-j} = \mathbf{y} - \sum_{i=1,i\neq j}^{m_0} g(\mathbf{x}; T_i^{'}, M_i^{'}) - \sum_{k=1}^{m_1} g(\mathbf{x}; T_k, M_{zk}),$$
$$R_{-k} = \mathbf{y} - \sum_{j=1}^{m_0} g(\mathbf{x}; T_j^{'}, M_j^{'}) - \sum_{i=1,i\neq k}^{m_1} g(\mathbf{x}; T_i, M_{zi}),$$

where $\mathbf{y}$, $\mathbf{x}$ refer to the observation vector, and the covariates of all the observations respectively. The parameter node is referred to as M_{zk} with $z \in \{0, 1\}$. Both R'_{-j} and R_{-k} refer to the residuals for all the units with $i = 1, \ldots, n$.

Before starting the backfitting algorithm, each tree is set to a root node and σ^2 is selected to be a positive random number prior to back-fitting. The sampling procedure consists of the following three parts.

Part 1: Cycle through $j = 1, \ldots m_0$, sample (T'_j, M'_j) given R'_{-j}, σ^2.

Part 2: Cycle through $k = 1, \ldots m_1$, sample (T_k, M_{0k}, M_{1k}) given R_{-k}, σ^2.

Part 3: Sample σ^2 given $\{(T'_j, M'_j), (T_k, M_{0k}, M_{1k})\}, j = 1, \ldots, m_0$, $k = 1, \ldots, m_1$.

Part 3 is the most straightforward. The full posterior conditional distribution of σ^2 is an inverse-gamma distribution, due to the choice of a conjugate prior.

Parts 1 and 2 are decomposed into two steps. For Part 1, we need to first sample from $p(T'_j | R'_{-j}, \sigma^2)$ and then sample $p(M'_j | T'_j, R'_{-j}, \sigma^2)$. Part 2 proceeds similarly. The second step is equivalent to updating the mean value for a normal distribution given a normal prior. The first step requires the MH algorithm, which we implement as follows.

C.1.3.1 Sample T'_j given R'_{-j}

The subscript j for T'_j, R'_{-j}, $j = 1, \ldots m_0$, is suppressed in this section for simplicity.

The acceptance ratio r of the MH algorithm is defined as,

$$r = \frac{p(T'_* \to T')p(T'_*|R',\sigma^2)}{p(T' \to T'_*)p(T'|R',\sigma^2)}.$$

The proposal for T, T_*, is either a grow or prune process. Details of the grow and prune processes are in Kapelner and Bleich (2013). The derivation for $p(T'|R',\sigma^2)$ is

$$p(T'|R',\sigma^2) = \frac{p(T',R'|\sigma^2)}{p(R'|\sigma^2)} = \frac{p(R'|T',\sigma^2)p(T'|\sigma^2)}{p(R'|\sigma^2)}.$$

The distribution of $p(T'|\sigma^2)$ is independent of σ^2, so it can be further simplified as $p(T')$. In combination, the formulation above gives the acceptance ratio

$$r = \frac{p(T'_* \to T')}{p(T' \to T'_*)} \times \frac{p(R'|T'_*,\sigma^2)}{p(R'|T',\sigma^2)} \times \frac{p(T'_*)}{p(T')};$$

r is the product of three components, each of which can be computed separately. The computation of the second term requires an integration over M'. The next section develops the full expression.

Computing $p(R'|T',\sigma^2)$

It is straightforward to show that $p(R'|T',M',\sigma^2)$ is a normal density. We obtain $p(R'|T',\sigma^2)$ by integrating out M' . Using the conditional independence of all observations given T', M', $p(R'|T',\sigma^2)$ is expressed as,

$$\begin{aligned} p(R'|T',\sigma^2) &= \int_{M'} p(R'|T',M',\sigma^2)p(M'|T',\sigma^2)dM' \\ &= \prod_{t=1}^{b} \int_{\mu'_t} p(R'_{\mu'_t}|T',\mu'_t,\sigma^2)p(\mu'_t|T',\sigma^2)d\mu'_t, \end{aligned}$$

where $M' = \{\mu'_1, \dots, \mu'_b\}$, $R'_{\mu'_t}$ stands for the residuals for the observation falling on the leaf with the node parameter μ'_t. The analytical solution of the integral is available as,

$$\begin{aligned} &\log \int_{\mu_{t'}} p(R'_{\mu_{t'}}|T',\mu'_t,\sigma^2)p(\mu'_t|T',\sigma^2)d\mu'_t = \\ &-\frac{1}{2}\log(\tau^2) - \frac{1}{2}\log A - \frac{1}{2}\frac{\sum r^2_{\mu'_t}}{\sigma^2} + \frac{1}{2A}\left(\frac{\sum r_{\mu'_t}}{\sigma^2}\right)^2, \end{aligned}$$

where $\mu'_t \sim N(0,\tau^2)$, $\sum r_{\mu'_t}$ is the sum of all the observations (residuals) falling on the leaf with node parameter μ'_t, and $A = \sigma^{-2}n_b + \tau^{-2}$, where n is the total number of observations (residuals) in μ'_t.

Computing $\dfrac{p(R'|T'_*, \sigma^2)}{p(R'|T', \sigma^2)}$ turns out to be simple by observing that T', T'_* only differ in one of the nodes as a consequence of the grow/prune process, which results in computing the likelihood for a small portion of observations only.

Computing $p(R|T, \sigma^2)$

The general procedure is similar to computing $p(R'|T', \sigma^2)$. The major difference is that T has the same structure for the treatment and the control groups, so the integration involves M_0 and M_1 respectively. The analytical result is not detailed here, as it is just a product of the probability from each group. It is worth noting that there must be a positive number of observations from the control and the treatment groups, given a specific tree structure. Otherwise, the acceptance ratio is set as zero for the proposal T_*. In the implementation, if the number of observations falling on the leaf is less than 5 based on T_*, we set the acceptance ratio to 0.

C.1.3.2 Sampling M'_j given T'_j, R'_{-j}, σ^2

We again suppress the subscript j. Recall that $M' = \{\mu'_1, \ldots, \mu'_b\}$. By exploiting the conditional independence between the μ's,

$$\begin{aligned} p(M'|T', R', \sigma^2) &= \prod_{t=1}^{b} p(\mu'_t|T', R', \sigma^2) \\ &\propto \prod_{t=1}^{b} p(R'_{\mu'_t}|T', \mu'_t, \sigma^2)p(\mu'_t), \end{aligned}$$

where $R'_{\mu'_t}$ denotes the residuals for the observations falling on the leaf with parameter μ'_t. As the prior on μ'_t and the likelihood are both normal distributions, the conditional posterior distribution for μ'_t is also a normal distribution with mean $\left(\dfrac{n_b}{\sigma^2} + \dfrac{1}{\tau^2}\right)^{-1} \dfrac{\sum r_{\mu'_t}}{\sigma^2}$, and variance $\left(\dfrac{n_b}{\sigma^2} + \dfrac{1}{\tau^2}\right)^{-1}$, where n_b is the number of observations in $r_{\mu'_t}$ and $\sum r_{\mu'_t}$ is the sum of all the observations (residuals) falling on the node parameter μ'_t.

C.1.3.3 Sampling M_{0k}, M_{1k} given T_k, R_{-k} , σ^2

The sampling procedure is almost identical to the one in the previous section. The node parameters M_{0k}, M_{1k} are updated separately.

References

Abadie, A. and Imbens, G. W. (2016). Matching on the estimated propensity score. *Econometrica*, 84(2):781–807.

Ai, H., Yongmiao, H., Lai, K. K., and Shouyang, W. (2008). Interval time series analysis with an application to the sterling-dollar exchange rate. *Journal of Systems Science and Complexity*, 21(4):558–573.

Allcott, H. and Knittel, C. (2019). Are consumers poorly informed about fuel economy? Evidence from two experiments. *American Economic Journal: Economic Policy*, 11(1):1–37.

Andrieu, C., De Freitas, N., Doucet, A., and Jordan, M. I. (2003). An introduction to MCMC for machine learning. *Machine Learning*, 50(1):5–43.

Andrieu, C. and Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, 37(2):697–725.

Andrieu, C. and Vihola, M. (2016). Establishing some order amongst exact approximations of MCMCs. *The Annals of Applied Probability*, 26(5):2661–2696.

Angrist, J. D. and Pischke, J.-S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Atchadé, Y. F., Lartillot, N., and Robert, C. (2013). Bayesian computation for statistical models with intractable normalizing constants. *Brazilian Journal of Probability and Statistics*, 27(4):416–436.

Atchadé, Y. F. and Rosenthal, J. S. (2005). On adaptive Markov chain Monte Carlo algorithms. *Bernoulli*, 11(5):815–828.

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46(3):399–424.

Bardenet, R., Doucet, A., and Holmes, C. (2017). On Markov chain Monte Carlo methods for tall data. *The Journal of Machine Learning Research*, 18(1):1515–1557.

Barnow, B. S., Cain, G. G., Goldberger, A. S., et al. (1980). Issues in the analysis of selectivity bias. *Evaluation Studies Review Annual*, 5:43–59.

Beaumont, M. A. (2003). Estimation of population growth or decline in genetically monitored populations. *Genetics*, 164(3):1139–1160.

Beranger, B., Lin, H., and Sisson, S. A. (2018). New models for symbolic data analysis. *arXiv preprint arXiv:1809.03659*.

Bertrand, P. and Goupil, F. (2000). Descriptive statistics for symbolic data. In *Analysis of Symbolic Data*, pages 106–124. Springer.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–225.

Betancourt, M. (2020). Robust Gaussian process modeling. *https://github.com/betanalpha/knitr_case_studies/tree/master/gaussian_processes, commite10083abbcdb65c745f840ab9d2da58229fa9af*.

Billard, L. (2006). Symbolic data analysis: What is it? In *Compstat 2006-Proceedings in Computational Statistics*, pages 261–269. Springer.

Billard, L. (2007). Dependencies and variation components of symbolic interval-valued data. In *Selected contributions in data analysis and classification*, pages 3–12. Springer.

Billard, L. (2008). Sample covariance functions for complex quantitative data. In *Proceedings of World IASC Conference, Yokohama, Japan*, pages 157–163.

Billard, L. and Diday, E. (2000). Regression analysis for interval-valued data. In *Data Analysis, Classification, and Related Methods*, pages 369–374. Springer.

Billard, L. and Diday, E. (2002). Symbolic regression analysis. In *Classification, Clustering, and Data Analysis*, pages 281–288. Springer.

Billard, L. and Diday, E. (2003). From the statistics of data to the statistics of knowledge: Symbolic data analysis. *Journal of the American Statistical Association*, 98(462):470–487.

Bock, H.-H. and Diday, E. (1999). *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data.* Springer Science & Business Media.

Botev, Z. I. (2017). The normal law under linear restrictions: Simulation and estimation via minimax tilting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(1):125–148.

Brito, P. (2002). Hierarchical and pyramidal clustering for symbolic data. *Journal of the Japanese Society of Computational Statistics*, 15(2):231–244.

Brito, P. (2014). Symbolic data analysis: Another look at the interaction of data mining and statistics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 4(4):281–295.

Brito, P. and Duarte Silva, A. P. (2012). Modelling interval data with normal and skew-normal distributions. *Journal of Applied Statistics*, 39(1):3–20.

Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. (2011). *Handbook of Markov Chain Monte Carlo.* CRC press.

Caimo, A. and Friel, N. (2011). Bayesian inference for exponential random graph models. *Social Networks*, 33(1):41–55.

Caliendo, M. and Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, 22(1):31–72.

Carter, L. L. and Cashwell, E. D. (1975). Particle-transport simulation with the Monte Carlo method. Technical report, Los Alamos Scientific Lab., N. Mex.(USA).

Cazes, P., Chouakria, A., Diday, E., and Schektman, Y. (1997). Extension de l'analyse en composantes principales à des données de type intervalle. *Revue de Statistique Appliquée*, 45(3):5–24.

Ceperley, D. and Dewing, M. (1999). The penalty method for random walks with uncertain energies. *The Journal of Chemical Physics*, 110(20):9812–9820.

Chavent, M. (1998). A monothetic clustering method. *Pattern Recognition Letters*, 19(11):989–996.

Chavent, M. and Lechevallier, Y. (2002). Dynamical clustering of interval data: Optimization of an adequacy criterion based on Hausdorff distance. In *Classification, Clustering, and Data Analysis*, pages 53–60. Springer.

Chipman, H. A., George, E. I., and McCulloch, R. E. (1998). Bayesian CART model search. *Journal of the American Statistical Association*, 93(443):935–948.

Chipman, H. A., George, E. I., and McCulloch, R. E. (2007). Bayesian ensemble learning. *Advances in Neural Information Processing Systems*, 19:265.

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298.

Chouakria, A., Diday, E., and Cazes, P. (1998). An improved factorial representation of symbolic objects. *Knowledge Extraction from Statistical Data*, 301:305.

Dawid, A. P. (2000). Causal inference without counterfactuals. *Journal of the American statistical Association*, 95(450):407–424.

De Carvalho, F. d. A., Neto, E. d. A. L., and Tenorio, C. P. (2004). A new method to fit a linear regression model for interval-valued data. In *Annual Conference on Artificial Intelligence*, pages 295–306. Springer.

De Carvalho, F. d. A. and Tenório, C. P. (2010). Fuzzy k-means clustering algorithms for interval-valued data based on adaptive quadratic distances. *Fuzzy Sets and Systems*, 161(23):2978–2999.

Del Moral, P., Doucet, A., and Jasra, A. (2006). Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436.

Deligiannidis, G., Doucet, A., and Pitt, M. K. (2018). The correlated pseudomarginal method. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(5):839–870.

Diday, E. (1989). Introduction à l'approche symbolique en analyse des données. *RAIRO-Operations Research-Recherche Opérationnelle*, 23(2):193–236.

Diday, E. and Vrac, M. (2005). Mixture decomposition of distributions by copulas in the symbolic data analysis framework. *Discrete Applied Mathematics*, 147(1):27–41.

Dietterich, T. G. et al. (2002). Ensemble learning. *The Handbook of Brain Theory and Neural Networks*, 2(1):110–125.

Dorie, V. (2021). dbarts: Discrete bayesian additive regression trees sampler. R package version 0.9-20.

Dorie, V., Hill, J., Shalit, U., Scott, M., and Cervone, D. (2019). Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1):43–68.

Doucet, A., Pitt, M. K., Deligiannidis, G., and Kohn, R. (2015). Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *Biometrika*, 102(2):295–313.

Dowe, D. L., Oliver, J. J., and Wallace, C. S. (1996). MML estimation of the parameters of the spherical fisher distribution. In *International Workshop on Algorithmic Learning Theory*, pages 213–227. Springer.

Efron, B. (1992). Bootstrap methods: Another look at the jackknife. In *Breakthroughs in Statistics*, pages 569–593. Springer.

Figueiredo, A. (2009). Discriminant analysis for the von Mises-Fisher distribution. *Communications in Statistics-Simulation and Computation*, 38(9):1991–2003.

Fisher, N. I., Lewis, T., and Embleton, B. J. (1993). *Statistical Analysis of Spherical Data*. Cambridge University Press.

Flury, T. and Shephard, N. (2011). Bayesian inference based only on simulated likelihood: Particle filter analysis of dynamic economic models. *Econometric Theory*, 27(5):933–956.

Fredrickson, G. H. and Andersen, H. C. (1984). Kinetic Ising model of the glass transition. *Physical Review Letters*, 53(13):1244.

Freedman, D. A. (2006). Statistical models for causation: What inferential leverage do they provide? *Evaluation Review*, 30(6):691–713.

Friel, N. and Pettitt, A. N. (2008). Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(3):589–607.

García-Ascanio, C. and Maté, C. (2010). Electric power demand forecasting using interval time series: A comparison between VAR and iMLP. *Energy Policy*, 38(2):715–725.

Garthwaite, P. H., Fan, Y., and Sisson, S. A. (2016). Adaptive optimal scaling of Metropolis–Hastings algorithms using the Robbins–Monro process. *Communications in Statistics-Theory and Methods*, 45(17):5098–5111.

Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409.

Gelman, A., Gilks, W. R., and Roberts, G. O. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, 7(1):110–120.

Gelman, A. and Meng, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, pages 163–185.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741.

Genz, A. (1992). Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, 1(2):141–149.

Geweke, J. and Zhou, G. (2015). Measuring the pricing error of the arbitrage pricing theory. *The Review of Financial Studies*, 9(2):557–587.

Geyer, C. J. (1991). Markov chain Monte Carlo maximum likelihood. *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, pages 156–163.

Gilks, W. R., Best, N. G., and Tan, K. K. (1995). Adaptive rejection Metropolis sampling within Gibbs sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 44(4):455–472.

Giordani, P. (2015). Lasso-constrained regression analysis for interval-valued data. *Advances in Data Analysis and Classification*, 9(1):5–19.

Giordani, P. and Kohn, R. (2010). Adaptive independent Metropolis–Hastings by fast estimation of mixtures of normals. *Journal of Computational and Graphical Statistics*, 19(2):243–259.

Gowda, K. C. and Diday, E. (1991). Symbolic clustering using a new dissimilarity measure. *Pattern Recognition*, 24(6):567–578.

Green, D. P. and Kern, H. L. (2012). Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. *Public Opinion Quarterly*, 76(3):491–511.

Haario, H., Saksman, E., and Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernoulli*, 7(2):223–242.

Hahn, P. R., Carvalho, C. M., Puelz, D., and He, J. (2018). Regularization and confounding in linear regression for treatment effect estimation. *Bayesian Analysis*, 13(1):163–182.

Hahn, P. R., Murray, J. S., and Carvalho, C. M. (2020). Bayesian regression tree models for causal inference: Regularisation, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15(3):965–1056.

Hainmueller, J. and Hazlett, C. (2014). Kernel regularised least squares: Reducing misspecification bias with a flexible and interpretable machine learning approach. *Political Analysis*, 22(2):143–168.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.

He, J., Yalov, S., and Hahn, P. R. (2019). XBART: Accelerated Bayesian additive regression trees. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, volume 89, pages 1130–1138. PMLR.

Heinrich, C., Maffioli, A., Vazquez, G., et al. (2010). A primer for applying propensity-score matching. *Inter-American Development Bank.*

Herrero, C. P. (2002). Ising model in small-world networks. *Physical Review E*, 65(6):066110.

Hill, J., Linero, A., and Murray, J. (2020). Bayesian additive regression trees: A review and look forward. *Annual Review of Statistics and Its Application*, 7:251–278.

Hill, J. and Su, Y.-S. (2013). Assessing lack of common support in causal inference using Bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children's cognitive outcomes. *The Annals of Applied Statistics*, 7(3):1386–1420.

Hill, J. L. (2011). Bayesian non-parametric modelling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960.

Hughes, J., Haran, M., and Caragea, P. C. (2011). Autologistic models for binary data on a lattice. *Environmetrics*, 22(7):857–871.

Hunter, D. R. and Handcock, M. S. (2006). Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics*, 15(3):565–583.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86(1):4–29.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Ising, E. (1925). Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik*, 31(1):253–258.

Jacob, P. E. and Thiery, A. H. (2015). On nonnegative unbiased estimators. *The Annals of Statistics*, 43(2):769–784.

Jerrum, M. and Sinclair, A. (1996). The Markov chain Monte Carlo method: An approach to approximate counting and integration. *Approximation Algorithms for NP-hard Problems, PWS Publishing*.

Kapelner, A. and Bleich, J. (2013). Bartmachine: Machine learning with Bayesian additive regression trees. *arXiv preprint arXiv:1312.2171*.

Kasarapu, P. (2015). Modelling of directional data using Kent distributions. *arXiv preprint arXiv:1506.08105*.

Kent, J. T. (1982). The Fisher-Bingham distribution on the sphere. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(1):71–80.

Kent, J. T., Ganeiber, A. M., and Mardia, K. V. (2013). A new method to simulate the Bingham and related distributions in directional data analysis with applications. *arXiv preprint arXiv:1310.8110*.

Kolmogorov, V. and Zabin, R. (2004). What energy functions can be minimized via graph cuts? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):147–159.

Kume, A. and Wood, A. T. (2005). Saddlepoint approximations for the Bingham and Fisher–Bingham normalising constants. *Biometrika*, 92(2):465–476.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165.

Lakshminarayanan, B., Roy, D., and Teh, Y. W. (2013). Top-down particle filtering for Bayesian decision trees. In *International Conference on Machine Learning*, pages 280–288. PMLR.

Le-Rademacher, J. and Billard, L. (2011). Likelihood functions and some maximum likelihood estimators for symbolic data. *Journal of Statistical Planning and Inference*, 141(4):1593–1602.

Lei, L. and Ding, P. (2021). Regression adjustment in completely randomised experiments with a diverging number of covariates. *Biometrika*, 108(4):815–828.

Lenz, W. (1920). Beitršge zum verstšndnis der magnetischen eigenschaften in festen kšrpern. *Physikalische Z*, 21:613–615.

Liang, F. (2010). A double Metropolis–Hastings sampler for spatial models with intractable normalizing constants. *Journal of Statistical Computation and Simulation*, 80(9):1007–1022.

Liang, F., Jin, I. H., Song, Q., and Liu, J. S. (2016). An adaptive exchange algorithm for sampling from distributions with intractable normalizing constants. *Journal of the American Statistical Association*, 111(513):377–393.

Lima Neto, E. d. A. and dos Anjos, U. U. (2015). Regression model for interval-valued variables based on copulas. *Journal of Applied Statistics*, 42(9):2010–2029.

Lin, H., Caley, M., and Sisson, S. A. (2017). Estimating global species richness using symbolic data meta-analysis. *arXiv preprint arXiv:1711.03202*.

Linero, A. R. (2018). Bayesian regression trees for high-dimensional prediction and variable selection. *Journal of the American Statistical Association*, 113(522):626–636.

Linero, A. R., Sinha, D., and Lipsitz, S. R. (2020). Semiparametric mixed-scale models using shared Bayesian forests. *Biometrics*, 76(1):131–144.

Linero, A. R. and Yang, Y. (2018). Bayesian regression tree ensembles that adapt to smoothness and sparsity. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(5):1087–1110.

Liu, H., Ong, Y.-S., Shen, X., and Cai, J. (2020). When Gaussian process meets big data: A review of scalable GPS. *IEEE Transactions on Neural Networks and Learning Systems*, 31(11):4405–4423.

Liu, Y., Ročková, V., and Wang, Y. (2021). Variable selection with ABC Bayesian forests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(3):453–481.

Lyne, A.-M., Girolami, M., Atchadé, Y., Strathmann, H., Simpson, D., et al. (2015). On Russian roulette estimates for Bayesian inference with doubly-intractable likelihoods. *Statistical Science*, 30(4):443–467.

Maia, A. L. S., de Carvalho, F. d. A., and Ludermir, T. B. (2008). Forecasting models for interval-valued time series. *Neurocomputing*, 71(16-18):3344–3352.

Marin, J.-M., Pudlo, P., Robert, C. P., and Ryder, R. J. (2012). Approximate Bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180.

McGree, J. M., Drovandi, C. C., White, G., and Pettitt, A. N. (2016). A pseudo-marginal sequential Monte Carlo algorithm for random effects models in Bayesian sequential design. *Statistics and Computing*, 26(5):1121–1136.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Møller, J., Pettitt, A. N., Reeves, R., and Berthelsen, K. K. (2006). An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika*, 93(2):451–458.

Morgan, S. L. and Winship, C. (2015). *Counterfactuals and Causal Inference*. Cambridge University Press.

Murray, I., Ghahramani, Z., and MacKay, D. J. C. (2006). MCMC for doubly-intractable distributions. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'06, page 359–366, Arlington, Virginia, USA. AUAI Press.

Murray, J. S. (2021). Log-linear Bayesian additive regression trees for multinomial logistic and count regression models. *Journal of the American Statistical Association*, 116(534):756–769.

Naesseth, A. C., Lindsten, F., and Schön, T. B. (2014). Sequential Monte Carlo for graphical models. *Advances in Neural Information Processing Systems*, 27:1862–1870.

Neal, P. and Roberts, G. (2006). Optimal scaling for partially updating MCMC algorithms. *The Annals of Applied Probability*, 16(2):475–515.

Neal, R. M. (2001). Annealed importance sampling. *Statistics and Computing*, 11(2):125–139.

Neiswanger, W., Wang, C., and Xing, E. (2013). Asymptotically exact, embarrassingly parallel MCMC. *arXiv preprint arXiv:1311.4780*.

Neto, E. d. A. L., de Carvalho, F. A., and Tenorio, C. P. (2004). Univariate and multivariate linear regression methods to predict interval-valued features. In *Australasian Joint Conference on Artificial Intelligence*, pages 526–537. Springer.

Neto, E. d. A. L. and de Carvalho, F. d. A. (2010). Constrained linear regression models for symbolic interval-valued variables. *Computational Statistics & Data Analysis*, 54(2):333–347.

Papaspiliopoulos, O. (2011). Monte Carlo probabilistic inference for diffusion processes: A methodological framework. *Bayesian Time Series Models*, pages 82–103.

Park, J. and Haran, M. (2018). Bayesian inference in the presence of intractable normalizing functions. *Journal of the American Statistical Association*, 113(523):1372–1390.

Park, J. and Haran, M. (2020). A function emulation approach for doubly intractable distributions. *Journal of Computational and Graphical Statistics*, 29(1):66–77.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Pitt, M. K., dos Santos Silva, R., Giordani, P., and Kohn, R. (2012). On some properties of Markov chain Monte Carlo simulation methods based on the particle filter. *Journal of Econometrics*, 171(2):134–151.

Pratola, M., Chipman, H., George, E., and McCulloch, R. (2017). Heteroscedastic BART using multiplicative regression trees. *arXiv preprint arXiv:1709.07542*.

Pratola, M. T., Chipman, H. A., Gattiker, J. R., Higdon, D. M., McCulloch, R., and Rust, W. N. (2014). Parallel Bayesian additive regression trees. *Journal of Computational and Graphical Statistics*, 23(3):830–852.

Propp, J. G. and Wilson, D. B. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures & Algorithms*, 9(1-2):223–252.

Quinonero-Candela, J. and Rasmussen, C. E. (2005). A unifying view of sparse approximate Gaussian process regression. *The Journal of Machine Learning Research*, 6:1939–1959.

Quiroz, M., Kohn, R., Villani, M., and Tran, M.-N. (2019). Speeding up MCMC by efficient data subsampling. *Journal of the American Statistical Association*, 114(526):831–843.

Quiroz, M., Tran, M.-N., Villani, M., Kohn, R., and Dang, K.-D. (2021). The block-Poisson estimator for optimally tuned exact subsampling MCMC. *Journal of Computational and Graphical Statistics*, 30(4):877–888.

Quiroz, M., Villani, M., Kohn, R., Tran, M.-N., and Dang, K.-D. (2018). Subsampling MCMC-An introduction for the survey statistician. *Sankhya A*, 80(1):33–69.

Rahman, P. A., Beranger, B., Roughan, M., and Sisson, S. A. (2020). Likelihood-based inference for modelling packet transit from thinned flow summaries. *arXiv preprint arXiv:2008.13424*.

Rice, G. E., Eide, I., Feder, P. I., and Gennings, C. (2018). Assessing human health risks using information on whole mixtures. In *Chemical Mixtures and Combined Chemical* *and Nonchemical Stressors: Exposure, Toxicity, Analysis, and Risk*, pages 421–463. Springer International Publishing.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407.

Roberts, G. O. and Rosenthal, J. S. (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16(4):351–367.

Roberts, G. O. and Rosenthal, J. S. (2009). Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics*, 18(2):349–367.

Robins, G., Pattison, P., Kalish, Y., and Lusher, D. (2007). An introduction to exponential random graph (p*) models for social networks. *Social Networks*, 29(2):173–191.

Ročková, V. and van der Pas, S. (2020). Posterior concentration for Bayesian regression trees and forests. *The Annals of Statistics*, 48(4):2108–2131.

Rokach, L. and Maimon, O. (2005). Top-down induction of decision trees classifiers-A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 35(4):476–487.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rossi, S., Heinonen, M., Bonilla, E., Shen, Z., and Filippone, M. (2021). Sparse Gaussian processes revisited: Bayesian approaches to inducing-variable approximations. In *International Conference on Artificial Intelligence and Statistics*, pages 1837–1845. PMLR.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 6(1):34–58.

Rubin, D. B. (1990). Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, 5(4):472–480.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392.

Rue, H. and Tjelmeland, H. (2002). Fitting Gaussian Markov random fields to Gaussian fields. *Scandinavian Journal of Statistics*, 29(1):31–49.

Schmidt, P. (1976). The non-uniqueness of the Australian Mesozoic palaeomagnetic pole position. *Geophysical Journal International*, 47(2):285–300.

Schmon, S. M., Deligiannidis, G., Doucet, A., and Pitt, M. K. (2021). Large-sample asymptotics of the pseudo-marginal method. *Biometrika*, 108(1):37–51.

Scott, S. L., Blocker, A. W., Bonassi, F. V., Chipman, H. A., George, E. I., and McCulloch, R. E. (2016). Bayes and big data: The consensus Monte Carlo algorithm. *International Journal of Management Science and Engineering Management*, 11(2):78–88.

Seeger, M. W., Williams, C. K., and Lawrence, N. D. (2003). Fast forward selection to speed up sparse Gaussian process regression. In *International Workshop on Artificial Intelligence and Statistics*, pages 254–261. PMLR.

Shephard, N. and Pitt, M. K. (1997). Likelihood analysis of non-Gaussian measurement time series. *Biometrika*, 84(3):653–667.

Sherlock, C., Thiery, A. H., Roberts, G. O., and Rosenthal, J. S. (2015). On the efficiency of pseudo-marginal random walk Metropolis algorithms. *The Annals of Statistics*, 43(1):238–275.

Snelson, E. and Ghahramani, Z. (2006). Sparse Gaussian processes using pseudo-inputs. *Advances in Neural Information Processing Systems*, 18:1257.

Snijders, T. A. (2002). Markov chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure*, 3(2):1–40.

Sparapani, R., Spanbauer, C., and McCulloch, R. (2021). Nonparametric machine learning and efficient computation with Bayesian additive regression trees: the BART R package. *Journal of Statistical Software*, 97(1):1–66.

Sparapani, R. A., Logan, B. R., McCulloch, R. E., and Laud, P. W. (2016). Nonparametric survival analysis using Bayesian additive regression trees (BART). *Statistics in Medicine*, 35(16):2741–2753.

Stoehr, J., Benson, A., and Friel, N. (2019). Noisy Hamiltonian Monte Carlo for doubly intractable distributions. *Journal of Computational and Graphical Statistics*, 28(1):220–232.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science: a Review Journal of the Institute of Mathematical Statistics*, 25(1):1–21.

Swiler, L. P., Gulian, M., Frankel, A. L., Safta, C., and Jakeman, J. D. (2020). A survey of constrained Gaussian process regression: Approaches and implementation challenges. *Journal of Machine Learning for Modeling and Computing*, 1(2).

Teles, P. and Brito, P. (2005). Modelling interval time series data. In *Proceedings of the 3rd IASC World Conference on Computational Statistics and Data Analysis, Limassol, Cyprus.*

Teles, P. and Brito, P. (2015). Modeling interval time series with space–time processes. *Communications in Statistics - Theory and Methods*, 44(17):3599–3627.

Tran, M.-N., Kohn, R., Quiroz, M., and Villani, M. (2016). The block pseudo-marginal sampler. *arXiv preprint arXiv:1603.02485*.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Wei, C. and Murray, I. (2017). Markov chain truncation for doubly-intractable inference. In *Artificial Intelligence and Statistics*, pages 776–784. PMLR.

Wendling, T., Jung, K., Callahan, A., Schuler, A., Shah, N., and Gallego, B. (2018). Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases. *Statistics in Medicine*, 37(23):3309–3324.

Whitaker, T., Beranger, B., and Sisson, S. A. (2020). Composite likelihood methods for histogram-valued random variables. *Statistics and Computing*, 30(5):1459–1477.

Williams, C. K. and Rasmussen, C. E. (2006). *Gaussian Processes for Machine Learning*, volume 2. MIT Press Cambridge, MA.

Wood, A. (1982). A bimodal distribution on the sphere. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 31(1):52–58.

Xu, W. (2010). *Symbolic Data Analysis: Interval-valued Data Regression.* PhD thesis, University of Georgia Athens, GA.

Zachary, W. W. (1977). An information flow model for conflict and fission in small groups. *Journal of anthropological research*, 33(4):452–473.

Zeldow, B., Re III, V. L., and Roy, J. (2019). A semiparametric modeling approach using Bayesian additive regression trees with an application to evaluate heterogeneous treatment effects. *The Annals of Applied Statistics*, 13(3):1989.

Zhang, X., Beranger, B., and Sisson, S. A. (2020). Constructing likelihood functions for interval-valued random variables. *Scandinavian Journal of Statistics*, 47(1):1–35.

Zhou, T., Elliott, M. R., and Little, R. J. (2019). Penalized spline of propensity methods for treatment comparison. *Journal of the American Statistical Association*, 114(525):1–19.