UNIVERSITY OF CALIFORNIA
RIVERSIDE

Fixed-Width Stopping Procedures for Markov Chain Monte Carlo

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Applied Statistics

by

Lei Gong

August 2015

Dissertation Committee:

Dr. James M. Flegal, Chairperson
Dr. Subir Ghosh
Dr. Stephen R. Spindler

The Dissertation of Lei Gong is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

It has been a long way since my first day in graduate school, and I have learned tremendously during my four years at University of California, Riverside.

First and foremost, I am fortunate to have studied with my advisor, Dr. James Flegal. I am gratitude for his guidance, friendship, patience and continuous support throughout my PhD study. His advice on research, career and life has been priceless to me. Working with him has been an inspiring experience. I could not have imagined having a better advisor and mentor.

I am thankful to the rest of my thesis and qualification committee: Dr. Xinping Cui, Dr. Subir Ghosh, Dr. Vagelis Hristidis, Dr. Daniel Jeske, Dr. Stephen Spindler and Dr. Shizhong Xu, for their encouragement, time, expertise and hard questions.

I also owe a debt of gratitude to the faculty, stuff, and fellow students in the Department of Statistics. In particular, I would like to thank Dr. Barry Arnold and Dr. Jun Li for challenging classes; Perla Fabelo and Paula Lemire for great assistance; Ashley Cacho, Hailu Chen, Shrabanti Chowdhury, Roberto Crackel, Zijian Huang, Jianan Hui, Ying Liu, Yue Liu, Yijia Wang, Zhen Xiao and Xin Zhang for helpful and interesting conversations.

I gratefully acknowledge the funding sources that supported my study: the Dean's Distinguished Fellowship from the Department of Statistics for the first three years and the Dissertation Year Fellowship from the Graduate Division for the last year.

Last but not least, I would like to thank my friends and family for their support and encouragement. I am especially grateful to my wife Chen Gao for her unfailing faith in me and allowing me to be as ambitious as I wanted. Her love and constant support have made the past years the best of my life. Thank you for being my wife and best friend. I owe you everything.

To my wife, Chen, and my parents,
who has always believed in me, even when I didn't.

ABSTRACT OF THE DISSERTATION

Fixed-Width Stopping Procedures for Markov Chain Monte Carlo

by

Lei Gong

Doctor of Philosophy, Graduate Program in Applied Statistics
University of California, Riverside, August 2015
Dr. James M. Flegal, Chairperson

Markov chain Monte Carlo (MCMC) simulations are commonly employed for estimating features of a target distribution, particularly for Bayesian inference. A fundamental challenge is determining when these simulations should stop. This dissertation begins by introducing relevant MCMC basics and discussing several existing techniques to terminate an MCMC simulation: the convergence diagnostics, using the effective sample size (ESS) as a stopping rule, and the fixed-width stopping rule (FWSR).

This dissertation continues by proposing the relative FWSRs that terminate the simulation when the width of a confidence interval is sufficiently small relative to the size of the target parameter. Specifically, we introduce two sequential stopping rules: the relative magnitude and the relative standard deviation FWSR in the context of MCMC. In each setting, we develop conditions to ensure the simulation will terminate with probability one and the resulting confidence intervals will have the proper coverage probability. The results are applicable in such MCMC estimation settings as expectation, quantile, or simultaneous multivariate estimation. We investigate the finite sample properties through a variety of examples, and provide some recommendations to practitioners.

New challenges present when the relative FWSRs are applied to terminate high-dimensional MCMC simulations. To this end, we propose using a modified relative standard deviation FWSR that terminates the simulation when the computational uncertainty is small relative to the posterior uncertainty. Further, we show this stopping rule is equivalent to stopping when the effective sample size is sufficiently large. Such a stopping rule has previously been shown to work well in settings with posteriors of moderate dimension. We

further illustrate its utility in high-dimensional simulations while overcoming some current computational issues. As examples, we consider two complex Bayesian analyses on spatially and temporally correlated datasets. The first involves a dynamic space-time model on weather station data and the second a spatial variable selection model on fMRI brain imaging data. The results show the modified sequential stopping rule is easy to implement, provides uncertainty estimates, and performs well in high-dimensional settings.

As a novel application, we propose using Bayesian model selection on linear mixed-effects models to compare multiple treatments with a control. A fully Bayesian approach is implemented to estimate the marginal posterior inclusion probability for each treatment, along with the model-averaged posterior distributions. It automatically traverses the model space and identifies subsets of predictors with nonzero fixed-effects coefficients; that is, it locates the model with the highest posterior probability. The resulting marginal inclusion probabilities provide a straightforward measure of the differences between treatments and the control. Default priors are proposed for model selection and a component-wise Gibbs sampler is developed for posterior computation. The proposed method is shown to work well using simulated data and the experimental data from a longitudinal study of mouse weight trajectories.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Markov chain Monte Carlo

Markov chain Monte Carlo (MCMC) methods are commonly employed in a Bayesian context to estimate features of complex and often high-dimensional posterior distributions, especially the Metropolis-Hastings algorithm (Hastings, 1970; Metropolis et al., 1953) and the Gibbs sampler (Gelfand and Smith, 1990; Geman and Geman, 1984). Properly implemented, MCMC techniques allow exploration of intractable probability distributions by constructing a Markov chain whose stationary distribution equals the desired distribution. Simply put, it is used to produce an estimate of some characteristics of a target distribution that is too complex to directly sample from.

Let π denote a probability distribution having support $\mathsf{X} \subseteq \mathbb{R}^d$, $d \geq 1$, about which we wish to make inference. This inference is usually based on various features (parameters) of π. For example, if $g : \mathsf{X} \to \mathbb{R}$, we may need to calculate

$$\mu_g := E_\pi[g(X)] = \int_{\mathsf{X}} g(x)\pi(dx) \ ,$$

or if $W \sim \pi$, then we might require quantiles of the distribution of $V = h(W)$ where $h : \mathsf{X} \to \mathbb{R}$. Specifically, if F_V denotes the cumulative distribution function of V, then we want to calculate

$$\xi_q := F_V^{-1}(q) = \inf\{v : F_V(v) \geq q\} \ .$$

In general, we denote $\theta \in \mathbb{R}^p$, $p \geq 1$ as a target parameter of interest with respect to π. Note that p can be smaller or larger than d, with large values of either indicating a

high-dimensional setting. Unfortunately, in most practically relevant settings we cannot calculate θ directly. Thus we are faced with a classical statistical problem; given a probability distribution π we want to estimate a fixed unknown feature.

Frequently π is such that MCMC is the only viable technique for estimating θ. The basic MCMC method entails constructing a time-homogeneous Harris ergodic Markov chain $X = \{X^{(0)}, X^{(1)}, \ldots\}$ on state space X with σ-algebra $\mathcal{B} = \mathcal{B}(\mathsf{X})$ and invariant distribution π. The popularity of MCMC methods result from the ease with which X can be simulated (Robert and Casella, 2004).

A major challenge for practitioners is determining how long to run an MCMC simulation. It is often difficult to decide when it is reasonable to believe that the samples are truly representative of the underlying stationary distribution of the Markov chain (Cowles and Carlin, 1996). On one hand, we definitely do not want premature termination that results in less reliable posterior inference; on the other hand, running an MCMC simulation too long is a waste of time and computational resources, especially in high-dimensional settings. Many experiments employ a fixed-time rule to terminate the simulation; that is, the procedure terminates after n iterations, where n is determined heuristically. Indeed, some simulations are so complex that this is the only practical approach, but that is not so for most experiments. This approach is unsatisfactory since practitioners would not have any confidence in the quality of the resulting estimates.

Alternatively, graphical methods are often utilized to evaluate if the chain has been run long enough. These include scatterplots, histograms, time series plots, autocorrelation plots and running mean plots (for a review see Geyer, 2011). In multivariate settings, a d-dimensional Markov chain is simulated to simultaneously estimate p-dimensional vector θ of π. These graphical techniques are obviously problematic when either d or p is large (Flegal and Jones, 2011). That is, good performance in marginal plots does not necessarily infer convergence in joint target distribution. Moreover, these plots soon become impractical to examine individually if either d pr p is large, which is often the case in modern Bayesian analysis.

Convergences diagnostics are also widely used among MCMC practitioners (for a review see Cowles and Carlin, 1996) to determine if n is sufficiently large. They assess the convergence of a chain statistically via outputs produced by the algorithm. Some of them are available in popular statistical softwares, e.g. R package `boa` (Smith, 2005) and `coda` (Best et al., 1995). Although "convergence diagnostic" was the central keyword of the

nineties in this area (Ceperley et al., 2012), these methods are mute about the quality of the resulting estimates (Flegal et al., 2008) and are essentially univariate. Moreover, they can introduce bias directly in to the estimates (Cowles et al., 1999).

In addition, some researchers use the effective sample size (ESS) as a run length diagnostic for MCMC simulations. The ESS, originally defined in survey sampling, measures the "effective number of independent sample" with respect to the correlated sample from an MCMC simulation. That is, it measures the size of an independently and identically distributed (i.i.d.) sample with the same standard error. A simulation is terminated once the ESS estimates are greater than a pre-specified threshold K (for e.g. see Atkinson et al., 2008; Drummond et al., 2006). Although the intuition behind this rule is clear, we are not aware of any theoretical discussions of the validity of using it as a stopping criteria for MCMC simulations.

Recently, a sequential fixed-width stopping rule (FWSR) is proposed by Jones et al. (2006) for MCMC. The FWSR terminates the simulation when an estimate is sufficiently accurate for the analytic purpose that motivates the inquiry. Intuitively, the simulation is terminated the first time a confidence interval width for a desired quantity is smaller than a user-specified absolute measure ϵ. Note that specifying a meaningful ϵ requires prior knowledge of the size of the quantity. It is an automated procedure and the total simulation effort will be random. Moreover, Flegal et al. (2008) and Jones et al. (2006) show this stopping rule is superior to using convergence diagnostics as a stopping criteria. However, it is impractical to require the users to know the size of the parameter of interest in advance and this becomes even more challenging in multivariate settings where a vector of ϵ's is needed.

Despite that the FWSR is theoretical validity and only constrained by a few assumptions, its implementation in practice is largely limited by the requirement of a pre-specified, absolute measure ϵ. To this end, we propose two variants of the FWSR, known as the relative FWSRs, that terminate a simulation once the width of a confidence interval is small relative to the size of the parameter. Specifically, we consider two measures of size, i.e. magnitude and standard deviation. The tuning parameter in the relative FWSRs is a relative measure ϵ that eliminates the requirement of the prior knowledge of the size of a desired quantity. Furthermore, practitioners only need to specify a single ϵ in multivariate settings.

We advocate the use of the relative standard deviation FWSR that terminates the

simulation when computational uncertainty is relatively small to the posterior uncertainty. Since modern Bayesian analysis often involves complex and high-dimensional posterior inference, we propose several modifications to improve the performance of the stopping rule, e.g. a strongly consistent variance estimator that significantly reduces memory usage and computational time. Also, we establish a connection between the relative standard deviation FWSR and using the ESS as a stopping rule, which justifies the theoretical validity of the latter.

The performance of the proposed stopping criterion is investigated and validated using several numerical examples, ranging from toy examples with a few parameters to a spatial Bayesian dynamic model with hundreds of parameters and a Bayesian fMRI model with thousands of parameters. A novel linear mixed-effects model that utilize Bayesian model selection techniques to compare multiple treatments to a control is proposed and studied using an experimental dataset from a longitudinal study of mouse weight trajectories.

The rest of this dissertation is organized as follows. Chapter 2 introduces two relative FWSRs that eliminate the requirement of the prior knowledge of the size of a desired quantity, and investigates the procedures for estimating expectation and quantiles. Chapter 3 proposes modifications for the relative standard deviation FWSR on high-dimensional settings, along with establishing the connection between the FWSR and using the ESS as a stopping criteria. Chapter 4 considers a novel application that utilizes Bayesian model selection techniques on linear mixed-effects model to compare multiple treatments with a control for a longitudinal study of mouse weight trajectories.

1.2 Existing Stopping Rules

Some details are given for the existing stopping rules introduced in Section 1.1 that include their theoretical assumptions, formulations and practicality of implementation. Specifically, three popular convergence diagnostics are described, i.e. Gelman and Rubin's diagnostic (Gelman and Rubin, 1992), Geweke's diagnostic (Geweke, 1992) and Raftery and Lewis's diagnostic (Raftery and Lewis, 1992b); so are using the ESS as a stopping rule and the fixed-width procedure.

Specificity requires some notation. Suppose we simulate X for n iterations, where n is finite. Define Z_n as an estimator of θ from the observed chain. Outside of toy examples,

no matter how long our simulation, there will be an unknown Monte Carlo error, $Z_n - \theta$. While it is impossible to assess this error directly, we can obtain its approximate sampling distribution if a Markov chain central limit theorem (CLT) holds. That is, if

$$\sqrt{n}\,(Z_n - \theta) \xrightarrow{d} \mathrm{N}(0, \sigma_\theta^2) \tag{1.1}$$

as $n \to \infty$ where $\sigma_\theta^2 \in (0, \infty)$. Let λ_θ^2 denote the posterior variance associated with θ. That is, if an i.i.d. sample from π is available then λ_θ^2 is the asymptotic variance in the CLT associated with θ. It is important to note that due to the correlation present in a Markov chain $\sigma_\theta^2 \neq \lambda_\theta^2$, except in trivial cases.

1.2.1 Convergence Diagnostics

There is an extensive literature on convergence diagnostics for MCMC simulations. Cowles and Carlin (1996) provide an excellent review and we direct interested readers to their paper for details. Due to their popularity in the statistical community and implementaion in statistical software, we discuss the convergence diagnostics of Gelman and Rubin (1992), of Geweke (1992) and of Raftery and Lewis (1992b).

Gelman and Rubin's Diagnostic

The Gelman and Rubin diagnostic (Gelman and Rubin, 1992) is a two-step method based on normal approximation to the posterior distribution. Before a simulation begins, the diagnostic requires an over-dispersed estimate of the target distribution, and a set of starting points for m independent chains. Suppose a simulation is run for $2n$ iterations, the Gelman and Rubin's method uses the last n iterations to re-estimate the target distribution of the quantity as a Student's t distribution that involves both the between-chain variance and the within-chain variance. That is, the shrink factor

$$\sqrt{\hat{R}} = \sqrt{\left(\frac{n-1}{n} + \frac{m+1}{mn}\frac{B}{W}\right)\frac{df}{df-2}}\,,$$

where B is the variance between the means from the m parallel chains, W is the average of the m within-chain variances, and df is the degrees of freedom of the approximately t distribution.

Gelman and Rubin (1992) suggest iteratively increasing iterations for the parallel chains and re-calculating the shrink factors for quantities of interest until all are near 1. At termination, assuming each parallel chain is of length $2n$, the posterior inference is carried out using the combined values from the last n iterations from all chains. Although the method is inherently univariate, the authors recommend to apply this procedure to -2 times the log of the posterior density to summarize convergence of the joint posterior density.

A few of the challenges faced with the implementation of the Gelman and Rubin diagnostic, e.g. its reliance on the user's ability to find an over-dispersed starting distribution and its inefficiency introduced by the requirement of multiple chains.

Geweke's Diagnostic

The Geweke diagnostic (Geweke, 1992) is based on a hypothesis test that the mean estimates of two non-overlapping parts of the Markov chain have converged. Unlike Gelman and Rubin's method, it requires only a single Markov chain. As a rule of thumb, Geweke (1992) suggested to take first 0.1 and last 0.5 proportions of the Markov chain. The resulting test statistic is univariate by its nature and the z-score is constructed as follows,

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\hat{s}_1(0)/n_1 + \hat{s}_2(0)/n_2}},$$

where $\bar{x}_1, \bar{x}_2$ are the sample average and $\hat{s}_1(0), \hat{s}_2(0)$ are spectral density estimates at zero frequency for the two parts of the Markov chain, respectively. In multivariate settings, given the hypothesis-testing nature of diagnostic, one way to confirm the convergence of the joint density is to counteract the problem of multiple comparisons using Bonferroni correction.

A number of criticisms of the Geweke diagnostic, e.g. its sensitivity to the choice of the spectral window and its lacking of detailed specification of the implementation by the author.

Raftery and Lewis's Diagnostic

The Raftery and Lewis Diagnostic (Raftery and Lewis, 1992b) is based on an evaluation of the accuracy of estimation of the percentiles q. It reports the number of samples needed to reach the desired accuracy of the percentiles. An initial chain of length

$N_{\min}$ needs to be run, where $N_{\min}$ is the minimum number of iterations to obtain the desired accuracy of estimation if the samples were independent. The implementation of this method in the R package `coda` (Best et al., 1995) takes in acceptable tolerance r for q and a probability s of being within the given tolerance, and outputs the number of iterations n that should be run and the length of the burn-in period necessary to satisfy the specified conditions.

Disadvantages of the Raftery and Lewis diagnostic include its variability in estimation given different initial chains for the same problem and its impracticality in requiring re-diagnosis for every quantile of interest.

1.2.2 Effective Sample Size

Given n iterations in a Markov chain, the ESS measures the size of an i.i.d. sample with the same standard error, or the "effective number of independent samples". This quantity is frequently used by practitioners as a run length diagnostic, terminating the simulation once ESS estimates are greater than a pre-specified threshold K (for e.g. see Atkinson et al., 2008; Drummond et al., 2006).

Note that the ESS is not uniquely defined. One way to define ESS is described in Kass et al. (1998) and Robert and Casella (2004),

$$\mathrm{ESS}_\theta = \frac{n}{1 + 2\sum_{k=1}^{\infty} \rho_k(g)},$$

where $\rho_k(g)$ is the autocorrelation of lag k for g. This calculation is implemented in many R packages, such as `coda` (Best et al., 1995) and `mcmcse` (Flegal and Hughes, 2012).

An alternative approach to define ESS as in the custom of survey sampling (Kish, 1965; Liu et al., 1998), where

$$\mathrm{ESS}_\theta = \frac{n}{\sigma_\theta^2/\lambda_\theta^2}.$$

In practice, we estimate this quantity by replacing the parameters with their strongly consistent estimates, i.e.

$$\widehat{\mathrm{ESS}}_n = \frac{n}{\hat{\sigma}_n^2/\hat{\lambda}_n^2}. \tag{1.2}$$

Therefore, the alternative approach provides a strongly consistent estimate of ESS.

In univariate settings, using ESS as a stopping rule is equivalent to terminating the simulation when the estimated ESS is above the threshold K. That is, the time at

which the simulation terminates is defined by

$$\tilde{T}(K) = \inf\left\{n \geq 0 : \widehat{\mathrm{ESS}}_n \geq K\right\} .$$

Although this stopping criteria is intuitively sound, its theoretical validity has not been fully studied. Also, the ESS estimate most commonly implemented in statistical softwares is not strongly consistent.

1.2.3 Fixed-width Procedure

Suppose a Markov chain CLT holds (1.1), one can construct a $(1-\delta)100\%$ confidence interval for θ with width

$$w_\delta = 2z_{\delta/2}\frac{\hat{\sigma}_n}{\sqrt{n}}, \tag{1.3}$$

where $z_{\delta/2}$ is a critical value from a standard Normal distribution and $\hat{\sigma}_n^2$ is a strongly consistent estimator of σ_θ^2. The width at (1.3) allows analysts to report the uncertainty in their estimates and users to assess the practical reliability.

Moreover, Jones et al. (2006) propose constructing a sequential fixed-width stopping rule based on w_δ for MCMC simulations. By controlling the width w_δ, the stopping criteria controls the computational uncertainty of the simulation. Under a few regularity conditions (for a review see Flegal and Gong, 2015; Jones et al., 2006), the FWSR has been proved to be theoretically valid, in the sense that the resulting confidence interval has the right coverage probability $1-\delta$. Specifically, the FWSR terminates an MCMC simulation the first w_δ is below a pre-specified threshold ϵ, i.e.

$$\tilde{T}(\epsilon, \delta) = \inf\left\{n \geq 0 : 2z_{\delta/2}\hat{\sigma}_n/\sqrt{n} + p(n) \leq \epsilon\right\} .$$

Notice that $p(n)$ is introduced in the stopping rule to prevent pre-mature termination due to poorly behaved estimate $\hat{\sigma}_n$, when the sample size n is small. It is a positive function that decreases monotonically such that $p(n) = o(n^{-1/2})$ as $n \to \infty$ and let n^* be the desired minimum simulation effort (a reasonable default is $p(n) = \epsilon I(n \leq n^*) + n^{-1}$).

In addition, for such a stopping rule to work well, we need to have prior knowledge of the size of the desired quantity θ, which is not always practical for real problems. In multivariate settings, it becomes even harder to specify a threshold ϵ for each parameter based on their sizes.

Chapter 2

Relative Fixed-width Stopping Rules

This chapter introduces two variants of the FWSR that eliminate the requirement for the prior knowledge of the size of the parameter. Specifically, two relative FWSRs, i.e. the relative magnitude FWSR and the relative standard deviation FWSR, are proposed that terminate an MCMC simulation once the width of a confidence interval is sufficiently small relative to the size of the target parameter. Conditions for asymptotic validity are developed and finite sample properties are studies through a variety of simulations. The content of this chapter is primarily contained in Flegal and Gong (2015).

2.1 Introduction

Markov chain Monte Carlo (MCMC) methods allow exploration of intractable probability distributions by constructing a Markov chain whose stationary distribution equals the desired distribution. A major challenge for practitioners is determining how long to run an MCMC simulation. Many experiments employ a fixed-time rule to terminate the simulation; that is, the procedure terminates after n iterations, where n is determined heuristically. Indeed, some simulations are so complex that this is the only practical approach, but that is not so for most experiments.

Alternatively, some practitioners use convergence diagnostics to determine if n is sufficiently large (for a review see Cowles and Carlin, 1996). Although practical, these methods are mute about the quality of the resulting estimates (Flegal et al., 2008). Moreover,

they can introduce bias directly in to the estimates (Cowles et al., 1999).

We instead advocate terminating the simulation when an estimate is sufficiently accurate for the analytic purpose that motivates the inquiry. In other words, the simulation is terminated the first time a confidence interval width for a desired quantity is sufficiently small. We refer to such a procedure as a sequential fixed-width stopping rule and note the total simulation effort will be random.

As we show later, fixed-width methods are especially desirable because they are theoretically justified and constrained by few assumptions. The simplest fixed-width rule, first studied in MCMC by Jones et al. (2006), stops the simulation when the width of a confidence interval based on an ergodic average is less than a user-specified value, say ϵ. Flegal et al. (2008) and Jones et al. (2006) show this stopping rule is superior to using convergence diagnostics as a stopping criteria.

In this chapter, we introduce relative fixed-width stopping rules that eliminate the need to specify an absolute value for ϵ. Specifically, the simulation is terminated the first time the width of a confidence interval is sufficiently small relative to the *size* of a target parameter. We consider two measures of size, magnitude and standard deviation. Further, we illustrate the utility of these rules for simultaneous estimation of multiple parameters.

Specificity requires some notation. Let π denote a probability distribution having support $\mathsf{X} \subseteq \mathbb{R}^d$, $d \geq 1$, about which we wish to make inference. This inference is usually based on various features (parameters) of π. For example, if $g : \mathsf{X} \to \mathbb{R}$, we may need to calculate

$$\mu_g := E_\pi[g(X)] = \int_{\mathsf{X}} g(x)\pi(dx) \, ,$$

or if $W \sim \pi$, then we might require quantiles of the distribution of $V = h(W)$ where $h : \mathsf{X} \to \mathbb{R}$. Specifically, if F_V denotes the cumulative distribution function of V, then we want to calculate

$$\xi_q := F_V^{-1}(q) = \inf\{v : F_V(v) \geq q\} \, .$$

In general, we denote $\theta \in \mathbb{R}$ as a target parameter of interest with respect to π. Unfortunately, in most practically relevant settings we cannot calculate θ directly. Thus we are faced with a classical statistical problem; given a probability distribution π we want to estimate a fixed unknown feature.

Frequently π is such that MCMC is the only viable technique for estimating θ. The basic MCMC method entails constructing a time-homogeneous Harris ergodic Markov chain

$X = \left\{X^{(0)}, X^{(1)}, \ldots\right\}$ on state space X with σ-algebra $\mathcal{B} = \mathcal{B}(\mathsf{X})$ and invariant distribution π. The popularity of MCMC methods result from the ease with which X can be simulated (Robert and Casella, 2004).

Suppose we simulate X for n iterations, where n is finite. Define Z_n as an estimator of θ from the observed chain. Outside of toy examples, no matter how long our simulation, there will be an unknown Monte Carlo error, $Z_n - \theta$. While it is impossible to assess this error directly, we can obtain its approximate sampling distribution if a Markov chain central limit theorem (CLT) holds. That is, if

$$\sqrt{n}\,(Z_n - \theta) \xrightarrow{d} \mathrm{N}(0, \sigma_\theta^2) \tag{2.1}$$

as $n \to \infty$ where $\sigma_\theta^2 \in (0, \infty)$.

Let $\hat{\sigma}_n^2$ denote an estimator of σ_θ^2. This allows construction of a $(1-\delta)100\%$ confidence interval for θ with width

$$w_\delta = 2 z_{\delta/2} \frac{\hat{\sigma}_n}{\sqrt{n}} \tag{2.2}$$

where $z_{\delta/2}$ is a critical value from a standard Normal distribution. The width at (2.2) allows analysts to report the uncertainty in their estimates and users to assess the practical reliability.

Moreover, we will use w_δ to construct sequential fixed-width stopping rules. First, we require a final bit of notation. Let λ_θ^2 denote the posterior variance associated with θ. That is, if an i.i.d. sample from π is available then λ_θ^2 is the asymptotic variance in the CLT associated with θ. It is important to note that due to the correlation present in a Markov chain $\sigma_\theta^2 \neq \lambda_\theta^2$, except in trivial cases. For estimation of μ_g it is easy to show $\lambda_\theta^2 = \mathrm{Var}[g(X)]$. For estimation of ξ_q, we have $\lambda_\theta^2 = q(1-q)/\left(f_V(\xi_q)\right)^2$ where f_V is the density associated with F_V.

Our work advocates stopping the simulation the first time w_δ is sufficiently small. We consider three distinct stopping rules: (i) an absolute precision rule that terminates when $w_\delta < \epsilon$, (ii) a relative magnitude rule that terminates when $w_\delta < \epsilon\,|\theta|$ and (iii) a relative standard deviation rule that terminates when $w_\delta < \epsilon\lambda_\theta$.

First, we investigate the theoretical properties of the three stopping rules. Previously, Glynn and Whitt (1992) established conditions for asymptotic validity of (i) and (ii).

Asymptotic validity is important since it implies the simulation will terminate w.p.1 and the resulting confidence intervals will have the right coverage probability. In this chapter, we extend these results to establish asymptotic validity of the stopping rule (iii).

Next, we consider applying fixed-width stopping rules in MCMC simulations. Flegal et al. (2008), Flegal and Jones (2010) and Jones et al. (2006) have previously investigated (i) for MCMC expectation estimation. We are not aware of any prior use of fixed-width methods for quantile estimation or any use of (ii) or (iii) as a stopping rule in MCMC. The rule (iii) has significant promise in Bayesian applications since the simulation terminates the first time the length of a confidence interval is less than an ϵth fraction of the magnitude of the standard deviation of θ. In other words, the simulation stops when an estimate of θ is sufficiently accurate relative to an associated posterior standard deviation. Another substantial benefit of rule (iii) is it is easy to implement in multivariate settings since ϵ can remain constant.

There are two main assumptions for asymptotic validity. First, we require a functional central limit theorem (FCLT) for the Monte Carlo error. Fortunately, Markov chains frequently enjoy a FCLT under identical conditions as those that ensure a CLT. Second, we require a strongly consistent estimator of the associated asymptotic variance, that is $\hat{\sigma}_n^2 \to \sigma_\theta^2$ almost surely as $n \to \infty$. Many commonly used MCMC estimators of σ_θ^2 can satisfy this condition, see e.g. Flegal and Jones (2010), Doss et al. (2014), Hobert et al. (2002), and Jones et al. (2006).

Finally, we investigate the finite sample properties of relative fixed-width stopping rules through three examples. Our first example considers an independence Metropolis sampler to explore an exponential random variable. Our second example considers exploring a mixture of bivariate Normal distributions with Metropolis Hastings and Gibbs samplers. While these are only toy examples, we will use true parameter values to illustrate the utility of our stopping rules. Our final example considers a Bayesian version of a logistic regression to model the presence or absence of the freshwater eel *Anguilla australis*.

Using these examples, we terminate the simulation with the three distinct fixed-width stopping rules and calculate confidence intervals for a vector of target parameters. Over replicated simulations, all the finite sample empirical coverage probabilities are close to a specified nominal level. Thus, fixed-width stopping rules provide a theoretically valid and practically accurate procedure to determine when to stop a MCMC simulation.

For Bayesian practitioners, we advocate the relative standard deviation fixed-width

stopping rule (iii) since it is easy to implement and applicable in multivariate settings without a priori knowledge of the target parameter size. Specifically in multivariate settings, one can terminate the first time the length of a confidence interval is sufficiently small for each parameter of interest. Given the natural appeal of such a stopping rule, some practitioners have likely already adopted a similar informal approach. As our examples show, setting $\epsilon = 0.02$ provides excellent results in a wide variety of univariate and multivariate settings.

The rest of this chapter is organized as follows. Section 2.2 formally introduces relative fixed-width stopping rules and establishes asymptotic validity. Section 2.3 investigates fixed-width stopping procedures when estimating expectations and quantiles. Section 2.4 studies the finite sample properties in three numerical examples and concludes with a discussion that provides some recommendations to practitioners.

2.2 Sequential fixed-width procedures

In this section, we obtain conditions that ensure asymptotic validity of fixed-width procedures. The primary assumptions are the limiting process must satisfy a FCLT and $\hat{\sigma}_n^2 \to \sigma_\theta^2$ w.p.1 as $n \to \infty$. Hence, our results can be applied very generally. Section 2.3 outlines checkable sufficient conditions for the most common MCMC settings, estimating expectations and quantiles.

Recall our goal is to estimate a parameter $\theta \in \mathbb{R}$. To this end, we assume there exists an $\mathbb{R}$-valued stochastic process $\{Z_n : n \geq 1\}$ called the estimation process for which $Z_n \to \theta$ in probability. Asymptotic validity requires the estimation process satisfies a FCLT as follows. For ease of exposition, we consider a slightly more general $\mathbb{R}$-valued stochastic process $Z = \{Z(t) : t \geq 0\}$ for which $Z(t) \to \theta$ in probability as $t \to \infty$. Let $D(0, \infty)$ denote the space of right-continuous $\mathbb{R}$-valued functions with left limits on the open interval $(0, \infty)$. We assume that Z has sample paths in $D(0, \infty)$ and consider the family of scaled processes in $D(0, \infty)$ for $\epsilon > 0$

$$\mathcal{Z}_\epsilon(t) = \epsilon^{-1/2} \left(Z(t/\epsilon) - \theta \right) \ , \text{ where } t > 0.$$

We will say a FCLT holds if there exists a constant $\sigma_\theta > 0$ such that as $\epsilon \to 0$

$$\mathcal{Z}_\epsilon(t) \xrightarrow{d} \sigma_\theta B(t)/t \ ,$$

in $D(0, \infty)$ where $B(t)$ denotes a standard Brownian motion process $\{B(t) : t \geq 0\}$. Fortu-

nately, in many situations a FCLT holds under the same conditions as those required for an ordinary CLT (as we will discuss in Section 2.3).

Next, define an interval

$$C[n] = \left(Z_n - z_{\delta/2}\hat{\sigma}_n/\sqrt{n}\ ,\ Z_n + z_{\delta/2}\hat{\sigma}_n/\sqrt{n}\right)\ .$$

If a CLT at (2.1) holds and $\hat{\sigma}_n$ is weakly consistent for σ_θ, then $C[n]$ achieves the nominal coverage level as the sample size $n \to \infty$. Thus we have a valid confidence interval provided the sample size is permitted to go to ∞.

Now consider a sequential procedure that terminates the simulation when the length of a confidence interval drops below a prescribed level ϵ. We will refer to this type of stopping rule as an absolute precision fixed-width stopping rule. For such a rule, the time at which the simulation terminates is defined by

$$\tilde{T}(\epsilon) = \inf\left\{n \geq 0 : 2z_{\delta/2}\hat{\sigma}_n/\sqrt{n} \leq \epsilon\right\}\ .$$

Unfortunately, use of this stopping rule is insufficient because $\tilde{T}(\epsilon)$ can terminate much too early if $\hat{\sigma}_n$ is poorly behaved for small n (Glynn and Whitt, 1992). Instead, suppose $p(n)$ is a positive function that decreases monotonically such that $p(n) = o(n^{-1/2})$ as $n \to \infty$ and let n^* be the desired minimum simulation effort (a reasonable default is $p(n) = \epsilon I(n \leq n^*) + n^{-1}$). Then an absolute precision stopping rule terminates the simulation at

$$T_1(\epsilon) = \inf\left\{n \geq 0 : 2z_{\delta/2}\hat{\sigma}_n/\sqrt{n} + p(n) \leq \epsilon\right\}\ .$$

The following result, an immediate consequence of Theorem 1 in Glynn and Whitt (1992), yields asymptotic validity of the sequential stopping rule $T_1(\epsilon)$. Note the desired coverage probability will be obtained in an asymptotic sense as $\epsilon \to 0$.

Proposition 1. *Suppose a FCLT for the Monte Carlo error holds. If $\hat{\sigma}_n \to \sigma_\theta$ w.p.1 as $n \to \infty$, then as $\epsilon \to 0$ the simulation will terminate w.p.1 and*

$$Pr\left(\theta \in C[T_1(\epsilon)]\right) \to 1 - \delta\ .$$

Remark 2. *Glynn and Whitt (1992) show weak consistency of $\hat{\sigma}_n$ is not enough to ensure asymptotic validity .*

The stopping rule $T_1(\epsilon)$ has previously been used for estimating expectations in MCMC (Flegal et al., 2008; Flegal and Jones, 2010; Jones et al., 2006). We further show this rule works well for MCMC estimation of quantiles in the following section. The challenge in both settings is finding a strongly consistent estimator of σ_θ.

One can consider a variant of the stopping rule $T_1(\epsilon)$ known as a relative precision stopping rule, which avoids having to choose an absolute value for ϵ. Simply put, the simulation is run until the length of a confidence interval is less than an ϵth fraction of the magnitude of the parameter of interest, θ. Using Z_n as an estimator of θ yields the following relative magnitude stopping rule

$$T_2(\epsilon) = \inf\left\{n \geq 0 : 2z_{\delta/2}\hat{\sigma}_n/\sqrt{n} + p(n) \leq \epsilon\,|Z_n|\right\} \ .$$

For large n, $T_2(\epsilon)$ will behave like $T_1(\epsilon|\theta|)$. The following obtains asymptotic validity of $T_2(\epsilon)$, which is a direct consequence of Theorem 3 in Glynn and Whitt (1992).

Proposition 3. *Suppose a FCLT for the Monte Carlo error holds and $|\theta| > 0$. If $Z_n \to \theta$ w.p.1 and $\hat{\sigma}_n \to \sigma_\theta$ w.p.1 as $n \to \infty$, then as $\epsilon \to 0$ the simulation will terminate w.p.1 and*

$$Pr\left(\theta \in C[T_2(\epsilon)]\right) \to 1 - \delta \ .$$

Note that Proposition 3 requires $Z_n \to \theta$ w.p.1 along with necessary conditions of Proposition 1. In general stochastic simulations, this condition does not immediately follow from (2.1) (see Example 2 of Glynn and Whitt, 1988) but is readily available when θ is an expectation via the Markov chain strong law of large numbers (SLLN).

While $T_2(\epsilon)$ has some support in the operations research literature, it makes little intuitive sense in Bayesian settings. Specifically, if $\theta = 0$ then $T_2(\epsilon)$ will be theoretically invalid and poorly behaved in finite simulations. In addition, $T_2(\epsilon)$ could be problematic even when θ is not equal to zero, which we illustrate through example in Section 2.4.

Given the popularity of MCMC in Bayesian settings, it is useful to consider another specifically designed variant of $T_1(\epsilon)$. To this end, we propose a stopping rule that terminates the simulation when the length of a confidence interval is less than an ϵth fraction of the magnitude of λ_θ, i.e. the posterior standard deviation of θ. Suppose $\hat{\lambda}_n$ is an estimator of

λ_θ and consider the following stopping rule

$$T_3(\epsilon) = \inf\left\{n \geq 0 : 2z_{\delta/2}\hat{\sigma}_n/\sqrt{n} + p(n) \leq \epsilon\hat{\lambda}_n\right\} \ .$$

For large n, $T_3(\epsilon)$ will behave like $T_1(\epsilon\lambda_\theta)$. The benefit of using $T_3(\epsilon)$ is that ϵ is selected as a fraction rather than in the units of the target parameter. Hence, a single value of ϵ would be appropriate for target parameters of any magnitude. Naturally, decreasing ϵ would decrease the uncertainty of the resulting estimates and could be done simultaneously for multiple parameters. The following establishes asymptotic validity of $T_3(\epsilon)$, which we prove in Section 2.5.1.

Theorem 4. *Suppose a FCLT for the Monte Carlo error holds and $\lambda_\theta > 0$. If $\hat{\lambda}_n \to \lambda_\theta$ w.p.1 and $\hat{\sigma}_n \to \sigma_\theta$ w.p.1 as $n \to \infty$, then as $\epsilon \to 0$ the simulation will terminate w.p.1 and*

$$Pr\left(\theta \in C[T_3(\epsilon)]\right) \to 1 - \delta \ .$$

Note the only additional condition required for Theorem 4 is a strongly consistent estimator of λ_θ. For expectations, an estimator is readily available via the Markov chain SLLN. In the case of quantiles, we discuss a viable estimator in the following section.

The benefit of the stopping rule $T_3(\epsilon)$ is twofold. First, one only needs to specify a relative ϵ, and hence no knowledge about the magnitude is required. Second, when estimating multiple parameters a single ϵ will suffice to obtain estimates whose uncertainty will be comparable relative to their standard deviations. In other words, we have developed a simple, yet informative, stopping criteria applicable in multivariate settings. In these settings, one could address the issue of multiplicity by adjusting the critical value appropriately. We illustrate this procedure via examples in Section 2.4, and show the resulting simultaneous confidence regions obtain at least the nominal coverage probability.

Remark 5. *Asymptotic validity of relative stopping rules can be established if a FCLT is replaced by a more general $\mathbb{R}$-valued stochastic process (Glynn and Whitt, 1992). The generalization enables consideration of θ that follow non-Normal asymptotic distributions.*

Remark 6. *A more general relative stopping rule that terminates when $w_\delta < \epsilon\nu_\theta$ can be established for any ν_θ such that $|\nu_\theta| > 0$ provided there exists an estimator $\hat{\nu}_n \to \nu_\theta$ w.p.1. Thus, one could consider relative stopping rules setting ν_θ as the interquartile range, the length of a Bayesian credible region, and so on.*

2.3 Applications

This section demonstrates that fixed-width stopping rules are appropriate for MCMC estimation of expectations and quantiles. This is an important contribution since we know of no other formal stopping criteria applicable in both settings. Raftery and Lewis (1992a) propose a heuristic approach to terminating an MCMC simulation when the primary interest is quantile estimation. However, Brooks and Roberts (1999) argue "in the case where quantiles themselves are not of interest, this method should be used with caution".

First, we require a bit more notation to describe sufficient mixing conditions for a Markov chain CLT and consistent estimation of the asymptotic variance. An interested reader is directed to Meyn et al. (2009) and Roberts and Rosenthal (2004) for more on Markov chain theory.

Recall X is a Harris ergodic Markov chain on state space X with σ-algebra $\mathcal{B} = (\mathsf{X})$ and invariant distribution π. Denote the n-step Markov kernel associated with X as $P^n(x, dy)$ for $n \in \mathbb{N}$. Then if $A \in \mathcal{B}(\mathsf{X})$ and $k \in \{0, 1, 2, \ldots\}$, $P^n(x, A) = \Pr(X_{k+n} \in A | X_k = x)$. Let $\|\cdot\|$ denote the total variation norm. Let $M : \mathsf{X} \mapsto \mathbb{R}^+$ and $\gamma : \mathbb{N} \mapsto \mathbb{R}^+$ be decreasing such that

$$\|P^n(x, \cdot) - \pi(\cdot)\| \leq M(x)\gamma(n) \; . \tag{2.3}$$

Polynomial ergodicity of order m where $m \geq 0$ means (2.3) holds with $E_\pi M < \infty$ and $\gamma(n) = n^{-m}$ for all $X_0 = x$. *Geometrical ergodicity* means (2.3) holds with $\gamma(n) = t^n$ for some $0 < t < 1$ for all $X_0 = x$. *Uniform ergodicity* means (2.3) holds with M bounded and $\gamma(n) = t^n$ for some $0 < t < 1$.

Establishing (2.3) directly can be challenging, but some constructive techniques are available (Jarner and Roberts, 2002; Meyn et al., 2009). Most literature on MCMC algorithms focuses on establishing geometric and uniform ergodicity, see e.g. Hobert (2011), Jones and Hobert (2001), Johnson et al. (2013), Mengersen and Tweedie (1996), Roberts and Tweedie (1996) and Tierney (1994). Less has been said concerning polynomial ergodicity, but an interested reader is directed to Douc et al. (2004), Fort and Moulines (2000), Fort and Moulines (2003), Jarner and Roberts (2002), Jarner and Roberts (2007) and Jarner and Tweedie (2003).

2.3.1 Expectations

For MCMC estimation of an expectation, one can obtain all the necessary conditions for asymptotic validity of fixed-width stopping rules. Let $g : \mathsf{X} \to \mathbb{R}$, then we consider estimation of

$$\mu_g := E_\pi[g(X)] = \int_{\mathsf{X}} g(x)\pi(dx) \; .$$

Estimating μ_g is natural by appealing a Markov chain SLLN, a special case of the Birkhoff Ergodic Theorem (p. 558 Fristedt and Gray, 1997). Specifically, if $E_\pi|g| < \infty$ then w.p.1

$$Z_n := \bar{g}_n := \frac{1}{n}\sum_{i=0}^{n-1} g(X^{(i)}) \to \mu_g \text{ as } n \to \infty \; .$$

Hence the SLLN yields strongly consistent estimators of μ_g and $\lambda^2_\theta = \text{Var}[g(X)]$ (provided $E_\pi g^2 < \infty$) necessary for Proposition 3 and Theorem 4, respectively.

We can obtain an approximate sampling distribution for the Monte Carlo error via a Markov chain CLT if

$$\sqrt{n}(\bar{g}_n - \mu_g) \xrightarrow{d} \text{N}(0, \sigma^2_g) \tag{2.4}$$

as $n \to \infty$ where $\sigma^2_g \in (0, \infty)$. Conditions that ensure (2.4) can be found in Chan and Geyer (1994), Jones (2004), Meyn et al. (2009), Roberts and Rosenthal (2004) and Tierney (1994). For example, if X is geometrically ergodic and $E_\pi|g|^{2+\epsilon} < \infty$ for some $\epsilon > 0$, then (2.4) holds. Fortunately, Markov chains frequently enjoy a FCLT under the same conditions (Ibragimov, 1962; Jones et al., 2006; Oodaira and Yoshihara, 1972).

There are many strongly consistent variance estimation techniques applicable for σ^2_g in MCMC settings including batch means (Flegal and Jones, 2010; Jones et al., 2006), spectral variance techniques (Flegal and Jones, 2010) and regenerative simulation (Hobert et al., 2002; Mykland et al., 1995). We consider only non-overlapping batch means (BM) because it is easy to implement and available in many software packages, e.g. the `mcmcse` package available on `CRAN`.

In BM the output is broken into a_n batches where each batch is b_n iterations in length. Suppose the algorithm is run for a total of $n = a_n b_n$ iterations and define

$$\bar{Y}_j := \frac{1}{b_n}\sum_{i=(j-1)b_n+1}^{jb_n} g(X_i) \qquad \text{for } j = 1, \ldots, a_n \; .$$

The BM estimate of σ_g^2 is

$$\hat{\sigma}_n^2 = \frac{b_n}{a_n - 1} \sum_{j=1}^{a_n} (\bar{Y}_j - \bar{g}_n)^2 \ . \tag{2.5}$$

In general, the BM estimator at (2.5) is not a consistent estimator of σ_g^2. However, Jones et al. (2006) establish necessary conditions for $\hat{\sigma}_n^2 \to \sigma_g^2$ with probability 1 as $n \to \infty$ if the batch size and the number of batches are allowed to increase as the overall length of the simulation increases. Setting $b_n = \lfloor n^\tau \rfloor$ and $a_n = \lfloor n/b_n \rfloor$, the regularity conditions require that X be geometrically ergodic, $E_\pi |g|^{2+\epsilon_1+\epsilon_2} < \infty$ for some $\epsilon_1 > 0$, $\epsilon_2 > 0$ and $(1 + \epsilon_1/2)^{-1} < \tau < 1$. A common choice of $\tau = 1/2$ (i.e., $b_n = \lfloor \sqrt{n} \rfloor$ and $a_n = \lfloor n/b_n \rfloor$) has been shown to work well in applications (Flegal et al., 2008; Flegal and Jones, 2010; Jones et al., 2006).

Remark 7. *Most sampling plans require storing the entire Markov chain to allow for recalculations as the batch size increases with n. If storage is a concern, one could consider increasing the batch size of the form $b_n \in \{2, 4, 8, ..., 2^k, ...\}$ in an effort to reduce memory usage. One can establish strong consistency for the BM variance estimator with such a sampling plan using results in Jones et al. (2006) and Bednorz and Latuszyński (2007).*

2.3.2 Quantiles

It is routine to estimate univariate quantiles associated with π, especially in Bayesian applications. To this end, let $W \sim \pi$ and recall $h : \mathsf{X} \to \mathbb{R}$. Setting $V = h(W)$, we consider estimation of the quantiles associated with the univariate distribution of V. Suppose F_V denotes the cumulative distribution function of V, then our goal is to obtain

$$\xi_q := F_V^{-1}(q) = \inf\{v : F_V(v) \geq q\} \ .$$

We further suppose that $F_V(x)$ is absolutely continuous and has continuous density function $f_V(x)$ such that $0 < f_V(\xi_q) < \infty$.

Little has been formally said regarding MCMC estimation of quantiles, but we outline the current state of understanding (for details see Doss et al., 2014). A natural estimator of ξ_q is the inverse of the empirical distribution function given by

$$Z_n := \hat{\xi}_{n,q} = Y_{n(j)} \quad \text{where} \quad j - 1 < nq \leq j \ , \tag{2.6}$$

where $Y_{n(j)}$ denotes the jth order statistic of $\{Y_0, \ldots, Y_{n-1}\} = \{h(X_0), \ldots, h(X_{n-1}),\}$. If X is Harris recurrent and then $\hat{\xi}_{n,q} \to \xi_q$ w.p.1 as $n \to \infty$ (Doss et al., 2014).

Under stronger mixing conditions on X, one can obtain a Markov chain CLT. To this end, define

$$\sigma^2(y) := \mathrm{Var}_\pi\left[I(Y_0 \leq y)\right] + 2\sum_{k=1}^{\infty} \mathrm{Cov}_\pi\left[I(Y_0 \leq y), I(Y_k \leq y)\right] \ .$$

Suppose X is polynomially ergodic of order $m > 11$ and $\sigma^2(\xi_q) > 0$, then as $n \to \infty$

$$\sqrt{n}(\hat{\xi}_{n,q} - \xi_q) \xrightarrow{d} \mathrm{N}(0, \gamma^2(\xi_q)) \ , \tag{2.7}$$

where $\gamma^2(\xi_q) = \sigma^2(\xi_q)/[f_V(\xi_q)]^2$ (Doss et al., 2014). A FCLT holds for uniformly ergodic chains via sufficient conditions in Sen (1972) that can be verified with results in Jones (2004). As a direction of future work, it is likely a FCLT holds under polynomial ergodicity combining results in Doss et al. (2014) and Sen (1972).

Estimation of the variance from the asymptotic Normal distribution at (2.7) is broken into two parts. First, we plug in $\hat{\xi}_{n,q}$ for ξ_q and separately consider estimating $f_V(\hat{\xi}_{n,q})$ and $\sigma^2(\hat{\xi}_{n,q})$. Estimating $f_V(\hat{\xi}_{n,q})$ uses a kernel density approach with a gaussian kernel, which we denote as $\hat{f}_V(\hat{\xi}_{n,q})$. There are well known conditions guaranteeing strongly consistent estimation of the density at a point (see e.g. Kim and Lee, 2005; Yu, 1993).

We will use BM for estimating $\sigma^2(\hat{\xi}_{n,q})$. Suppose we have $n = a_n b_n$ iterations, then for $k = 0, \ldots, a_n - 1$ define $\bar{U}_k(\hat{\xi}_{n,q}) := b_n^{-1} \sum_{i=0}^{b_n - 1} I(Y_{kb_n+i} \leq \hat{\xi}_{n,q})$. The BM estimate of $\sigma^2(\hat{\xi}_{n,q})$ is

$$\hat{\sigma}^2_{BM}(\hat{\xi}_{n,q}) = \frac{b_n}{a_n - 1} \sum_{k=0}^{a_n - 1} \left(\bar{U}_k(\hat{\xi}_{n,q}) - \bar{U}_n(\hat{\xi}_{n,q})\right)^2 \ .$$

Combining $\hat{f}_V(\hat{\xi}_{n,q})$ and $\hat{\sigma}^2_{BM}(\hat{\xi}_{n,q})$, we estimate $\gamma^2(\xi_q)$ with

$$\hat{\gamma}^2(\hat{\xi}_{n,q}) := \frac{\hat{\sigma}^2_{BM}(\hat{\xi}_{n,q})}{[\hat{f}_V(\hat{\xi}_{n,q})]^2} \ .$$

This approach is implemented in the R package `mcmcse` which is used to perform the computations in our examples. Doss et al. (2014) outline the conditions that ensure strong consistency of this estimator.

The relative standard deviation fixed-width stopping rule of Theorem 4 requires a

estimation of

$$\lambda_\theta = \frac{\sqrt{q(1-q)}}{f_V(\xi_q)} \, .$$

We use the same kernel density estimate resulting in

$$\hat{\lambda}_n = \frac{\sqrt{q(1-q)}}{\hat{f}_V(\hat{\xi}_{n,q})} \, .$$

2.4 Numerical studies

This section investigates the finite sample properties of fixed-width stopping rules through a variety of simulations. In each example, we independently repeat the MCMC simulation to evaluate the resulting finite sample confidence intervals. Naturally, this evaluation requires the true parameter values. In our first two examples, the true values are readily available. In our final example, the truth was estimated using an independent long run of the MCMC sampler. Overall, the empirical coverage probabilities obtained via fixed-width stopping rules are remarkably close to the nominal level.

Each simulation considered both expectations and quantiles with the following common methodology. For a single replication, the same MCMC draws were used in applying the three stopping rules. Further, we uniformly set $p(n) = \epsilon I(n < n^*) + n^{-1}$ and estimate σ_θ^2 via BM methods with $b_n = \lfloor \sqrt{n} \rfloor$ calculated with the `mcmcse` package. Finally, standard errors for the empirical coverage probabilities equal $\sqrt{\hat{p}(1-\hat{p})/r}$ where r is the number of replications.

2.4.1 Exponential distribution

Consider an Exp(1) target distribution, i.e. $f(x) = e^{-x} I(x > 0)$. It is easy to show that $E[X] = 1$ and $F^{-1}(q) = \log\left\{(1-q)^{-1}\right\}$, which we use to evaluate finite sample confidence intervals obtained via fixed-width methods. We will sample from $f(x)$ using an independence Metropolis sampler with an Exp(1/2) proposal and note this chain is geometrically ergodic (Jones and Hobert, 2001).

First, consider estimation of $E[X]$ using each combination of $T_i(\epsilon)$ for $i \in \{1, 2, 3\}$ and $\epsilon \in \{0.10, 0.05, 0.02\}$. The chain was started from 1 and ran for a minimum of $n^* = 1000$ iterations. If the stopping criteria was not met, an additional 500 iterations were added to the chain before checking again. The simulation was repeated for 2000 replications to

evaluate the resulting coverage probabilities.

Table 2.1 summarizes the mean and standard deviation of the number of iterations at termination along with the resulting coverage probabilities. All the coverage probabilities are close to the 0.90 nominal level suggesting all three stopping rules are preforming well. Note the mean iterations are approximately equal, which is expected since $E[X] = 1$ and $\lambda_\theta = \text{Var}[X] = 1$.

	Length (SD)	$E[X]$	Length (SD)	$\xi_{.5}$
$T_1(0.10)$	2.44E3 (4.9E2)	0.8840	2.70E3 (5.9E2)	0.8580
$T_1(0.05)$	8.89E3 (1.2E3)	0.8940	1.01E4 (1.5E3)	0.8805
$T_1(0.02)$	5.36E4 (4.7E3)	0.8875	6.17E4 (5.4E3)	0.8775
$T_2(0.10)$	2.44E3 (4.8E2)	0.8895	5.40E3 (9.4E2)	0.8800
$T_2(0.05)$	8.90E3 (1.2E3)	0.8910	2.07E4 (2.4E3)	0.8820
$T_2(0.02)$	5.35E4 (4.7E3)	0.8870	1.29E5 (9.1E3)	0.8830
$T_3(0.10)$	2.45E3 (4.7E2)	0.8885	2.79E3 (5.2E2)	0.8650
$T_3(0.05)$	8.90E3 (1.2E3)	0.8880	1.03E4 (1.3E3)	0.8820
$T_3(0.02)$	5.35E4 (4.6E3)	0.8895	6.23E4 (5.2E3)	0.8770

Table 2.1: Summary of coverage probabilities for estimation of $E[X]$ and $\xi_{.5}$ based on 2000 replications and 0.90 nominal level.

Next, consider estimation of the median, $\xi_{.5}$, using the same simulation settings. Table 2.1 summarizes the results from 2000 replications. Again the results are very close to the 0.90 nominal level, though slightly lower than those for estimating the mean. Here we have $\xi_{.5} = 0.693$ and $\sqrt{0.5(1-0.5)}/e^{-\xi_{.5}} = 1$, hence for fixed ϵ we expect $T_1(\epsilon)$ and $T_3(\epsilon)$ to be similar and $T_2(\epsilon)$ to be larger.

Finally, consider estimating the mean and an 80% Bayesian credible region simultaneously, which we denote as $\Phi = (E[X], \xi_{.1}, \xi_{.9})$. Due to increased computation time, each chain was run for a minimum of $n^* = 10000$ iterations with an additional 5000 added between checks. The simulation was terminated the first time the length of a confidence interval was sufficiently small for each parameter in Φ. To adjust for multiplicity, we apply a Bonferonni approach. Specifically, we set individual confidence intervals to have a coverage probability of $0.90^{1/3} = 0.9655$ resulting in a simultaneous confidence region with coverage probability of at least 0.90.

The simulation was repeated for 2000 replications with each combination of $T_i(\epsilon)$ for $i \in \{1, 2, 3\}$ and $\epsilon \in \{0.10, 0.05, 0.02\}$. Table 2.2 summarizes the simulation results.

We can see the individual coverage probabilities improve as ϵ decreases, especially in the case of $\xi_{.1}$. For $\epsilon = 0.02$, all the individual coverage probabilities are remarkably close to the nominal level of 0.9655. Note the observed confidence region coverage probabilities are above the 0.90 nominal level, which is unsurprising due to correlation between parameters in Φ.

	Length (SD)	$E[X]$	$\xi_{.1}$	$\xi_{.9}$	Region
$T_1(0.10)$	2.88E4 (3.9E3)	0.963	0.989	0.963	0.930
$T_1(0.05)$	1.07E5 (9.7E3)	0.965	0.979	0.962	0.923
$T_1(0.02)$	6.53E5 (3.3E4)	0.965	0.967	0.968	0.917
$T_2(0.10)$	6.71E4 (5.9E3)	0.969	0.979	0.964	0.925
$T_2(0.05)$	2.29E5 (1.4E4)	0.966	0.974	0.963	0.920
$T_2(0.02)$	1.29E6 (5.0E4)	0.964	0.963	0.970	0.915
$T_3(0.10)$	1.00E4 (0)	0.962	0.991	0.955	0.927
$T_3(0.05)$	2.31E4 (2.9E3)	0.963	0.983	0.958	0.921
$T_3(0.02)$	1.30E5 (9.1E3)	0.961	0.970	0.965	0.914

Table 2.2: Summary of coverage probabilities for estimation of Φ based on 2000 replications. Individual confidence intervals have a 0.9655 nominal level, resulting in a 0.90 nominal level confidence region.

2.4.2 Mixture of bivariate Normals

Consider a mixture of bivariate Normals $\mathbf{X} = [X_1, X_2]^T = p\mathbf{Y}_1 + (1-p)\mathbf{Y}_2$, where

$$\mathbf{Y}_1 = \begin{bmatrix} Y_{11} \\ Y_{12} \end{bmatrix} \sim \mathbb{N}_2 \left(\begin{bmatrix} \mu_{11} \\ \mu_{12} \end{bmatrix}, \begin{bmatrix} \sigma_{11}^2 & 0 \\ 0 & \sigma_{12}^2 \end{bmatrix} \right) \quad \text{and} \quad \mathbf{Y}_2 = \begin{bmatrix} Y_{21} \\ Y_{22} \end{bmatrix} \sim \mathbb{N}_2 \left(\begin{bmatrix} \mu_{21} \\ \mu_{22} \end{bmatrix}, \begin{bmatrix} \sigma_{21}^2 & 0 \\ 0 & \sigma_{22}^2 \end{bmatrix} \right).$$

In this example, we choose $p = 0.25$, $\mu_{11} = 1$, $\mu_{12} = 10$, $\mu_{21} = 2.5$, $\mu_{22} = 25$, $\sigma_{11} = 0.5$, $\sigma_{12} = 5$, $\sigma_{21} = 0.7$ and $\sigma_{22} = 7$.

We first sample from $f(\mathbf{X})$ with two different component-wise Metropolis random walk algorithms, one with Uniform proposals and another with Normal proposals. For the Uniform proposals, we apply a $Unif(-3, 3)$ and $Unif(-30, 30)$ random walk for the X_1 and X_2 dimensions, respectively. For the Normal proposals, we apply a $N(0, 3^2)$ and $N(0, 30^2)$ random walk for the X_1 and X_2 dimensions, respectively. It can be shown that these chains are geometrically ergodic (Jarner and Hansen, 2000).

Consider estimation of $\Phi = (E[X], \xi_{.1}, \xi_{.9})$ using fixed-width stopping rules $T_i(\epsilon)$

for $i \in \{1, 2, 3\}$ and $\epsilon \in \{0.10, 0.05, 0.02\}$. We ran the chain for a minimum of $n^* = 5000$ iterations and added 1000 iterations between checking the stopping criteria. This simulation was repeated for 1000 independent replications.

Table 2.3 summarizes the mean and standard deviation of the number of iterations at termination along empirical coverage probabilities from the Uniform and Normal proposals. Notice for both samplers, the coverage probabilities improve as ϵ decreases and are close to the 0.95 nominal level once $\epsilon = 0.02$. It appears the Metropolis random walk with Normal proposals is mixing faster since the overall simulation effort is substantially lower than that of the Uniform proposals. This difference in simulation effort illustrates the importance of specifying a good proposal distribution in MCMC simulations.

Uniform	Length (SD)	$E[X_1]$	$\xi_{.1,X_1}$	$\xi_{.9,X_1}$	$E[X_2]$	$\xi_{.1,X_2}$	$\xi_{.9,X_2}$
$T_1(0.10)$	14,658 (3.4E3)	0.930	0.932	0.917	0.936	0.945	0.937
$T_1(0.05)$	59,869 (9.1E3)	0.934	0.922	0.939	0.940	0.934	0.953
$T_1(0.02)$	391,566 (3.1E4)	0.956	0.944	0.945	0.956	0.948	0.953
$T_2(0.10)$	20,897 (5.0E3)	0.929	0.933	0.911	0.931	0.936	0.938
$T_2(0.05)$	85,401 (1.2E4)	0.950	0.926	0.934	0.929	0.925	0.942
$T_2(0.02)$	556,821(3.9E4)	0.953	0.946	0.954	0.950	0.938	0.956
$T_3(0.10)$	8,827 (1.0E3)	0.926	0.928	0.899	0.920	0.922	0.920
$T_3(0.05)$	35,733 (2.9E3)	0.924	0.938	0.931	0.934	0.928	0.937
$T_3(0.02)$	233,312 (1.3E4)	0.954	0.955	0.959	0.948	0.958	0.956

Normal	Length (SD)	$E[X_1]$	$\xi_{.1,X_1}$	$\xi_{.9,X_1}$	$E[X_2]$	$\xi_{.1,X_2}$	$\xi_{.9,X_2}$
$T_1(0.10)$	8,028 (1.5E3)	0.946	0.939	0.939	0.934	0.943	0.937
$T_1(0.05)$	29,844 (3.7E3)	0.927	0.936	0.948	0.917	0.932	0.953
$T_1(0.02)$	186,061 (1.3E4)	0.952	0.936	0.952	0.943	0.946	0.938
$T_2(0.10)$	11,307 (2.1E3)	0.949	0.933	0.940	0.940	0.944	0.943
$T_2(0.05)$	42,338 (4.6E3)	0.911	0.943	0.956	0.937	0.934	0.951
$T_2(0.02)$	261,741 (1.6E4)	0.940	0.938	0.956	0.949	0.938	0.945
$T_3(0.10)$	5,114 (3.2E2)	0.944	0.950	0.933	0.936	0.936	0.924
$T_3(0.05)$	17,654 (1.8E3)	0.922	0.930	0.943	0.925	0.921	0.939
$T_3(0.02)$	112,626 (7.6E3)	0.933	0.946	0.941	0.941	0.930	0.940

Table 2.3: Summary of coverage probabilities for estimations of Φ using a Metropolis random walk with Uniform and Normal proposals based on 1000 replications and a 0.95 nominal level.

Next, we consider a Gibbs sampler using the full conditional densities, i.e.

$$f_{X_1|X_2}(x_1|x_2) = P_{X_2}Y_{11} + (1 - P_{X_2})Y_{21} \text{ and}$$
$$f_{X_2|X_1}(x_2|x_1) = P_{X_1}Y_{12} + (1 - P_{X_1})Y_{22} \,,$$

where

$$P_{X_2} = \left(1 + \frac{(1-p)\sigma_{12}}{p\sigma_{22}} \exp\left\{\frac{1}{2}\left(\left(\frac{x_2 - \mu_{12}}{\sigma_{12}}\right)^2 - \left(\frac{x_2 - \mu_{22}}{\sigma_{22}}\right)^2\right)\right\}\right)^{-1},$$

and

$$P_{X_1} = \left(1 + \frac{(1-p)\sigma_{11}}{p\sigma_{21}} \exp\left\{\frac{1}{2}\left(\left(\frac{x_1 - \mu_{11}}{\sigma_{11}}\right)^2 - \left(\frac{x_1 - \mu_{21}}{\sigma_{21}}\right)^2\right)\right\}\right)^{-1}.$$

Note, $X_1|X_2 = x_2$ and $X_2|X_1 = x_1$ are easy to sample from since they are mixtures of Normal random variables.

Table 2.4 summarizes the results for the Gibbs sampler. Notice, the coverage probabilities do not uniformly improve as ϵ decreases. However, they are all close to the nominal 0.95 level using significantly fewer total iterations, suggesting the Gibbs sampler mixes better than either of the Metropolis random walk samplers.

As a final comparison, we performed additional simulations via i.i.d. sampling (not shown). The resulting empirical coverage probabilities were similar to using the Gibbs sampler, albeit with slightly fewer iterations.

Gibbs	Length (SD)	$E[X_1]$	$\xi_{.1,X_1}$	$\xi_{.9,X_1}$	$E[X_2]$	$\xi_{.1,X_2}$	$\xi_{.9,X_2}$
$T_1(0.10)$	1,930 (3.7E2)	0.941	0.940	0.937	0.954	0.958	0.927
$T_1(0.05)$	5,727 (8.7E2)	0.946	0.958	0.941	0.942	0.945	0.940
$T_1(0.02)$	31,170 (2.8E3)	0.935	0.945	0.961	0.937	0.937	0.944
$T_2(0.10)$	2,465 (5.4E2)	0.935	0.939	0.939	0.954	0.950	0.937
$T_2(0.05)$	7,865 (1.1E3)	0.950	0.959	0.943	0.955	0.954	0.952
$T_2(0.02)$	43,756 (3.6E3)	0.933	0.936	0.959	0.936	0.959	0.946
$T_3(0.10)$	1,182 (3.9E2)	0.929	0.936	0.942	0.936	0.936	0.924
$T_3(0.05)$	3,786 (6.2E2)	0.956	0.951	0.944	0.940	0.940	0.935
$T_3(0.02)$	20,289 (2.0E3)	0.945	0.947	0.954	0.940	0.943	0.952

Table 2.4: Summary of coverage probabilities for estimations of Φ using a Gibbs sampler based on 1000 replications and a 0.95 nominal level.

2.4.3 Bayesian logistic regression

Our final example considers the *Anguilla* eel data provided in the `dismo` R package (see e.g. Elith et al., 2008; Hijmans et al., 2010). The data consists of 1,000 observations from a New Zealand survey of site-level presence or absence for the short-finned eel (*Anguilla australis*). We selected six out of twelve covariates as in Leathwick et al. (2008). Five are continuous variables: SegSumT, DSDist, USNative, DSMaxSlope and DSSlope; one is a categorical variable: Method, with five levels Electric, Spo, Trap, Net and Mixture.

Let x_i be the regression vector of covariates for the ith observation of length k and $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_9)$ be the vector regression coefficients. For the ith observation, suppose $Y_i = 1$ denotes presence and $Y_i = 0$ denotes absence of *Anguilla australis*. Then the Bayesian logistic regression model is given by

$$\begin{aligned} Y_i &\sim Bernoulli(p_i) \ , \\ p_i &\sim \frac{\exp(x_i^T\boldsymbol{\beta})}{1+\exp(x_i^T\boldsymbol{\beta})} \ \text{ and,} \\ \boldsymbol{\beta} &\sim N(\mathbf{0}, \sigma_\beta^2 \mathbf{I}_k) \ , \end{aligned}$$

where $\mathbf{I}_k$ is the $k \times k$ identity matrix. For the analysis, $\sigma_\beta^2 = 100$ was chosen to represent a diffuse prior distribution on $\boldsymbol{\beta}$ (Boone et al., 2014). Further, we use the `MCMClogit` function in the `MCMCpack` package to sample from the target Markov chain.

Suppose we are interested in estimating the posterior mean along with an 80% Bayesian credible interval for each regression coefficient in the model. Given that we are working with real data, the true values are naturally unknown. Instead, we ran 1000 independent chains for 1E6 iterations to obtain an accurate estimate, which we treat as the truth (Table 2.5).

Consider estimating $\Phi_j = \left(\beta_j, \xi_{.1}^{(j)}, \xi_{.9}^{(j)}\right)$ for $j = 0, \ldots, 9$ using fixed-width stopping rules $T_i(\epsilon)$ for $i \in \{1, 2, 3\}$. From the magnitudes in Table 2.5, it is easy to see a single ϵ would be problematic for $T_1(\epsilon)$. Instead, we will specify an ϵ for each Φ_j with respect to its magnitude. Specifically, we choose three simulation settings such that $\boldsymbol{\epsilon}_1 =$ (1, 0.01, 0.001, 0.1, 0.1, 0.1, 0.1, 0.1, 0.01, 0.01), $0.5\boldsymbol{\epsilon}_1$ and $0.2\boldsymbol{\epsilon}_1$.

A single ϵ value for $T_2(\epsilon)$ will also be problematic since there are parameters with very small absolute values (e.g. DSDist). We instead specify an ϵ for each Φ_j. In this case,

Variable	β_j	$\xi_{.1}^{(j)}$	$\xi_{.9}^{(j)}$
Intercept	-10.463 (2.7E-5)	-12.224 (3.9E-4)	-8.730 (3.7E-4)
SegSumT	0.657 (1.5E-5)	0.559 (2.1E-5)	0.757 (2.2E-5)
DSDist	-4.02E-3 (3.3E-7)	-6.15E-3 (4.9E-7)	-1.93E-3 (4.4E-7)
USNative	-1.170 (7.1E-5)	-1.625 (9.9E-5)	-0.718 (1.0E-4)
MethodMixture	-0.468 (6.8E-5)	-0.910 (9.8E-5)	-0.028 (9.8E-5)
MethodNet	-1.525 (8.2E-5)	-2.026 (1.2E-4)	-1.035 (1.1E-4)
MethodSpo	-1.831 (1.3E-4)	-2.623 (2.2E-4)	-1.798 (1.4E-4)
MethodTrap	-2.594 (1.1E-4)	-3.285 (1.8E-4)	-1.937 (1.3E-4)
DSMaxSlope	-0.170 (1.1E-5)	-0.244 (1.7E-5)	-0.099 (1.5E-5)
USSlope	-0.052 (3.7E-6)	-0.076 (5.5E-6)	-0.028 (5.1E-6)

Table 2.5: Summary of estimated true values with standard errors for the Bayesian logistic regression example.

we choose three simulation settings such that $\boldsymbol{\epsilon}_2 = (0.1, 0.1, 1, 0.1, 1, 0.1, 0.1, 0.1, 0.1, 1)$, $0.5\boldsymbol{\epsilon}_2$ and $0.2\boldsymbol{\epsilon}_2$.

For both $T_1(\epsilon)$ and $T_2(\epsilon)$, it becomes overwhelmingly tedious to specify appropriate ϵ vectors when the number of parameters becomes large. However, for the stopping rule $T_3(\epsilon)$ we can use a single ϵ for the 30 dimensional target parameter vector. Specifically, we choose three simulation settings such that $\epsilon_3 \in \{0.10, 0.05, 0.02\}$.

For the two larger ϵ settings, we set $n^* = 10000$ and added 1000 iterations between checks. For the smallest ϵ setting, we set $n^* = 1E5$ and added 10000 iterations between checks due to increased computational demands. Each simulation setting was repeated 1000 times independently.

Table 2.6 summarizes the empirical coverage probabilities. We can see the coverage probabilities for each stopping rule increase towards the nominal level of 0.95 as ϵ decreases, suggesting that all the stopping rules perform well. For high dimensional settings such as this, $T_3(\epsilon)$ provides a distinct practical advantage since a practitioner can specify a single ϵ value.

To adjust for multiplicity, we again apply a Bonferonni approach. We set individual confidence intervals to have a nominal level of $0.80^{1/10} = 0.9779$ resulting in simultaneous confidence region with nominal level of at least 0.80. We only considered estimating the posterior mean of the 10 dimensional vector $\boldsymbol{\beta}$ using $T_3(\epsilon)$ with $\epsilon \in \{0.20, 0.10, 0.05, 0.02\}$. The minimum simulation effort was $n^* = 1E5$ iterations with an additional 1000 added

	$T_1(\boldsymbol{\epsilon}_1)$			$T_1(0.5\boldsymbol{\epsilon}_1)$			$T_1(0.2\boldsymbol{\epsilon}_1)$		
Variable	β_j	$\xi_{.1}^{(j)}$	$\xi_{.9}^{(j)}$	β_j	$\xi_{.1}^{(j)}$	$\xi_{.9}^{(j)}$	β_j	$\xi_{.1}^{(j)}$	$\xi_{.9}^{(j)}$
Intercept	0.936	0.933	0.912	0.937	0.942	0.942	0.946	0.946	0.930
SegSumT	0.932	0.922	0.916	0.942	0.941	0.934	0.953	0.944	0.936
DSDist	0.987	0.969	0.979	0.976	0.969	0.960	0.956	0.954	0.952
USNative	0.927	0.929	0.917	0.939	0.933	0.943	0.948	0.939	0.944
MethodMixture	0.930	0.928	0.920	0.946	0.948	0.938	0.935	0.953	0.940
MethodNet	0.946	0.922	0.936	0.941	0.948	0.932	0.943	0.939	0.935
MethodSpo	0.913	0.913	0.927	0.931	0.929	0.931	0.943	0.942	0.926
MethodTrap	0.928	0.906	0.937	0.938	0.930	0.927	0.941	0.947	0.947
DSMaxSlope	0.932	0.930	0.921	0.942	0.943	0.945	0.953	0.958	0.951
USSlope	0.921	0.928	0.935	0.951	0.927	0.954	0.957	0.952	0.962
Length (SD)	19,521 (3.8E3)			76,894 (9.5E3)			492,910 (3.4E4)		

	$T_2(\boldsymbol{\epsilon}_2)$			$T_2(0.5\boldsymbol{\epsilon}_2)$			$T_2(0.2\boldsymbol{\epsilon}_2)$		
Variable	β_j	$\xi_{.1}^{(j)}$	$\xi_{.9}^{(j)}$	β_j	$\xi_{.1}^{(j)}$	$\xi_{.9}^{(j)}$	β_j	$\xi_{.1}^{(j)}$	$\xi_{.9}^{(j)}$
Intercept	0.928	0.938	0.915	0.950	0.948	0.947	0.945	0.949	0.938
SegSumT	0.923	0.916	0.937	0.953	0.955	0.948	0.944	0.947	0.947
DSDist	0.985	0.968	0.975	0.970	0.958	0.958	0.956	0.955	0.947
USNative	0.921	0.936	0.921	0.946	0.933	0.945	0.940	0.956	0.941
MethodMixture	0.941	0.938	0.933	0.942	0.945	0.916	0.935	0.933	0.942
MethodNet	0.942	0.920	0.922	0.940	0.942	0.939	0.942	0.944	0.935
MethodSpo	0.919	0.901	0.924	0.936	0.923	0.937	0.947	0.956	0.947
MethodTrap	0.935	0.910	0.936	0.939	0.939	0.931	0.941	0.933	0.941
DSMaxSlope	0.937	0.942	0.916	0.948	0.942	0.950	0.942	0.954	0.955
USSlope	0.935	0.933	0.930	0.949	0.936	0.941	0.949	0.944	0.943
Length (SD)	37,667 (3.5E4)			151,276 (8.9E4)			1,161,400 (2.6E5)		

	$T_3(0.10)$			$T_3(0.05)$			$T_3(0.02)$		
Variable	β_j	$\xi_{.1}^{(j)}$	$\xi_{.9}^{(j)}$	β_j	$\xi_{.1}^{(j)}$	$\xi_{.9}^{(j)}$	β_j	$\xi_{.1}^{(j)}$	$\xi_{.9}^{(j)}$
Intercept	0.932	0.944	0.929	0.943	0.950	0.943	0.943	0.954	0.934
SegSumT	0.932	0.935	0.941	0.942	0.934	0.946	0.942	0.934	0.946
DSDist	0.981	0.969	0.969	0.968	0.966	0.955	0.957	0.954	0.950
USNative	0.939	0.942	0.923	0.941	0.948	0.954	0.942	0.943	0.940
MethodMixture	0.939	0.928	0.920	0.947	0.943	0.933	0.927	0.947	0.928
MethodNet	0.929	0.922	0.931	0.939	0.939	0.934	0.930	0.938	0.939
MethodSpo	0.915	0.902	0.925	0.924	0.933	0.926	0.948	0.946	0.935
MethodTrap	0.930	0.909	0.920	0.941	0.937	0.933	0.939	0.935	0.948
DSMaxSlope	0.941	0.932	0.930	0.940	0.950	0.943	0.958	0.955	0.951
USSlope	0.939	0.928	0.940	0.953	0.937	0.955	0.954	0.957	0.958
Length (SD)	24,404 (1.4E3)			78,886 (4.2E3)			439,260 (1.7E4)		

Table 2.6: Summary of coverage probabilities for Bayesian logistic regression example with 1000 independent replicates. The coverage probabilities have a 0.95 nominal level.

between checks. Again, for the smallest ϵ setting, we set $n^* = 1E6$ with an additional 10000 added between checks. The simulation was terminated the first time $T_3(\epsilon)$ was met and repeated 1000 times independently.

Table 2.7 summarizes the simulation results. We can see that, as ϵ decreases, all the individual coverage probabilities are remarkably close to the nominal level of 0.9779. Note the observed confidence region coverage probabilities approach the nominal level of 0.80 as expected. However, it is bit surprising how close this is to the nominal 0.80 level given possible correlation among parameters. To this end, we investigated the correlation between pairs of target parameters. We found that most pairs have low correlation, except for strong correlation between (Intercept, SegSumT) and moderate correlation between (USNative, USSlope). Given the lack of correlation, the confidence region coverages are very encouraging.

	$T_3(0.20)$	$T_3(0.10)$	$T_3(0.05)$	$T_3(0.02)$
Variable	β_j	β_j	β_j	β_j
Intercept	0.959	0.975	0.976	0.973
SegSumT	0.960	0.971	0.979	0.974
DSDist	0.995	0.989	0.993	0.979
USNative	0.948	0.978	0.970	0.973
MethodMixture	0.950	0.973	0.967	0.968
MethodNet	0.962	0.962	0.976	0.973
MethodSpo	0.946	0.954	0.968	0.979
MethodTrap	0.950	0.960	0.970	0.978
DSMaxSlope	0.966	0.971	0.977	0.974
USSlope	0.964	0.965	0.973	0.982
Region	0.693	0.763	0.792	0.805
Length (SD)	10,082(2.7E2)	29,729(1.8E3)	100,261(5.2E3)	583,488(1.9E4)

Table 2.7: Summary of coverage probabilities for $\boldsymbol{\beta}$ based on $T_3(\epsilon)$ with 1,000 replicates. The coverage probabilities have a 0.9779 nominal level, resulting in a 0.80 nominal level confidence region.

2.4.4 Discussion

This chapter considers absolute precision, relative magnitude, and relative standard deviation fixed-width stopping rules in the context of MCMC simulations. Under limited assumptions, we show fixed-width stopping rules obtain a desired coverage proba-

bility in an asymptotic sense as ϵ tends to 0. Moreover, we illustrate these rules perform well in a variety of finite sample settings provided ϵ is specified to be small enough.

A practical MCMC stopping rule should be applicable for a large number of parameters since practitioners usually report multiple expectation and quantile estimates. Unfortunately, choosing a single ϵ could be problematic for absolute precision and relative magnitude stopping rules. These stopping rules would be better served by specifying an $\boldsymbol{\epsilon}$ vector, which can be tedious when the number of parameters becomes large.

Instead, we advocate use of the relative standard deviation stopping rule since it is easy to implement and applicable in multivariate settings without a priori knowledge of the target parameter size. Simply put, this rule terminates an MCMC simulation when estimates of target parameters are sufficiently accurate relative to their associated posterior standard deviations. The resulting estimates are approximately ϵ^{-1} more accurate than their posterior standard deviations. We recommend using $\epsilon = 0.02$, which provided excellent results in the wide variety of examples considered here. However, a smaller ϵ may be appropriate when the accuracy of estimation is critical.

When estimating multiple quantities simultaneously, we have focused on controlling the width of each of the marginal confidence intervals. We also investigated the impact of a Bonferonni correction in the case of multiplicity. Alternatively, one could consider multiple quantities jointly by controlling the volume of confidence region, which is the subject of ongoing research. In this setting, one should be able to establish asymptotic validity for a relative fixed-volume approach using techniques presented here and in Glynn and Whitt (1992).

In any MCMC simulation, a key component is choosing a Markov chain that mixes well while sufficiently exploring the state space. As in the mixture of bivariate Normals, the sampler choice affects the performance significantly in terms of coverage probabilities. Moreover, the computational effort to achieve a reasonable accuracy varies depending on the sampling scheme. In practice, the true parameters values are unknown and thus poorly behaved samplers may lead to suspicious inference. We have offered limited guidance in this direction, but note this is usually the most challenging aspect of an MCMC simulation. An interested reader is directed to Brooks et al. (2010) and the references therein for advice on sampling schemes.

Finally, our examples only consider BM to estimate the asymptotic variance from a CLT since it is the most popular technique and widely available. Improving the variance

estimation step might be possible using alternative methods such as overlapping batch means, spectral variance, or subsampling bootstrap methods (Doss et al., 2014; Flegal, 2012; Flegal and Jones, 2010), which are currently available in the `mcmcse` package.

2.5 Proofs and Calculations

2.5.1 Proof of Theorem 4

The proof follows techniques introduced in Glynn and Whitt (1992). Define $z = z_{\delta/2}$ and recall

$$T_3(\epsilon) = \inf\left\{n \geq 0 : 2z\hat{\sigma}_n/\sqrt{n} + p(n) \leq \epsilon\hat{\lambda}_n\right\}$$

and note $T_3(\epsilon) \to \infty$ w.p.1 as $\epsilon \to 0$. The following two facts will be utilized multiple times. First, since $\hat{\sigma}_n \to \sigma_\theta$ w.p.1 as $n \to \infty$, we have $\hat{\sigma}_{T_3(\epsilon)} \to \sigma_\theta$ w.p.1 as $\epsilon \to 0$. Second, since $\hat{\lambda}_n \to \lambda_\theta$ w.p.1 as $n \to \infty$, we have $\hat{\lambda}_{T_3(\epsilon)} \to \lambda_\theta$ w.p.1 as $\epsilon \to 0$.

Define $V(n) = 2z\hat{\sigma}_n/\sqrt{n} + p(n)$, where $p(n) = o(n^{-1/2})$. Then $T_3(\epsilon)$ can be denoted as $T_3(\epsilon) = \inf\left\{n \geq 0 : V(n) \leq \epsilon\hat{\lambda}_n\right\}$. Recall $\sigma_\theta^2 \in (0, \infty)$, then it is easy to verify that

$$n^{1/2}V(n) \to 2z\sigma_\theta > 0 \text{ w.p.1 as } \; n \to \infty. \tag{2.8}$$

By definition of $T_3(\epsilon)$, $V(T_3(\epsilon) - 1) > \epsilon\hat{\lambda}_{T_3(\epsilon)-1}$ and $V(T_3(\epsilon)) \leq \epsilon\hat{\lambda}_{T_3(\epsilon)}$. Using (2.8) we have

$$\limsup_{\epsilon\to 0} \epsilon T_3(\epsilon)^{1/2} \leq \limsup_{\epsilon\to 0} T_3(\epsilon)^{1/2} V(T_3(\epsilon) - 1)/\hat{\lambda}_{T_3(\epsilon)-1} = 2z\sigma_\theta/\lambda_\theta \text{ w.p.1.}$$

Similarly,

$$\liminf_{\epsilon\to 0} \epsilon T_3(\epsilon)^{1/2} \geq \liminf_{\epsilon\to 0} T_3(\epsilon)^{1/2} V(T_3(\epsilon))/\hat{\lambda}_{T_3(\epsilon)} = 2z\sigma_\theta/\lambda_\theta \text{ w.p.1.}$$

Thus, we have

$$\lim_{\epsilon\to 0} \epsilon T_3(\epsilon)^{1/2} = 2z\sigma_\theta/\lambda_\theta \text{ w.p.1.} \tag{2.9}$$

Using (2.9) with properties of $\hat{\sigma}_{T_3(\epsilon)}$ and $\hat{\lambda}_{T_3(\epsilon)}$, we have

$$\lim_{\epsilon\to 0} \epsilon^{-1} T_3(\epsilon)^{-1/2} 2z\hat{\sigma}_{T_3(\epsilon)}/\hat{\lambda}_{T_3(\epsilon)} = 1 \text{ w.p.1.} \tag{2.10}$$

Let $\beta = 2z\sigma_\theta/\lambda_\theta$ and set $\tau_\epsilon(t) = T_3(\epsilon)\epsilon^2\beta^{-2}t$ for $t \geq 0$. Note that $\tau_\epsilon \to e$ as $\epsilon \to 0$ w.p.1 pointwise, where $e(t) = t$. Then it follows from the FCLT and a standard random-time-change argument (p. 151 Billingsley, 1999) that

$$\mathcal{Z}_{\epsilon^2\beta^{-2}}(\tau_\epsilon(1)) \xrightarrow{d} \sigma_\theta B(e(1))/e(1) = \sigma_\theta B(1) \text{ as } \epsilon \to 0\,, \tag{2.11}$$

where

$$\mathcal{Z}_{\epsilon^2\beta^{-2}}(\tau_\epsilon(1)) = \beta\epsilon^{-1}\left(Z_{T_3(\epsilon)} - \theta\right).$$

Slutsky's theorem with (2.10) and (2.11) yield

$$T_3(\epsilon)^{1/2}/\hat{\sigma}_{T_3(\epsilon)}\left(Z_{T_3(\epsilon)} - \theta\right) \xrightarrow{d} B(1) \text{ as } \epsilon \to 0.$$

Finally, we have

$$\begin{aligned}\Pr\left(\theta \in C[T_3(\epsilon)]\right) &= \Pr\left(Z_{T_3(\epsilon)} - \theta \in (-z\hat{\sigma}_{T_3(\epsilon)}/T_3(\epsilon)^{1/2},\ z\hat{\sigma}_{T_3(\epsilon)}/T_3(\epsilon)^{1/2})\right)\\ &= \Pr\left(T_3(\epsilon)^{1/2}/\hat{\sigma}_{T_3(\epsilon)}(Z_{T_3(\epsilon)} - \theta)) \in (-z,\ z)\right) \to 1 - \delta \text{ as } \epsilon \to 0.\end{aligned}$$

Chapter 3

Fixed-width Procedure in High Dimensions

This chapter proposes modifications for the relative standard deviation FWSR in high-dimensional settings, including a strongly consistent variance estimator that significantly improves computational efficiency and a novel sampling scheme that automates the adjustment of the frequency of which the stopping rule is checked with respect to the total simulation effort. It also establishes a connection between the relative standard deviation FWSR and using the ESS as a stopping rule. Two modern Bayesian applications with high-dimensionality are used to evaluate the performance of the stopping criteria. The content of this chapter is primarily contained in Gong and Flegal (2015).

3.1 Introduction

Markov chain Monte Carlo (MCMC) simulations are commonly employed in a Bayesian context to estimate features of a posterior distribution by constructing a Markov chain with the target as its stationary distribution. A fundamental challenge is determining when to terminate the simulation, especially for the often high-dimensional problems encountered in modern Bayesian analyses. For instance, the visual inspection of trace plots and running means (see e.g. Flegal and Jones, 2011) is extremely challenging in high-dimensions. Further, convergence diagnostics (see e.g. Cowles and Carlin, 1996) were designed for problems of at most moderate dimension and can be essentially impossible to implement in high-dimensions. Given these problems, most practitioners resort to a fixed-time rule to

terminate the simulation. That is, the procedure terminates after n iterations where n is determined heuristically. In this chapter, we present a simple and theoretically valid sequential stopping rule applicable for high-dimensional MCMC.

As applications, we consider the analysis of large spatially and temporally correlated data sets routinely collected by the scientific community. A common framework to effectively incorporate spatial-temporal associations is by building multiple hierarchies in the model (Banerjee et al., 2004). The posterior analysis of Bayesian hierarchical models often involves the implementation of high-dimensional MCMC. There is considerable literature in this direction, for example, Huerta et al. (2004) develop a time-varying regression model for studying ozone levels; Gelfand et al. (2005) propose spatial process modeling for dynamic data with an application to climate data; Finley et al. (2012) use Gaussian predictive processes to model large space-time data; Woolrich et al. (2004) implement a fully spatio-temporal model for the noise process in fMRI data; Smith and Fahrmeir (2007) and Lee et al. (2014) develop spatial Bayesian variable selection models to study brain images.

With important economic, ecological and public health implications, these analyses require accurate assessment of their inferential uncertainties. However, few of these studies, which often involve thousands of parameters, carefully describe the stopping criterion utilized. Among them, some use convergence diagnostics (see e.g. Gelfand et al., 2005) and some report Monte Carlo standard errors (MCSEs) to assess the quality of estimates (see e.g. Lee et al., 2014). We assume the rest employ a fixed-time stopping rule where n is determined heuristically. Unfortunately, choosing too small an n can lead to inaccurate statistical inference.

As mentioned, many practitioners utilize convergence diagnostics and visual inspections to evaluate if the chain has been run long enough. While these methods can be useful to assess sampler performance and detect obvious multimodality, they are barely tenable in truly high-dimensional settings. For example, as stated in Gössl et al. (2001), "With this high-dimensional data, convergence diagnostics were reduced to a selection of randomly chosen parameter chains".

Instead, we advocate terminating the simulation using a relative standard deviation fixed-width stopping rule (FWSR), which is easy to implement and theoretically justified (Flegal and Gong, 2015). The main idea is to stop the simulation when an estimate is sufficiently accurate relative to its posterior uncertainty. That is, the simulation is terminated the first time a confidence interval width is less than an ϵth fraction of the

posterior standard deviation. In this chapter, we show such a stopping rule is equivalent to stopping when an effective sample size (ESS) is sufficiently large. In addition, we show relative FWSRs work well in truly high-dimensional problems since a single ϵ can be used for multiple parameters without any a priori knowledge.

Use of the relative standard deviation FWSR in high-dimensional settings requires overcoming some computational issues, which we address here. The proposed computational modifications provide significant improvements with minor tradeoffs. As we show later, the main benefits are a significant reduction in computer memory usage and improved computational efficiency. To our best knowledge, there are no previous attempts to formally address how long to run a MCMC simulation in such high-dimensional settings. Specifically, we extend the previous application (Flegal and Gong, 2015) of the stopping rule for estimating tens of parameters to a spatial Bayesian dynamic model with hundreds of parameters and a more complicated Bayesian fMRI model with thousands of parameters. Finally, we compare our results to a convergence diagnostic used as a stopping criterion and show the latter tends to terminate the simulations prematurely.

The two distinct high-dimensional Bayesian hierarchical analyses considered here are (i) the univariate dynamic space-time regression models introduced by Gelfand et al. (2005) applied to weather station data collected over the northeastern United States (Finley et al., 2012) and (ii) the spatial variable selection models proposed by Lee et al. (2014) applied to the StarPlus fMRI datasets (Carpenter et al., 1999; Keller et al., 2001; Wang and Mitchell, 2002). Both applications clearly demonstrate the potential of the relative standard deviation FWSR in general high-dimensional settings. Moreover, they illustrate the rule is easily implemented in an almost automated fashion while providing uncertainty estimates with confidence.

The rest of the chapter is organized as follows. Section 3.2 formally introduces the relative standard deviation FWSR, proposes modifications for modern applications, and illustrates its connection to ESS. Section 3.3 investigates finite sample properties for two high-dimensional MCMC simulations related to Bayesian hierarchical models. This section summarizes the models, experimental datasets and simulation studies. Section 3.4 concludes with a discussion and recommendations for practitioners.

3.2 A sequential stopping procedure

Suppose we want to make inference about a probability distribution π with support $\mathsf{X} \subseteq \mathbb{R}^d$, $d \geq 1$. Such inference is often based on expectations with respect to π. To this end, our goal is to calculate $\mathbb{E}_\pi \boldsymbol{g} = (\mathbb{E}_\pi g_1, \cdots, \mathbb{E}_\pi g_p)^T \in \mathbb{R}^p$, where, for $i = 1, \cdots, p$, $g_i : \mathsf{X} \to \mathbb{R}$ and

$$\mathbb{E}_\pi g_i = \int_{\mathsf{X}} g_i(x)\pi(dx).$$

Note that p can be smaller or larger than d, with large values of either indicating a high-dimensional setting.

Unfortunately, in most practical settings we cannot calculate $\mathbb{E}_\pi \boldsymbol{g}$ analytically and frequently π is such that MCMC is the only viable technique for estimating $\mathbb{E}_\pi \boldsymbol{g}$. MCMC methods entail constructing a time-homogeneous Harris ergodic Markov chain $X = \left\{X^{(0)}, X^{(1)}, \ldots\right\}$ on state space X with invariant distribution π (Robert and Casella, 2004).

Suppose n is finite and we simulate X for n steps. Let

$$\bar{\boldsymbol{g}}(n) := \frac{1}{n}\sum_{j=0}^{n-1} \boldsymbol{g}\left(X^{(j)}\right) = (\bar{g}_1(n), \ldots, \bar{g}_p(n))^T$$

be an estimator of $\mathbb{E}_\pi \boldsymbol{g}$ from the observed chain. Under certain regularity conditions (Chan and Geyer, 1994; Jones, 2004; Roberts and Rosenthal, 2004; Tierney, 1994), we can obtain a marginal Markov chain central limit theorem (CLT) for the sampling distribution of an unknown MCSE. That is for $i = 1, \ldots, p$,

$$\sqrt{n}\left(\bar{g}_i(n) - \mathbb{E}_\pi g_i\right) \xrightarrow{d} \mathrm{N}\left(0, \sigma_i^2\right) \tag{3.1}$$

as $n \to \infty$ where $\sigma_i^2 \in (0, \infty)$. One could also consider a multivariate Markov chain CLT for $\bar{\boldsymbol{g}}(n) - \mathbb{E}_\pi \boldsymbol{g}$. However, the often high-dimensionality of the associated asymptotic covariance matrix creates additional challenges and extracting useful information from it is a direction of future research.

For $i = 1, \ldots, p$, let $\hat{\sigma}_i(n)$ denote an estimator of σ_i. Then the CLT at (3.1) allows construction of p $(1-\delta)100\%$ marginal confidence intervals with widths

$$w_i(n, \delta) = 2z_{\delta/2}\frac{\hat{\sigma}_i(n)}{\sqrt{n}} \quad \text{for } i = 1, \ldots, p\,, \tag{3.2}$$

where $z_{\delta/2}$ is a critical value from the standard Normal distribution. We can use the widths at (3.2) to construct sequential FWSRs that terminate the simulation when they fall below specific values.

3.2.1 A relative fixed-width stopping rule

Suppose ϵ is a pre-specified value, then the simplest FWSR terminates the simulation when $w_i < \epsilon$ for all $i = 1, \ldots, p$. Asymptotic validity of such a rule was established by Glynn and Whitt (1992) and first used in MCMC simulations by Jones et al. (2006). Asymptotic validity is important because it ensures the simulation will terminate w.p.1 and the resulting confidence intervals will have the right coverage probability (as $\epsilon \to 0$).

Jones et al. (2006) and Flegal et al. (2008) show the simple FWSR is superior to using convergence diagnostics as stopping criteria. Unfortunately, such a rule is difficult to implement in high-dimensional settings without a priori knowledge of the magnitudes of the components in $\mathbb{E}_\pi \boldsymbol{g}$. Further, a single ϵ value is unlikely to be suitable across multiple dimensions.

Instead, we consider a relative standard deviation FWSR proposed by Flegal and Gong (2015). The main idea is to terminate the simulation when an estimator's computational uncertainty is small relative to its posterior uncertainty. As we will illustrate, this is equivalent to terminating the simulation when the ESS is sufficiently large.

To this end, let λ_i^2 denote the posterior variance associated with $\mathbb{E}_\pi g_i$. That is, λ_i^2 is the i-th diagonal element of $Var_\pi[\boldsymbol{g}(X)]$. Due to correlation in the Markov chain, $\sigma_i^2 \neq \lambda_i^2$ in general. We further suppose $\hat{\lambda}_i^2(n)$ is an estimator of λ_i^2, usually

$$(\hat{\lambda}_1^2(n), \ldots, \hat{\lambda}_p^2(n))^T = \frac{1}{n-1} \sum_{j=0}^{n-1} \left(\boldsymbol{g}\left(X^{(j)}\right) - \bar{\boldsymbol{g}}(n) \right)^2 .$$

Note that exponentiation on a vector is taken element-wise.

A relative standard deviation FWSR terminates the simulation when the length of all the confidence intervals are less than an ϵth fraction of the magnitude of their posterior standard deviations. That is, when $w_i < \epsilon \hat{\lambda}_i$ for all $i = 1, \ldots, p$. Formally, the time at which the simulation terminates is defined by

$$T(\epsilon, \delta) = \sup_{i \in \{1,\ldots,p\}} \inf \left\{ n \geq 0 : w_i(n, \delta) + p(n) \leq \epsilon \hat{\lambda}_i(n) \right\} ,$$

where $p(n) \geq 0$. The role of $p(n)$ is to ensure that the simulation is not terminated prematurely based on poor estimates of the σ_i^2s. A reasonable default is $p(n) = \epsilon I(n \leq n^*) + n^{-1}$ (Glynn and Whitt, 1992; Jones et al., 2006), where n^* is the desired minimum simulation effort. The user-specified starting value n^* is often based on the complexity of the problem at hand and ϵ reflects the desired accuracy for the analytical purpose. In our experience, setting $n^* = 1E4$ works well in practice. However, we caution that one should also examine trace plots, autocorrelation plots, and convergence diagnostics to determine the minimum simulation effort. Such care should also be used to determine if the MCMC sampler itself is performing well, see e.g. Flegal and Jones (2011).

Sufficient conditions for asymptotic validity of $T(\epsilon, \delta)$ are established in Flegal and Gong (2015). In short, they require the limiting process must satisfy a functional CLT and estimators of the associated asymptotic variance and posterior variance must be strongly consistent. While not trivial, one can establish these conditions in many complex practical MCMC settings. An interested reader is directed to Flegal and Gong (2015), Flegal and Jones (2010), and Jones et al. (2006).

The relative standard deviation FWSR (outlined in Algorithm 1) is appealing because it provides a simple, yet informative automated stopping criterion applicable in multivariate settings. One only needs to specify a relative ϵ and hence no prior knowledge about the magnitude of the parameters is needed. Moreover, a single ϵ will suffice in multivariate settings, whereas other fixed-width approaches require a vector of values.

Algorithm 1 Relative standard deviation FWSR

Require: $0 < \epsilon, \delta < 1$ and $n^*, m > 0$ ▷ m: # of iterations between checks

1: UPDATE the chain by n^* iterations
2: **while** `True` **do**
3: ESTIMATE $\hat{\sigma}$ and $\hat{\lambda}$
4: **if** stopping criterion is met **then**
5: **break**
6: UPDATE the chain by m iterations

The frequency with which the criterion should be checked is still an open question. Checking too often may substantially increase the computational burden. Instead, it is sufficient to check every m iterations, where m is a pre-specified gap determined by an estimated simulation effort.

Variance estimation modification

The MCMC community has expended considerable effort establishing strongly consistent estimators for the asymptotic variance at (3.1) including batch means (Flegal and Jones, 2010; Jones et al., 2006), spectral variance estimation (Flegal and Jones, 2010) and regenerative simulation (Hobert et al., 2002; Mykland et al., 1995). In this section, we propose a modified non-overlapping batch means (BM) estimator that does not require storage of the entire chain.

In standard BM the output is broken into a_n batches of equal size b_n. Suppose the algorithm is run for a total of $n = a_n b_n$ iterations and define for $j = 1, \ldots, a_n$,

$$\boldsymbol{Y}_j = \frac{1}{b_n} \sum_{k=(j-1)b_n}^{jb_n-1} \boldsymbol{g}\left(X^{(k)}\right) \ .$$

The BM estimate of the asymptotic variance from the CLT at (3.1) is

$$\left(\hat{\sigma}_1^2(n), \ldots, \hat{\sigma}_p^2(n)\right)^T = \frac{b_n}{a_n - 1} \sum_{j=1}^{a_n} \left(\boldsymbol{Y}_j - \bar{\boldsymbol{g}}(n)\right)^2 \ .$$

Jones et al. (2006) establish necessary conditions for $\hat{\sigma}_i^2(n) \to \sigma_i^2$ w.p.1, $i = 1, \cdots, p$, as $n \to \infty$. In short, they require the batch size and the number of batches to increase as the overall simulation length increases. Setting $b_n = \lfloor n^\tau \rfloor$ and $a_n = \lfloor n/b_n \rfloor$, the regularity conditions require that X be geometrically ergodic, $E_\pi |g|^{2+\epsilon_1+\epsilon_2} < \infty$ for some $\epsilon_1 > 0$, $\epsilon_2 > 0$ and $(1 + \epsilon_1/2)^{-1} < \tau < 1$. A common choice of $\tau = 1/2$ has been shown to work well in applications (Flegal et al., 2008; Jones et al., 2006). We denote the BM estimate with such a sampling plan as the consistent batch means (CBM) estimate.

Unfortunately, most sampling plans including CBM require storage of the entire Markov chain to allow recalculations as b_n grows with n. Given a target vector of dimension p, this means a matrix of size $p \times n$ will have to be stored in the memory. Clearly, computer memory soon becomes a serious limitation, which one can solve by writing parts of the chain in-and-out of memory. However, given the frequency that $T(\epsilon, \delta)$ is checked and the already computationally intense task of updating the chain, we prefer a simpler solution.

To this end, we propose a new sampling plan that utilizes less memory while still providing a strongly consistent variance estimator. Specifically, set $\tilde{b}_n = \inf\left\{2^k : 2^k \geq n^\tau, k \in \mathbb{Z}^+\right\}$ and $\tilde{a}_n = \lfloor n/\tilde{b}_n \rfloor$. Notice $\tilde{b}_n$ is bounded by $n^\tau \leq \tilde{b}_n \leq 2n^\tau$. Hence, it is easy to establish

strong consistency for $\hat{\sigma}_i^2(n)$ with such a sampling plan using results in Jones et al. (2006) and Bednorz and Latuszyński (2007). We denote this BM estimate with $\tilde{b}_n$ as the low-cost batch means (LCBM) estimate.

Notice that $\tilde{b}_n$ increases by doubling the batch size, i.e. in the form of $\{2, 4, 8, \ldots, 2^k, \ldots\}$. It then becomes possible to record only the batch means $\boldsymbol{Y}_j$s and merge every two batches by averaging their means when the batch size increases twofold. The size of the required storage then reduces significantly from $O(n)$ to $O(\tilde{a}_n) = O(n^{1-\tau})$. Moreover, calculations at each checking point take less time since the batch means are already in memory. In practice, this change significantly reduces computational effort and memory as we illustrate later. Finally, using the new sampling plan with $T(\epsilon, \delta)$ requires a standard recursive calculation of $\hat{\lambda}_i(n)$ as n increases. An interested reader is directed to the technique studied by Biesel (1977).

Use of the proposed sampling plan only requires storage of the current state and the batch means. The unit of interest is then per batch rather than per iteration. Thus, a natural adjustment to the frequency with which $T(\epsilon, \delta)$ should be checked is to examine the criterion every m batches. As before, m is pre-specified by the user but is likely much smaller than used previously. The gap between checks is then m batches, or equivalently $m\tilde{b}_n$ iterations. Hence, the number of iterations between each check increases in accordance to the magnitude of the simulation effort. Note that occasionally an additional batch is needed between checks to ensure there are an even number of batches. Such variation enables adjacent batch means to be merged when the batch size increases twofold.

The two proposed modifications fit naturally with each other and enable implementation of relative standard deviation FWSR in high-dimensional settings. The modified procedure with LCBM calculation is presented in Algorithm 2. Note that it only requires the in-memory storage of the batch means. In addition, they yield improvements in computational efficiency measured by clock time and stopping procedure automation. A drawback of the new sampling plan is one can not consider estimation problems that require storing the entire chain. One way to circumvent this is to periodically write (out of memory) a copy of the entire chain before only storing the means (in memory).

Remark 8. *A lower bound LCBM is defined by setting* $\tilde{b}_n^* = \sup\{2^k : 2^k \leq n^\tau, k \in \mathbb{Z}^+\}$ *and* $\tilde{a}_n^* = \lfloor n/\tilde{b}_n^* \rfloor$. *One can establish strong consistency for such a sampling plan because* $\tilde{b}_n^*$ *is then bounded by* $n^\tau/2 \leq \tilde{b}_n^* \leq n^\tau$. *We advocate the upper bound LCBM since it produces*

Algorithm 2 Relative standard deviation FWSR with LCBM

Require: $0 < \epsilon, \delta < 1$, n^* and $m > 0$ ▷ m: # of batches between checks

1: CALCULATE *batch_size* based on n^*
2: UPDATE the chain to get $\lceil n^*/batch\_size \rceil$ batches of iterations
3: INITIALIZE $counter = \lceil n^*/batch\_size \rceil \times batch\_size$
▷ Note that *counter* ($\geq n^*$) keeps track of the total number of iterations
4: UPDATE $\hat{\lambda}$ using recursive techniques
5: STORE the mean of each of the $\lceil n^*/batch\_size \rceil$ batches to a container: *batch_means*
6: **while** `True` **do**
7: ESTIMATE $\hat{\sigma}$ from *batch_means*
8: **if** stopping criterion is met **then**
9: **break**
10: UPDATE the chain to get m (or $m+1$) batches of iterations
▷ Variation is to ensure the feasibility of Line 15
11: UPDATE $\hat{\lambda}$ using recursive techniques
12: APPEND the mean of each of the new batches to *batch_means*
13: INCREMENT *counter* by $m \times batch\_size$ (or $(m+1) \times batch\_size$)
14: **if** *batch_size* changes based on the updated *counter* **then**
15: RESHAPE *batch_means* by averaging neighbors

more conservative estimates and thus better performances in terminating simulations (see Section 3.3 for numerical comparisons).

3.2.2 Connections with effective sample size

Given n iterations in a Markov chain, the ESS measures the size of an i.i.d. sample with the same standard error, or the "effective number of independent samples". This quantity is frequently used by practitioners as a run length diagnostic, terminating the simulation once ESS estimates are greater than a pre-specified threshold K (for e.g. see Atkinson et al., 2008; Drummond et al., 2006). Although the intuition behind this rule is clear, we are not aware of any theoretical discussions of its validity. Here we show the relative standard deviation FWSR and the ESS stopping rule are equivalent. Thus, we establish theoretical validity of using the ESS as a stopping rule for MCMC simulations. The main assumption required is a strongly consistent estimate of the ESS. The connection also provides practitioners with another intuitive way to look at the relative standard deviation FWSR.

Note that the ESS is not uniquely defined. One way to define ESS is described in

Kass et al. (1998) and Robert and Casella (2004), for $i = 1, \cdots, p$,

$$\mathrm{ESS}_i(n) = \frac{n}{1 + 2\sum_{k=1}^{\infty} \rho_k(g_i)},$$

where $\rho_k(g_i)$ is the autocorrelation of lag k for g_i. This calculation is implemented in many R packages, such as `coda` (Best et al., 1995) and `mcmcse` (Flegal and Hughes, 2012).

An alternative approach to define ESS as in the custom of survey sampling (Kish, 1965; Liu et al., 1998), where for $i = 1, \cdots, p$,

$$\mathrm{ESS}_i(n) = \frac{n}{\sigma_i^2/\lambda_i^2}.$$

In practice, we estimate this quantity by replacing the parameters with their strongly consistent estimates, i.e.

$$\widehat{\mathrm{ESS}}_i(n) = \frac{n}{\hat{\sigma}_i^2(n)/\hat{\lambda}_i^2(n)}. \tag{3.3}$$

The two ESS calculations produce comparable results in various simulation studies. As a toy example we consider an independence Metropolis sampler with an EXP(0.5) proposal to sample from an EXP(1) target distribution. Note that this Markov chain is uniformly ergodic. Figure 3.1 shows that these methods behave similarly as the number of iterations increases, although the alternative ESS calculation based on either CBM or LCBM tends to fluctuate more due to changes in batch size (with $\tau = 1/2$). Notice that LCBM produces slightly more stable estimates than CBM since its batch size changes less frequently. Despite that, the alternative ESS calculations, especially implemented using LCBM, enable the estimation of ESS in memory intensive and high-dimensional settings. Further, in the presence of high correlations, the alternative ESS calculations tend to be more conservative in the sense that they produce smaller ESS estimates.

In multivariate settings, using ESS as a stopping rule is equivalent to terminating the simulation when the estimated ESS for every parameter is above the threshold K. That is, the time at which the simulation terminates is defined by

$$\tilde{T}(K) = \sup_{i \in \{1,\ldots,p\}} \inf \left\{ n \geq 0 : \widehat{\mathrm{ESS}}_i(n) \geq K \right\} .$$

As mentioned, when implemented together with the alternative ESS calculation,

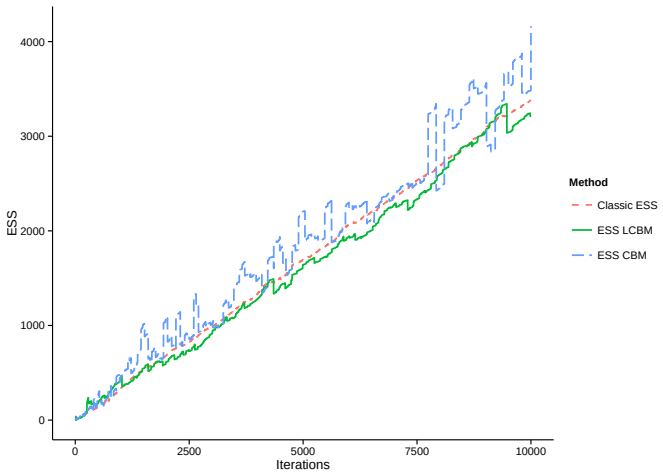

Figure 3.1: Comparison of ESS estimates for an independence Metropolis sampler with a EXP(0.5) proposal used to sample from an EXP(1) target.

$\tilde{T}(K)$ provides additional insights into the relative standard deviation FWSR. From the definition of $T(\epsilon, \delta)$, one can easily show that at termination

$$\epsilon\hat{\lambda}_i(n) \geq 2z_{\delta/2}\hat{\sigma}_i(n)/\sqrt{n} + p(n) \approx 2z_{\delta/2}\hat{\sigma}_i(n)/\sqrt{n}. \tag{3.4}$$

Combining (3.3), (3.4) and the definition of $\tilde{T}(K)$, setting $K = 4z^2_{\delta/2}/\epsilon^2$, we have $T(\epsilon, \delta) \approx \tilde{T}(K)$. That is, the relative standard deviation FWSR is equivalent to terminating a simulation when the smallest ESS is above a pre-specified level. For instance, setting $\epsilon = 0.124$ and $\delta = 0.05$ in FWSR is equivalent to set $K = 1000$ in ESS. Given the equivalency, we can establish the asymptotic validity of the ESS stopping rule under the same conditions for the relative standard deviation FWSR (see Flegal and Gong, 2015) provided the threshold K goes to infinity. This equivalency is valid only under the alternative ESS calculation since a strongly consistent estimate of ESS is required.

3.2.3 An alternative stopping criterion

Convergence diagnostics are widely employed by practitioners as stopping criteria. Particularly, we are interested in the Geweke diagnostic (GD) from Geweke (1992), which we will compare with the relative standard deviation FWSR in the next section. Our simulations use the GD implementation from the R package `coda` (Best et al., 1995). The GD is based on a hypothesis test that the mean estimates of two non-overlapping parts of

the Markov chain have converged. As a rule of thumb, Geweke (1992) suggested to take first 0.1 and last 0.5 proportions of the Markov chain. The resulting test statistic is univariate by its nature and the z-score is constructed as follows,

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\hat{s}_1(0)/n_1 + \hat{s}_2(0)/n_2}},$$

where $\bar{x}_1, \bar{x}_2$ are the sample average and $\hat{s}_1(0), \hat{s}_2(0)$ are spectral density estimates at zero frequency for the two parts of the Markov chain, respectively. In multivariate settings, given the hypothesis-testing nature of GD, one needs to confirm if the percentage of p-values below a pre-specified threshold δ (e.g. 0.1 or 0.05) at termination is greater than $1 - \delta$.

The GD requires a single Markov chain, which is close in spirit to the current work. It is also more practical in high-dimensional settings than the popular Gelman-Rubin diagnostic (Brooks and Gelman, 1998; Gelman and Rubin, 1992), which requires parallel chains. Jones et al. (2006) note that the GD is based on a Markov chain CLT and hence does not apply more generally than a FWSR that is based on the calculation of MCSE.

3.3 Applications

In this section, we evaluate the performance of relative standard deviation FWSR in finite sample settings using spatial-temporal Bayesian applications. Particularly, we consider the spatial Bayesian dynamic models of Gelfand et al. (2005) applied to a weather station dataset of Finley et al. (2012) and the spatial Bayesian variable selection models of Lee et al. (2014) applied to an experimental fMRI dataset of Carpenter et al. (1999).

3.3.1 Bayesian dynamic space-time model

This application considers the monthly temperature data collected over 356 weather stations in the northeastern United States starting in January 2000 to September 2011, which is available in the R package `spBayes` (Finley and Banerjee, 2013). We fit the univariate Bayesian dynamic space-time regression model proposed by Gelfand et al. (2005) to a subset of the dataset for illustrative purposes. Note that the modeling approach is limited to settings where space is continuous but time is taken to be discrete.

The response $y_t(s)$ denotes the temperature at location s and time t. It is modeled through a measurement equation that provides a regression specification with a space-time varying intercept. In addition, the model considers serially and spatially uncorrelated zero-centered Gaussian disturbances as measurement error $\epsilon_t(s)$. A transition equation introduces a $p \times 1$ coefficient vector $\boldsymbol{\beta}_t$, which is a strictly temporal component, and a spatial-temporal component $u_t(s)$. The overall model is given by

$$y_t(s) = \boldsymbol{x}_t(s)^T \boldsymbol{\beta}_t + u_t(s) + \epsilon_t(s),\ t = 1, 2, \ldots, N_t,$$

$$\epsilon_t \sim N(0, \tau_t^2),$$

$$\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} + \boldsymbol{\eta}_t;\ \boldsymbol{\eta}_t \sim N_p(0, \Sigma_\eta),$$

$$u_t(s) = u_{t-1}(s) + w_t(s);\ w_t(s) \sim GP(0, C_t(\cdot, \psi_t)).$$

The $GP(0, C_t(\cdot, \psi_t))$ denotes a spatial Gaussian process with covariance function $C_t(\cdot; \psi_t)$. We specify $C(s_1, s_2; \psi_t) = \sigma_t^2 \rho(s_1, s_2; \phi_t)$, where $\psi_t = \{\sigma_t^2, \phi_t\}$ and $\rho(\cdot; \phi)$ is an exponential correlation function with ϕ controlling the correlation decay and σ_t^2 represents the spatial variance component.

The prior specifications and MCMC schemes follow the `spDynLM` function in the `spBayes` package and we use it to sample from the Markov chain. Interested readers are directed to Finley and Banerjee (2013) for details. Specifically, we are only interested in the data recorded from 10 weather stations in the year 2000 and estimating the posterior mean of $\{\tau_t^2, \sigma_t^2, \phi_t, \boldsymbol{\beta}_t, \Sigma_\eta, u_t(s)\}$ with $p = 186$ dimensions.

Two measurements to evaluate the performance of the stopping criteria are utilized. One is the average of coverage probabilities over all 186 parameters. The other is

$$\max_{i \leq p} \mathbb{P}\left(|\bar{g}_i(n) - \mathbb{E}_\pi g_i| \geq \epsilon \lambda_i\right), \tag{3.5}$$

which should be less than δ when n is chosen according to $T(\epsilon, \delta)$. Since the true values are unknown, we use estimates of g_i and λ_i, $i = 1, \cdots, p$, obtained via 1000 parallel runs of 1E6 iterations and treated these as the "truth".

Terminating the simulation

We terminated 1000 parallel simulations and conducted comparative studies of a set of stopping criteria. Specifically, for each independent run, the ESS stopping rule was implemented with three batch means estimates (CBM, LCBM and LCBM*) and three threshold values (1000, 2000 and 4000). Note that LCBM* stands for the lower bound LCBM. Using the same data, the relative standard deviation FWSR was implemented with the same batch means estimates (CBM, LCBM and LCBM*) and three ϵ values (0.123, 0.088 and 0.062) with $\delta = 0.05$. Under such settings, the relative standard deviation FWSR should be equivalent to the ESS stopping rule. Since the proposed sampling plan leads to the batch size of the form 2^k, $k \in \mathbb{Z}^+$, we set $n^* = 2^{14} = 16,384$ and added 20 or 21 batches between checks. Two values of added batches were to ensure an even number of batches when the stopping criteria are checked. As a comparison, we used the GD as a stopping rule with a threshold p-value of 0.05 starting with 15000 iterations after 5000 burn-in.

Criteria	Estimator	Threshold	Length(SD)	Memory(SD)	Coverage	Maximum
ESS	LCBM*	1000	$1.73E5(4.01E3)$	1.01(0.02)	0.822	0.018
	CBM	1000	$2.70E5(1.20E4)$	402.32(17.73)	0.889	0.004
	LCBM	1000	$3.57E5(3.85E4)$	0.56(0.07)	0.925	0.001
	LCBM*	2000	$5.34E5(1.26E4)$	1.55(0.04)	0.897	0.005
	CBM	2000	$6.60E5(2.51E4)$	982.22(37.42)	0.916	0.002
	LCBM	2000	$7.20E5(2.61E4)$	1.05(0.04)	0.925	0.001
	LCBM*	4000	$1.36E6(1.64E5)$	2.27(0.37)	0.919	0.001
	CBM	4000	$1.50E6(5.01E4)$	2231.75(74.59)	0.931	0.001
	LCBM	4000	$1.68E6(6.60E4)$	1.22(0.05)	0.941	0.001
FWSR	LCBM*	0.124	$1.73E5(4.01E3)$	1.01(0.02)	0.822	0.018
	CBM	0.124	$2.70E5(1.20E4)$	402.32(17.73)	0.889	0.004
	LCBM	0.124	$3.57E5(3.85E4)$	0.56(0.07)	0.925	0.001
	LCBM*	0.088	$5.34E5(1.26E4)$	1.55(0.04)	0.897	0.005
	CBM	0.088	$6.60E5(2.51E4)$	982.22(37.42)	0.916	0.002
	LCBM	0.088	$7.20E5(2.61E4)$	1.05(0.04)	0.925	0.001
	LCBM*	0.062	$1.36E6(1.64E5)$	2.27(0.37)	0.919	0.001
	CBM	0.062	$1.50E6(5.01E4)$	2231.75(74.59)	0.931	0.001
	LCBM	0.062	$1.68E6(6.60E4)$	1.22(0.05)	0.941	0.001
GD	–	0.05	$1.50E4(0)$	22.32(0)	0.720	–

Table 3.1: Summary statistics for three stopping criteria based on 1000 independent replications and 0.95 nominal level. Memory usage is measured in megabytes.

Table 3.1 summarizes comparative statistics for the three stopping criteria utilized. Notice that results from the relative standard deviation FWSR and the ESS stopping rule are almost identical. We will therefore limit our discussion to only the ESS stopping criteria. As the threshold K increases, both the coverage probabilities and the maximum probabilities in (3.5) improve. For three batch means estimates, all the coverage probabilities are close to the 0.95 nominal level and the maximum probabilities are well below 0.05 indicating the ESS is performing well under these threshold values. However, LCBM, with significantly less computer memory usage, achieves slightly better coverage probabilities than CBM, which is a major advantage in high-dimensional settings. LCBM also outperforms LCBM* in these settings. The better performances of LCBM is due to its uniformly larger batch size, i.e. $\tilde{b}_n^* \leq b_n \leq \tilde{b}_n$. In short, the relative standard deviation FWSR and the ESS stopping rule are equivalent and all perform well in terminating the simulations. The proposed modification to the batch means estimator reduces memory usage while maintaining overall performance. On the contrary, the GD as a stopping rule produces poor coverage probabilities with far less simulation effort at termination. Its results indicate a premature termination as pointed out by Cowles and Carlin (1996).

3.3.2 Spatial Bayesian variable selection model

This application considers the Bayesian analysis of a functional Magnetic Resonance Imaging (fMRI) study. It studies the physiological changes that accompany brain activation via the blood oxygenation level dependent (BOLD) signal contrast. During the course of a typical fMRI experiment, a single patient performs a set of tasks in response to one or several external stimulus while a series of three dimensional brain images are acquired. Our goal is to detect activated brain regions associated with external stimulus through the image intensities. Imagine that the patient's brain can be divided into tiny voxels on a 3D regular lattice. The time series BOLD response is collected at each voxel resulting in enormous observations of spatio-temporally correlated structures. The Bayesian analysis of fMRI data often involves high-dimensional models and extensive computation.

For voxel $v = 1, \ldots, N$, let $\{y_{v,i}; i = 1, \ldots, t\}$ be the BOLD image intensities at t time points. Although other alternatives are possible, a conventional voxelwise regression analysis assumes a linear model with a balance between model complexity and computa-

tional feasibility (Friston et al., 1995; Smith and Fahrmeir, 2007),

$$y_{v,i} = z_i^T a_v + x_{v,i}\beta_v + \epsilon_{v,i}.$$

Linear combination $z_i^T a_v$ is the baseline trend to remove stimulus-independent effects. β_v is the activation amplitude and $x_{v,i}$ is the transformed stimulus (see Figure 3.2). In many experiments, the external stimulus $\{s_i; i = 1, \ldots, t\}$ alternates activation/inactivation in a 0-1 'boxcar' pattern. However, instead of proceeding in a 0-1 'boxcar' function, the brain produces a fairly fixed, stereotyped blood flow response with delay d_v every time a stimulus hits it, where d_v is estimated in a preprocessing step. The so-called hemodynamic response function (HRF) is used to characterizes this process. There are several formulations of HRF (see e.g. Friston et al., 1998; Glover, 1999; Gössl et al., 2001).

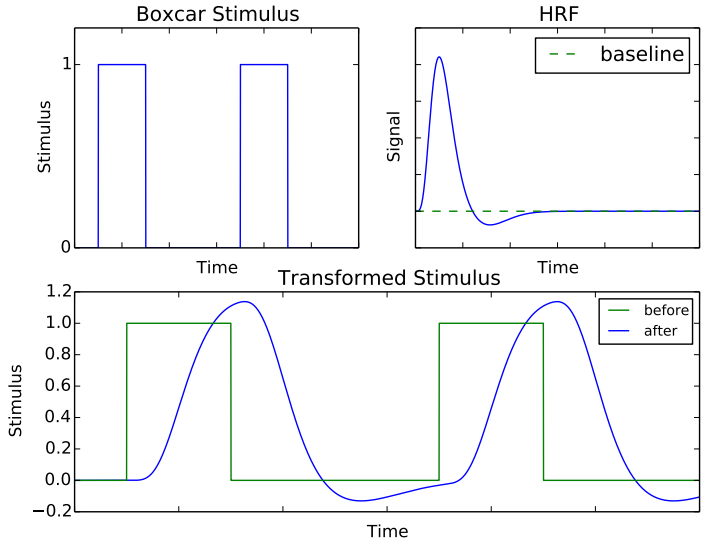

Figure 3.2: The transformed stimulus is obtained by convolving the original 0-1 'boxcar' stimulus and the HRF.

One approach is to use a canonical HRF consisting of a difference of two gamma functions (Lindquist et al., 2009),

$$h(t) = A\left(\frac{t^{\alpha_1-1}\beta_1^{\alpha_1}e^{-\beta_1 t}}{\Gamma(\alpha_1)} - c\frac{t^{\alpha_2-1}\beta_2^{\alpha_2}e^{-\beta_2 t}}{\Gamma(\alpha_2)}\right),$$

where $\alpha_1 = 6$, $\alpha_2 = 16$, $\beta_1 = \beta_2 = 1$ and $c = 1/6$. The only unknown parameter, i.e. the amplitude A, is estimated in a preprocessing step. We can transform the orignal 'boxcar' stimulus by a convolution with the HRF,

$$x_{v,i} = \sum_{k=0}^{i-d_v} h(k) s_{i-d_v-k}.$$

The measurement error is denoted by $\epsilon_{v,i}$. Appropriate distributional assumptions about $\epsilon_{v,i}$ can be made to incorporate temporal correlation and specific priors can be chosen to reflect spatial dependence.

In this chapter, we consider the spatial Bayesian variable selection models for single subject (Lee et al., 2014). This approach is shown to incorporate temporal-spatial correlation and allow for the task-related changes in BOLD response while mitigates the computational burden. It also possesses the ability to account for anatomic prior information. A general MCMC algorithm is designed to perform the large dimensional posterior inference. Here we summarize the model formulation and estimation process from Lee et al. (2014). An interested reader is directed to their paper for more details.

Denote $\boldsymbol{y}_v = (y_{v,1}, \ldots, y_{v,t})^T$ as the BOLD image intensity at time $i = 1, \ldots, t$ for voxel $v = 1, \ldots, N$. Let X_v be a $t \times p$ design matrix of transformed stimulus and $\boldsymbol{\beta}_v = (\beta_{v,1}, \ldots, \beta_{v,p})^T$ be a vector of p regression coefficients for each voxel. We formulate a linear regression mode,

$$\boldsymbol{y}_v = X_v \boldsymbol{\beta}_v + \boldsymbol{\epsilon}_v, \quad \boldsymbol{\epsilon}_v \sim N_t\left(\mathbf{0}, \sigma_v^2 \Lambda_v\right). \tag{3.6}$$

Notice that the detection of voxel activation is equivalent to the identification of nonzero $\boldsymbol{\beta}_v$s. To this end, we introduce 0/1 binary indicators $\boldsymbol{\gamma}_v = (\gamma_{v,1}, \ldots, \gamma_{v,p})$, $v = 1, \ldots, N$, such that $\beta_{v,j} = 0$ if $\gamma_{v,j} = 0$ and $\beta_{v,j} \neq 0$ if $\gamma_{v,j} = 1$. The $\gamma_{v,j}$ is used to indicate whether the voxel v is activated by input stimulus j. Given $\boldsymbol{\gamma}_v$, let $\boldsymbol{\beta}_v(\boldsymbol{\gamma}_v)$ be the vector of nonzero regression coefficients and $X_v(\boldsymbol{\gamma}_v)$ be the corresponding design matrix. Then, the model (3.6) can be rewritten as

$$\boldsymbol{y}_v = X_v(\boldsymbol{\gamma}_v)\, \boldsymbol{\beta}_v(\boldsymbol{\gamma}_v) + \boldsymbol{\epsilon}_v.$$

Further, we assume the independence among σ_v^2 and set its prior $\pi\left(\sigma_v^2\right) \propto 1/\sigma_v^2$.

Zellner's g-prior on $\boldsymbol{\beta}_v(\boldsymbol{\gamma}_v)|\boldsymbol{\gamma}_v$ is placed to undertake variable selection or model averaging. The parameter g is adjusted to obtain similar results with those if BIC were used,

$$\boldsymbol{\beta}_v(\boldsymbol{\gamma}_v)|\boldsymbol{y}_v, \sigma_v^2, \Lambda_v, \boldsymbol{\gamma}_v \sim N\left(\hat{\boldsymbol{\beta}}_v(\boldsymbol{\gamma}_v), T_v\sigma_v^2\left[X_v^T(\boldsymbol{\gamma}_v)\Lambda_v^{-1}X_v(\boldsymbol{\gamma}_v)\right]^{-1}\right),$$

where

$$\hat{\boldsymbol{\beta}}_v(\boldsymbol{\gamma}_v) = \left[X_v^T(\boldsymbol{\gamma}_v)\Lambda_v^{-1}X_v(\boldsymbol{\gamma}_v)\right]^{-1}X_v^T(\boldsymbol{\gamma}_v)\Lambda_v^{-1}\boldsymbol{y}_v. \tag{3.7}$$

Define the corresponding sum of squares for posterior inference

$$S(\rho_v, \boldsymbol{\gamma}_v) = \left(\boldsymbol{y}_v - X_v(\boldsymbol{\gamma}_v)\hat{\boldsymbol{\beta}}_v(\boldsymbol{\gamma}_v)\right)^T\Lambda_v^{-1}\left(\boldsymbol{y}_v - X_v(\boldsymbol{\gamma}_v)\hat{\boldsymbol{\beta}}_v(\boldsymbol{\gamma}_v)\right).$$

We incorporate the temporal dependence between observations on a given voxel through the specification of the structure of Λ_v. The AR(1) dependence, i.e. $\Lambda_v(i,j) = \rho_v^{|i-j|}$, is an effective compromise between inferential efficacy and computational efficiency. We specify a point mass prior for $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_N)$ at a fixed point $\hat{\boldsymbol{\rho}}$ using maximum likelihood methods.

We incorporate the spatial dependence, as well as the anatomical information, by using a binary Markov random field (MRF) prior, i.e. the Ising prior, on $\boldsymbol{\gamma}_v$. Let $\boldsymbol{\gamma}_{(j)} = (\gamma_{1,j}, \ldots, \gamma_{N,j})^T$ be the vector of indicators for regressor j over all voxels. Then, let $w_{v,k}$ be pre-specified constants that weigh the interaction between voxels v and k and let ν_j be parameter to measure the strength of the interaction between voxels for regressor j. We denote $v \sim k$, if two voxels v and k are defined as neighbors by the user. In this chapter, we employ a widely used three-dimensional structure containing the six immediate neighbors: 1 above, 1 below and 4 adjacent. The weight $w_{v,k}$ is set to be the reciprocal of the Euclidean distance between voxel v and k. Then, the spatial interaction is described as $\nu_j\sum_{v=1}^N\sum_{v\sim k} w_{v,k}I(\gamma_{v,j} = \gamma_{k,j})$, where $I(x)$ is the usual 0/1 indicator function. A linear "external field" $\sum_{v=1}^N \alpha_{v,j}\gamma_{v,j}$ is specified to incorporate anatomical prior information, where $\alpha_{v,j}$ is chosen to reflect prior knowledge.

We consider the prior on $\boldsymbol{\gamma}$ to be $\pi(\boldsymbol{\gamma}|\boldsymbol{\nu}) = \prod_{j=1}^p \pi\left(\boldsymbol{\gamma}_{(j)}|\nu_j\right)$, where

$$\pi\left(\boldsymbol{\gamma}_{(j)}|\nu_j\right) \propto \exp\left\{\sum_{v=1}^N \alpha_{v,j}\gamma_{v,j} + \nu_j\sum_{v=1}^N\sum_{v\sim k} w_{v,k}I(\gamma_{v,j} = \gamma_{k,j})\right\}.$$

The remaining prior to be addressed is the distribution of $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_p)$. A uniform prior is placed $\pi(\boldsymbol{\nu}) \propto \prod_{j=1}^{p} I(0 < \nu_j < \nu_{max})$, where Moller and Waagepetersen (2003) suggests to use $\nu_{max} \leq 2.0$.

The posterior density is characterized by

$$q\left(\boldsymbol{\beta}(\boldsymbol{\gamma}), \boldsymbol{\gamma}, \boldsymbol{\rho}, \boldsymbol{\nu}, \boldsymbol{\sigma}^2 | y\right) \propto p\left(y | \boldsymbol{\beta}(\boldsymbol{\gamma}), \boldsymbol{\gamma}, \boldsymbol{\sigma}^2, \Lambda\right) \times \pi\left(\boldsymbol{\beta}(\boldsymbol{\gamma}) | y, \boldsymbol{\sigma}^2, \Lambda, \boldsymbol{\gamma}\right) \pi(\boldsymbol{\gamma} | \boldsymbol{\nu}) \pi(\boldsymbol{\rho}) \pi\left(\boldsymbol{\sigma}^2\right) \pi(\boldsymbol{\nu}).$$

We follow the two-step component-wise Metropolis-hastings algorithm designed by Lee et al. (2014) to update $\boldsymbol{\gamma}$ and $\boldsymbol{\nu}$. Particularly, we are interested in estimating the posterior mean of $\boldsymbol{\theta} = \{\boldsymbol{\gamma}, \boldsymbol{\nu}\}$.

Particularly, we are interested in the StarPlus experiment of Carpenter et al. (1999). The experiment was designed to investigate brain activities related to high level cognition, i.e. language comprehension and visuospatial processing. Snapshots were taken every 0.5 seconds resulting in about 54 images throughout the experiment. Data were pre-processed using standard techniques such as slice timing and spatial smoothing (for a review see Lindquist, 2008) and were registered in standardized space with $64 \times 64 \times 8$ dimensions for 54 time points.

Based on the settings of the StarPlus experiment, we rewrite the linear model (3.6) as

$$\boldsymbol{y}_v = \alpha_0 \boldsymbol{z}_0 + \alpha_1 \boldsymbol{z}_1 + \beta_1 \boldsymbol{x}_1 + \beta_2 \boldsymbol{x}_2 + \boldsymbol{\epsilon}_v,$$

where $\alpha_i, \boldsymbol{z}_i$s are the baseline signal, β_is are the activation amplitude corresponding to the two tasks "Semantic" and "Symbol", respectively, The binary indicator $\gamma_v = \{1, 1, \gamma_{v,3}, \gamma_{v,4}\}$ is used in the variable selection problem described previously. Notice that we assume all α_is nonzero and set $\nu_{max} = 1.0$ as in Lee et al. (2014). Figure 3.3 visualize the design matrix for this linear model as we described previously.

We followed the component-wise Metropolis-hastings algorithm introduced in Lee et al. (2014) to update the 9398-dimensional posterior distribution. Given the computational challenges that arise from the high-dimensionality, it is not practicable to estimate the "truth" and subsequent coverage probabilities via multiple parallel runs using the resources at hand. Instead, the simulation study presented is from a single run of the Markov chain.

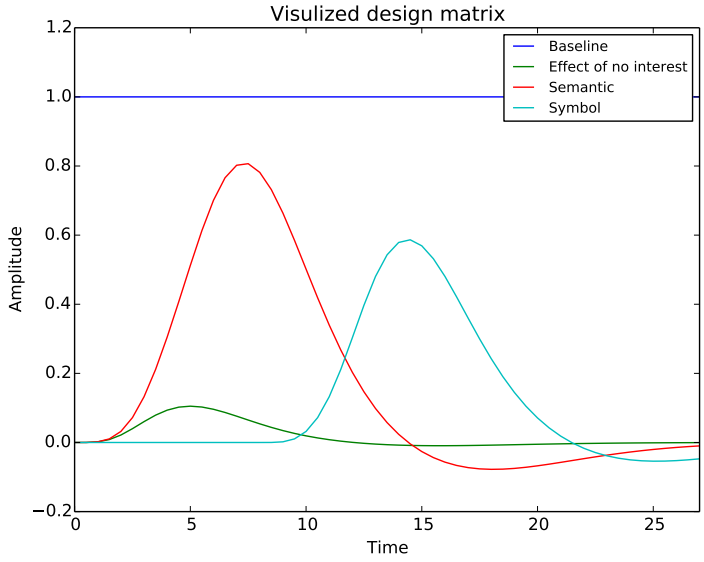

Figure 3.3: The visualization of the design matrix for the experimental dataset.

Terminating the simulation

Since the equivalency between the relative standard deviation FWSR and the ESS stopping rule has been established in the previous sections, we restrict our attention to relative standard deviation FWSR in this application. Given the dimension of the problem, in-memory storage of the entire chain becomes infeasible when the simulation approaches 1E6 iterations. Thus, the use of CBM in the stopping criteria is not an option. At the same time, by only keeping the summarized information, LCBM offers a practical solution with minimal tradeoffs. We, therefore, implemented $T(\epsilon, \delta)$ with LCBM as the variance estimator in this study.

The relative standard deviation FWSR was implemented with $\delta = 0.05$ and $\epsilon = 0.062$, which is equivalent to setting the threshold of the ESS stopping rule to $K = 4000$. We set $n^* = 2^{14} = 16,384$ and added 20 or 21 batches between checks. The simulation was terminated after $368,640$ iterations, which was 360 batches under LCBM. The storage required was approximately 84 megabytes, where it would have required over 84 gigabytes if we were to use CBM. Figure 3.4 and 3.5 are the estimated activation maps in eight brain slices for two tasks "Semantic" and "Symbol", respectively. The red voxels are identified

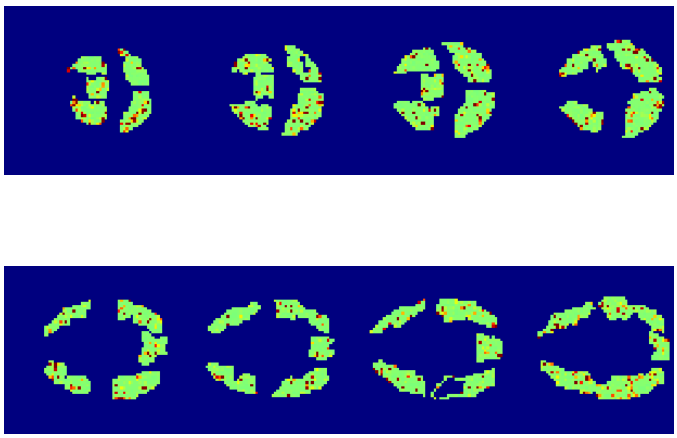

Figure 3.4: The activation map for all eight slices when perform task "Semantic".

as activated.

Again, we compared terminating simulations by $T(\epsilon, \delta)$ and through the GD. For the GD, we generated 20,000 iterations and discarded the first 5,000 as burn-in. We confirmed the chain convergence with a p-value equals 0.05 over all parameters. A parameter-wise estimation of ESS revealed that the smallest ESS from GD is 398, while the $T(0.062, 0.05)$ guarantees a minimal ESS greater than 4000 in the above setting. Moreover, it took about 3.5 gigabytes to store the entire chain for the GD, which is 100 times more memory usage than the relative standard deviation FWSR with LCBM implemented. We also note around 10% of the voxels stayed active or inactive, i.e. a sequence of constant 0 or 1, throughout the simulation terminated by GD. On the contrary, there were only 3.5% of voxels behaving in such a way in the simulation terminated by FWSR. This suggests that the GD as a stopping rule tends to result in a premature termination.

Anatomical knowledge suggests that certain areas of the brain, known as the region of interest (ROI), are more likely to be activated during the experiment. Looking at a particular ROI called 'LT', we found that the differences between the estimated percentage of activated voxels from two simulations are considerable (see Table 3.2), given the small

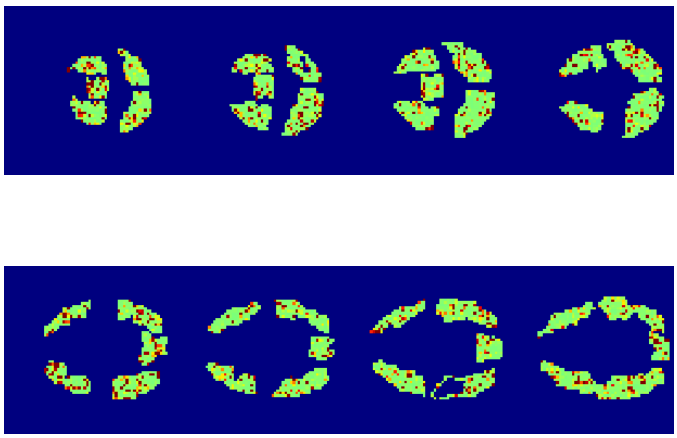

Figure 3.5: The activation map for all eight slices when perform task "Symbol".

'LT'	FWSR	GD
Semantic	3.93%	4.26%
Symbol	12.46%	12.79%

Table 3.2: Comparisons of the activated voxels in ROI based on FWSR and GD.

proportion of activation.

3.4 Discussion

This chapter considers a relative standard deviation FWSR in the context of truly high-dimensional MCMC simulations. In our viewpoint, a practical stopping rule should achieve three properties: (1) it is easy to implement in an automated fashion with a few tuning parameters; (2) it attains confidence in resulting estimates; and (3) it is applicable in both low- and high-dimensional settings. With such properties, practitioners can then apply the stopping rule on a routine basis.

We advocate use of the modified relative standard deviation FWSR since it meets all the properties and is especially applicable in high-dimensional Bayesian settings without

prior knowledge of the magnitude of the target parameters. It is controlled by one tuning parameter ϵ that measures the accuracy of the estimates. Simply put, the estimates are approximately ϵ^{-1} more accurate than their posterior standard deviation. Another way to understand ϵ is through the alternative ESS calculation. In high-dimensional settings, we suggest setting the tuning parameter $\epsilon = 0.062, \delta = 0.05$ which leads to an ESS of 4000. However, the choice of ϵ significantly affects the total simulation effort. For instance, in the described fMRI study, $\epsilon = 0.062, \delta = 0.05$ results in 368,640 iterations, while $\epsilon = 0.02, \delta = 0.05$ results in 1,419,264 iterations. Thus, there should be a balance between the accuracy of estimation and the cost of simulation.

The proposed LCBM sampling plan summarizes information along the simulation. On one hand, it eliminates the requirement of storing the entire chain in memory to allow recalculation and also reduces computing time. On the other hand, it is not applicable for more general estimation problems such as quantile estimation (Doss et al., 2014). Such tradeoffs are necessary in order to overcome challenges that arise in high-dimensional settings.

A natural extension of the stopping rule is to consider simultaneous multivariate estimation. Flegal and Gong (2015) apply a Bonferonni approach to adjust for multiplicity. However, the standard Bonferonni approach will not work for a large dimension p, since the individual confidence interval needs to be set to a nominal level of $0.95^{1/p}$. Clearly, a more sophisticated method is required to adjust for multiplicity in high-dimensional settings. One direction of future research is to control the volume of a desired confidence region rather than the width of multiple confidence intervals separately.

The utility of FWSRs remains an open question when MCMC is used as part of an optimization algorithm. For example, one could consider the relative standard deviation FWSR in conjunction with the Monte Carlo EM algorithm, see e.g. Caffo et al. (2005). In this setting, a sequence of decreasing ϵ values could be used in order to increase accuracy when the EM algorithm is near convergence. One could also consider FWSRs in the context of maximum likelihood using MCMC, see e.g. Geyer and Thompson (1992).

Finally, how well a chain mixes and explores the state space is a vital component in MCMC simulations. In our opinion, this is the role of convergence diagnostics and visual inspections. However, this remains a challenging problem for Bayesian practitioners regardless of dimension. We direct interested readers to Brooks et al. (2010) for more information.

Chapter 4

Bayesian model selection on linear mixed-effects models

This chapter proposes a novel Bayesian model selection on linear mixed-effects models for comparisons between multiple treatments and a control. A fully Bayesian solution provides practitioners with marginal inclusion probability for each treatment that directly measures its significance, along with model-averaged posterior distributions. It extends the existing literature by incorporating multiple group effects with unbalanced subjects into the stochastic search variable selection framework. Default priors are proposed for model selection and a component-wise Gibbs sampler is developed for posterior computation. A simulation study and a longitudinal experiment of mouse weight trajectories (Spindler et al., 2014a,b, 2013a,b, 2014c) are used to evaluate the performance of the proposed method. This application also serves as an example to advocate the use of the relative standard deviation FWSR for careful posterior inference.

4.1 Introduction

Experiments are run by researchers in medicine, biology, and various other scientific fields, to compare multiple treatments with a control or standard treatment over a period of time. Often these studies result in unbalanced repeated measured data that is widely analyzed by the flexible linear mixed-effects model (LMM). The LMM allows for some subsets of the regression parameters to vary among subjects, thereby accounting for sources of natural heterogeneity in the population. It models the mean response as a com-

bination of population characteristics (fixed-effects), that are assumed to be shared by all subjects, and subject-specific characteristics (random-effects) that are unique to a particular subject. For comparisons between multiple treatment groups and a control group, it is common to introduce a set of fixed-effects to model the treatment effects for each group (see e.g. Fitzmaurice et al., 2004). The comparison between groups is then equivalent to compare parameter estimation between the set of fixed-effects.

A difficult question is how to decide which treatments are significantly different from the control. Standard model selection criteria and test procedures can be implemented to solve this problem (see e.g. Bolker et al., 2009; Fitzmaurice et al., 2004) with certain disadvantages. One can select models by using hypothesis tests (Stephens et al., 2005); that is, test simpler nested models against more complex models and report corresponding p-values. Although the likelihood ratio test (LRT) is widely used to determine the contribution of a factor in a model throughout statistics, it is not recommended by Pinheiro and Bates (2006) for testing fixed-effects in LMM, because of its unreliability for small to moderate sample size. Also, when the focus is to compare multiple treatments with the control, Burnham and Anderson (2002) criticize that such a pairwise comparison as an abuse of hypothesis testing. Another extensively used approach is the information-theoretic model selection procedure that allows comparison of multiple models (see e.g. Burnham and Anderson, 2002). This method relies on information criteria, such as Akaike information criterion (AIC) and Bayesian information criterion (BIC), that use deviance as a measure of fit with a penalization on more complex models. Instead of reporting p-values, it estimates the magnitude of difference between models in expected predictive power that can sometimes be difficult to fully understand and explain by practitioners.

Motivated by these practical challenges faced by frequentist approaches, we resort to Bayesian model selection methods (for a review see e.g. Clyde and George, 2004; George and McCulloch, 1997; Kuo and Mallick, 1998). In the Bayesian framework, this problem can be transformed to the form of parameter estimation (O'Hara et al., 2009). That is, the marginal posterior probability that a variable should be in the model, i.e. the marginal inclusion probability, which is usually calculated directly from the posterior inference using an Markov chain Monte Carlo (MCMC) simulation. This inclusion probability provides practitioners with an intuitive understanding of the significance of each fixed-effect and a way to combine prior knowledge of different treatments.

There is an extensive literature on Bayesian model selection for fixed-effects.

George and McCulloch (1993); Geweke et al. (1996) develop the stochastic search variable selection (SSVS) technique for linear regression models that uses a Gibbs sampler to traverse the model space. Smith and Kohn (1996) extend its application to nonparametric regression models and show how integrating the regression parameters is essential to reliable convergence of a Gibbs sampler. Kohn et al. (2001) propose a more efficient single-site Metropolis-Hastings sampler. Holmes et al. (2002) consider selection and smoothing for a series of seemingly unrelated regressions. Chen and Dunson (2003); Kinney and Dunson (2007) develop variable selection for both fixed and random effects in generalized LMM. Recently, Bayesian model selection methods are extended to a series of spatially linked regression for functional magnetic resonance imaging (fMRI) analysis (see e.g. Lee et al., 2014; Smith and Fahrmeir, 2007). However, we are unaware of any work to extend Bayesian model selection on LMMs with multiple group effects that involve unbalanced subjects.

In this chapter, we develop a novel Bayesian model selection approach on LMMs to accommodate multiple group effects. The method includes a re-parameterization of the fixed-effects to attribute part of each treatment effect to a baseline, i.e. an effect of the control group. A modification of the fractional prior (Smith and Kohn, 1997) is proposed to undertake model selection and averaging. This prior is related to Zellner's g-prior (Zellner, 1986), and is critical to incorporate information of subjects in the same treatment group and any prior knowledge of that treatment. A component-wise Gibbs sampler is then developed for efficient posterior computation. As an example, we consider a longitudinal experiment of mouse weight trajectories (see e.g. Spindler et al., 2014a,b, 2013a,b, 2014c) that aims to identify significant treatments with respect to the control. This application provides both a clear demonstration of our approach and intuitive results for researchers to understand which treatments significantly affect weight trajectories of mouse. The method is general and applicable to most experiments that are interested in comparing treatment groups to a control group.

4.1.1 Experimental Data

As a motivating example, the experimental data is from a longitudinal study on the life span of an F1 hybrid mouse (see e.g. Spindler et al., 2014a,b, 2013a,b, 2014c). The study is part of a compound screening program designed to identify potential longevity therapeutics, and it was approved by the Institutional Animal Care and Use Committee

at the University of California, Riverside. It utilized an unbalanced statistical design to compare the life span of multiple treatment groups to that of one larger control group (Jeske et al., 2014). In this chapter, we use a part of the data sets that record mouse weight trajectories throughout the experiment.

In the study, 2266 male C3B6F1 mice were initially on ad libitum chow feeding. At 12 month of age (Day 365), 297 mice were shifted to daily feeding with 13.3 kcal/day/mouse of the control diet (Diet No.99), and the rest were shifted to 56 different treatment groups. All mice were fed daily and weighted bimonthly, but the number of mice progressively declined as the study progressed. The data are censored at extreme old age (Day 1369), when only 1% of the mice remained.

The control and drug-treated mice gradually lost weight when they were shifted from ad libitum chow feeding to the defined diets. Figure 4.1 shows spaghetti plot for the control diet, and Figure 4.2 shows the days on diet versus mean weight trajectories for all 56 treatment diets and the control diet. Note that the mean weight estimates become more unstable as days on diet increases since mice die off within the study. Our main interest is to determine which alternative diets would affect lifetime weight trajectories. That is, researchers are interested in if any deviation from the trajectory of the control group (Diet No.99) is statistically significant and is caused by the difference between diets.

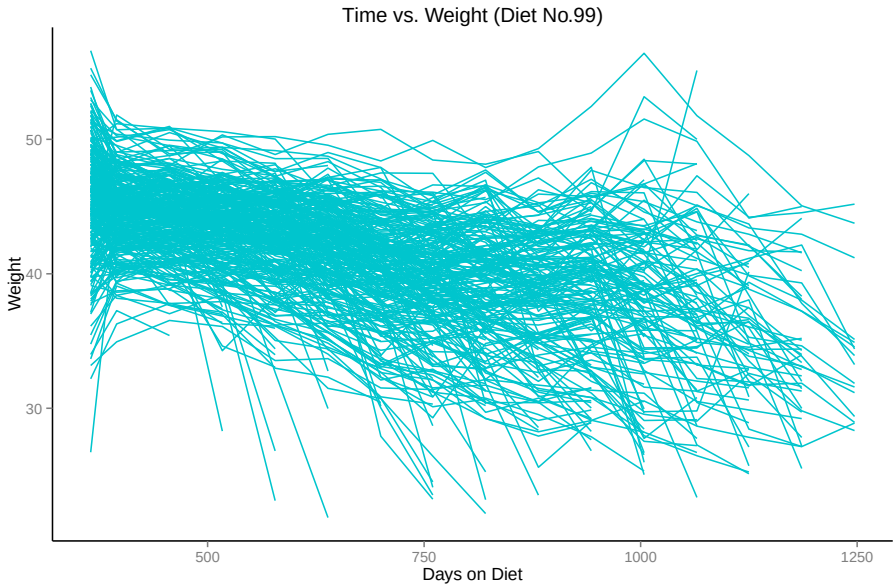

Figure 4.1: Spaghetti plot for the control diet in the experimental dataset.

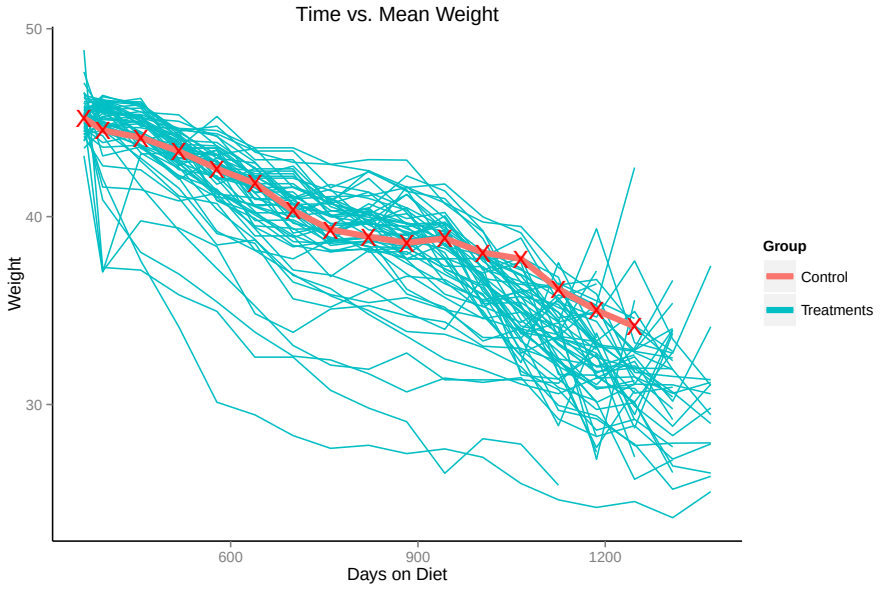

Figure 4.2: Combined plot of time on diet versus mean weight for 57 diets in the experimental dataset.

The rest of the chapter is organized as follows. Section 4.2 formally introduces the Bayesian variable selection methodology. It outlines the re-parameterization of a LMM,

associated with the prior specification, the MCMC sampling schemes and the stopping criterion utilized. A simulation study is also detailed to evaluate the performance of the proposed method. Section 4.5 contains the empirical results from the analysis of the motivating example. Section 4.4 concludes with a discussion.

4.2 Model Selection on Linear Mixed-effects Models

In general, suppose that we have n subjects from G experimental groups under study, each with n_i observations taken repeatedly over time, $i = 1, \cdots, n$, and let $\boldsymbol{y}_i = (y_{i,1}, \cdots, y_{i,n_i})^T$ denote the response vector for the i-th subject. Assume the i-th subject is from the g-th group, for $i = 1, \cdots, n$, $g = 1, \cdots, G$, let X_i and Z_i be two $n_i \times p$ design matrices, then a LMM (Fitzmaurice et al., 2004; McCullagh and Nelder, 1989) is denoted as

$$\boldsymbol{y}_i = X_i\boldsymbol{\alpha}_g + Z_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \sim N_{n_i}(\mathbf{0}, \sigma^2 I), \tag{4.1}$$

where $\boldsymbol{\alpha}_g = (\alpha_{g,0}, \cdots, \alpha_{g,p-1})^T$ are the fixed effects shared by subjects in the same experimental group. Further, denote $\boldsymbol{b}_i = (b_{i,0}, \cdots, b_{i,p-1})^T \sim N_p(\mathbf{0}, \lambda_D^{-1} I)$ as the random effects that are unique to a particular subject, and hence we allow subject specific trajectories.

Note that, among the G groups, there is one control group and $G-1$ treatment groups. Without loss of generality, let us assume the G-th group to be the control group, and $g = 1, \cdots, G-1$ are the treatment groups. A primary goal for many of these experiments is to determine which alternative treatments are statistically significantly differ from the control group. To this end, we propose a re-parameterization of the fixed effects $\boldsymbol{\alpha}_g$'s in (4.1), $g = 1, \cdots G$. Let W_i, X_i and Z_i be three $n_i \times p$ design matrices, the re-parameterized mixed-effects model is denoted as, for $i = 1, \cdots, n$, $g = 1, \cdots, G$,

$$\boldsymbol{y}_i = W_i\boldsymbol{\alpha} + X_i\boldsymbol{\beta}_g + Z_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \sim N_{n_i}(\mathbf{0}, \sigma^2 I), \tag{4.2}$$

where $\boldsymbol{b}_i = (b_{i,0}, \cdots, b_{i,p-1})^T \sim N_p(\mathbf{0}, \lambda_D^{-1} I)$ are the random effects as in (4.1), and $\boldsymbol{\alpha} = (\alpha_0, \cdots, \alpha_{p-1})^T$ are the fixed effects of the control group, $\boldsymbol{\beta}_g = (\beta_{g,0}, \cdots, \beta_{g,p-1})^T$ are the difference between the fixed effects of the g-th group and the control group. That is, the group effect $\boldsymbol{\alpha}_g$ is re-written as $\boldsymbol{\alpha} + \boldsymbol{\beta}_g$, for $g = 1, \cdots, G$. Also, it is straightforward to set $\boldsymbol{\beta}_G = (0, \cdots, 0)^T$ for the control group in this re-parameterization as the baseline.

Under the re-parameterization, the detection of significant diets is equivalent to

the identification of nonzero $\boldsymbol{\beta}_g$'s. To this end, we introduce 0/1 binary indicators $\boldsymbol{\gamma}_g = (\gamma_{g,0}, \cdots, \gamma_{g,p-1})^T, g = 1, \cdots, G$, such that $\beta_{g,j} = 0$ if $\gamma_{g,j} = 0$ and $\beta_{g,j} \neq 0$ if $\gamma_{g,j} = 1$. The $\gamma_{g,j}$ is used to indicate whether the fixed effect on the j-th predictor of the g-th is significantly differ from that fixed effect of the control group. Given $\boldsymbol{\gamma}_g$, let $\boldsymbol{\beta}_g(\boldsymbol{\gamma}_g)$ be the vector of nonzero fixed effects and $X_i(\boldsymbol{\gamma}_g)$ be the corresponding design matrix. Then, the model (4.2) can be written as

$$\boldsymbol{y}_i = W_i\boldsymbol{\alpha} + X_i(\boldsymbol{\gamma}_g)\boldsymbol{\beta}_g(\boldsymbol{\gamma}_g) + Z_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i. \tag{4.3}$$

This formulation allows us to look at the problem form the Bayesian stochastic search variable selection (SSVS) perspective (George and McCulloch, 1993). The SSVS searches for models having high posterior probability by traversing the model space using MCMC techniques. Moreover, it allows us to calculate the posterior distributions of parameters by marginalizing over all the over variables. In this way, the marginal posterior inclusion probabilities can be obtained to measure the significance of each diet.

Note that, (4.3) is a very general setting that is applicable to a wide range of applications. It is possible to impose specific structures on γ to suit different settings to further simplify modeling procedure. For example, given the setups of the motivating experiment, it is reasonable to assume a common intercept for the fixed effects since all mice were on the same diet at the first measurement; that is, $\beta_{g,0} = 0$, for $g = 1, \cdots, G$. Therefore, it is desirable to impose the following settings on γ for this application, where the primary goal is to compare the treatment groups to the baseline group,

$$\gamma_{1,0} = \cdots = \gamma_{G-1,0} = 0,$$

$$\boldsymbol{\gamma}_G = (\gamma_{G,0}, \cdots, \gamma_{G,p-1})^T = (0, \cdots, 0)^T.$$

4.2.1 Prior Specification

A proper prior must be placed on the nonzero coefficients $\boldsymbol{\beta}_g(\boldsymbol{\gamma}_g)$ to undertake model averaging (see e.g. George and McCulloch, 1993; Kohn et al., 2001; Mitchell and Beauchamp, 1988; Smith and Kohn, 1996). In particular, Kohn et al. (2001); Smith and Fahrmeir (2007) propose a conditional prior for the coefficients by setting it proportional to a fraction of the likelihood. This fractional prior is related to the g-prior in Zellner

(1986), and is located and scaled in line with the information from the likelihood. We propose an extension of this idea to accommodate multiple subjects within a group by setting $\pi\left(\boldsymbol{\beta}_g(\boldsymbol{\gamma}_g)|y,\boldsymbol{\alpha},\boldsymbol{\gamma}_g,b,\sigma^2\right) \propto \Pi_{i\in g}\ p\left(\boldsymbol{y}_i|\boldsymbol{\alpha},\boldsymbol{\beta}_g(\boldsymbol{\gamma}_g),\boldsymbol{\gamma}_g,\boldsymbol{b}_i,\sigma^2\right)^{1/n_i}$, so that

$$\boldsymbol{\beta}_g(\boldsymbol{\gamma}_g)|y,\boldsymbol{\alpha},\boldsymbol{\gamma}_g,b,\sigma^2 \sim N\left(\hat{\boldsymbol{\beta}}_g(\boldsymbol{\gamma}_g),\sigma^2\left(\sum_{i\in g}\frac{1}{n_i}X_i^T(\boldsymbol{\gamma}_g)X_i(\boldsymbol{\gamma}_g)\right)^{-1}\right), \tag{4.4}$$

where $\hat{\boldsymbol{\beta}}_g(\boldsymbol{\gamma}_g) = \left(\sum_{i\in g}\frac{1}{n_i}X_i^T(\boldsymbol{\gamma}_g)X_i(\boldsymbol{\gamma}_g)\right)^{-1}\left(\sum_{i\in g}\frac{1}{n_i}X_i^T(\boldsymbol{\gamma}_g)\left(\boldsymbol{y}_i - W_i\boldsymbol{\alpha} - Z_i\boldsymbol{b}_i\right)\right)$, and $\sum_{i\in g}$ stands for summation over all the subjects that belong to the g-th group.

This prior is proportional to the variance of the least squares estimate of β, and enjoys a number of attractive properties as pointed out by Kohn et al. (2001). The prior (4.4) is rescaled automatically if the design matrix X or the data y is rescaled because of its structure and the presence of σ^2. Moreover, this prior is invariant to location changes in X and y given the basis term $(1,\cdots,1)^T$ is included in X. Also it is data-based since $\hat{\boldsymbol{\beta}}_g(\boldsymbol{\gamma}_g)$ depends on y, which allows proper centering of β.

We consider the prior on γ to be $\pi(\boldsymbol{\gamma}_g|\pi_g) = \prod_{j=0}^{p-1}\pi(\gamma_{g,j}|\pi_g), g = 1,\cdots,G$, where $\pi(\gamma_{g,j}|\pi_g) \sim Bernoulli(\pi_g)$ and $\boldsymbol{\pi} = (\pi_1,\cdots,\pi_G)^T$ is a vector of hyper-parameters that represents prior knowledge for all experimental groups. For instance, we find a sensible setting, when there is little prior knowledge of the effects of the alternative treatments, to be letting $\pi_G = 0$ for the control group, and $\pi_1 = \cdots = \pi_{G-1} = 0.5$ for the $G-1$ treatment groups.

We assume standard priors in Bayesian hierarchical models (see e.g. Gelman et al., 2004; Johnson and Jones, 2010; Smith and Kohn, 1996) for the rest of the parameters, i.e. $\boldsymbol{\alpha}, b, \lambda_D, \sigma^2$,

$$\boldsymbol{\alpha}|\boldsymbol{d}_3, d_4 \sim N_p(\boldsymbol{d}_3, d_4^{-1})$$

$$\boldsymbol{b}_i|\lambda_D \sim N_p(\boldsymbol{0}, \lambda_D^{-1}I), i = 1,\cdots,n$$

$$\lambda_D|d_1,d_2 \sim \Gamma(d_1,d_2)$$

$$\pi(\sigma^2) \propto 1/\sigma^2$$

where $d_1, d_2, \boldsymbol{d}_3, d_4$ are hyper-parameters to be pre-specified.

4.2.2 Posterior Inference

Combining the priors and likelihoods, the full joint posterior density for $\theta = (\boldsymbol{\alpha}, \beta, \gamma, b, \sigma^2, \lambda_D)$ is characterized by

$$q(\boldsymbol{\alpha}, \beta, \gamma, b, \sigma^2, \lambda_D | y) \propto \left[\prod_{g=1}^{G} \left[\prod_{i \in g} p\left(\boldsymbol{y}_i | \boldsymbol{\alpha}, \boldsymbol{\beta}_g, \boldsymbol{\gamma}_g, \boldsymbol{b}_i, \sigma^2\right) \pi\left(\boldsymbol{b}_i | \lambda_D\right) \right] \pi\left(\boldsymbol{\beta}_g | \boldsymbol{\alpha}, \boldsymbol{\gamma}_g, b, \sigma^2\right) \pi\left(\boldsymbol{\gamma}_g\right) \right] \times \pi(\boldsymbol{\alpha}) \pi(\lambda_D) \pi(\sigma^2). \tag{4.5}$$

This distribution has a complex form which we cannot sample from directly; instead, we resort to MCMC methodology for the posterior inference and employ a component-wise strategy (Johnson et al., 2013). To this end, we need the full conditional posterior distributions of each of the parameters in θ to update the Markov chains. The derivation of all full conditional posterior distributions follows from (4.5) using straightforward algebraic route (see Section 4.5).

Specifically, we can set up a six-variable component-wise Gibbs sampler; that is, if we let $\theta = (\gamma, \beta, \boldsymbol{\alpha}, \sigma^2, b, \lambda_D)$ be the current state and $\theta' = (\gamma', \beta', \boldsymbol{\alpha}', (\sigma^2)', b', \lambda'_D)$ be the future state, we iteratively sample from the full conditional posterior distributions,

$$\begin{aligned}(\gamma, \beta, \boldsymbol{\alpha}, \sigma^2, b, \lambda_D) &\to (\gamma', \beta, \boldsymbol{\alpha}, \sigma^2, b, \lambda_D) \to (\gamma', \beta', \boldsymbol{\alpha}, \sigma^2, b, \lambda_D) \to (\gamma', \beta', \boldsymbol{\alpha}', \sigma^2, b, \lambda_D) \\ &\to (\gamma', \beta', \boldsymbol{\alpha}', (\sigma^2)', b, \lambda_D) \to (\gamma', \beta', \boldsymbol{\alpha}', (\sigma^2)', b', \lambda_D) \to (\gamma', \beta', \boldsymbol{\alpha}', (\sigma^2)', b', \lambda'_D).\end{aligned}$$

Step 1. Consider updating γ using a Gibbs sampler. Schematically, the transition $\gamma \to \gamma'$ consists of $G \times p$ steps

$$\begin{aligned}(\gamma_{1,0}, \gamma_{1,1}, \cdots, \gamma_{1,p-1}, \cdots, \gamma_{G,0}, \cdots, \gamma_{G,p-1}) &\to ({\gamma'}_{1,0}, \gamma_{1,1}, \cdots, \gamma_{1,p-1}, \cdots, \gamma_{G,0}, \cdots, \gamma_{G,p-1}) \\ &\to ({\gamma'}_{1,0}, {\gamma'}_{1,1}, \cdots, \gamma_{1,p-1}, \cdots, \gamma_{G,0}, \cdots, \gamma_{G,p-1}) \\ &\vdots \\ &\to ({\gamma'}_{1,0}, {\gamma'}_{1,1}, \cdots, {\gamma'}_{1,p-1}, \cdots, {\gamma'}_{G,0}, \cdots, {\gamma'}_{G,p-1}).\end{aligned}$$

From the Section 4.5, we have, for $g = 1, \cdots, G$ and $j = 0, \cdots, p-1$,

$$q(\gamma_{g,j}|\boldsymbol{\alpha}, \boldsymbol{\gamma}_{-(g,j)}, b, \sigma^2, y) \propto \pi_g^{\gamma_{g,j}}(1-\pi_g)^{1-\gamma_{g,j}} \left(\frac{|\sum_{i\in g} \frac{1}{n_i} X_i^T(\boldsymbol{\gamma}_g) X_i(\boldsymbol{\gamma}_g)|}{|\sum_{i\in g}(1+\frac{1}{n_i}) X_i^T(\boldsymbol{\gamma}_g) X_i(\boldsymbol{\gamma}_g)|} \right)^{\frac{1}{2}}$$
$$\times \exp\left\{ -\frac{1}{2\sigma^2}\Big[\sum_{i\in g} \boldsymbol{\phi}_i^T \boldsymbol{\phi} + (\sum_{i\in g} \frac{1}{n_i} X_i^T(\boldsymbol{\gamma}_g)\boldsymbol{\phi}_i)^T (\sum_{i\in g} \frac{1}{n_i} X_i^T(\boldsymbol{\gamma}_g) X_i(\boldsymbol{\gamma}_g))^{-1} (\sum_{i\in g} \frac{1}{n_i} X_i^T(\boldsymbol{\gamma}_g)\boldsymbol{\phi}_i) \right.$$
$$\left. - (\sum_{i\in g}(1+\frac{1}{n_i}) X_i^T(\boldsymbol{\gamma}_g)\boldsymbol{\phi}_i)^T (\sum_{i\in g}(1+\frac{1}{n_i}) X_i^T(\boldsymbol{\gamma}_g) X_i(\boldsymbol{\gamma}_g))^{-1} (\sum_{i\in g}(1+\frac{1}{n_i}) X_i^T(\boldsymbol{\gamma}_g)\boldsymbol{\phi}_i) \Big] \right\}, \tag{4.6}$$

where $\boldsymbol{\gamma}_{-(g,j)} = (\gamma_{g,0}, \cdots, \gamma_{g,j-1}, \gamma_{g,j+1}, \cdots, \gamma_{g,p-1})^T$ and $\boldsymbol{\phi}_i = \boldsymbol{y}_i - W_i\boldsymbol{\alpha} - Z_i\boldsymbol{b}_i$.

At each step, an update is simulated from ${\gamma_{g,j}}' \sim q(\gamma_{g,j}|\boldsymbol{\alpha}, \boldsymbol{\gamma}_{-(g,j)}, b, \sigma^2, y)$. Since $\gamma_{g,j}$ is binary, i.e. $\gamma_{g,j} \in \{0, 1\}$, the conditional posterior distribution $q(\gamma_{g,j}|\boldsymbol{\alpha}, \boldsymbol{\gamma}_{-(g,j)}, b, \sigma^2, y)$ is easily normalized by evaluating (4.6) for $\gamma_{g,j} = 0$ and $\gamma_{g,j} = 1$.

Step 2. Consider updating β using a Gibbs sampler. Schematically, the transition $\beta \to \beta'$ consists of G steps

$$\begin{aligned}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \cdots, \boldsymbol{\beta}_G) &\to ({\boldsymbol{\beta}_1}', \boldsymbol{\beta}_2, \cdots, \boldsymbol{\beta}_G) \\ &\to ({\boldsymbol{\beta}_1}', {\boldsymbol{\beta}_2}', \cdots, \boldsymbol{\beta}_G) \\ &\vdots \\ &\to ({\boldsymbol{\beta}_1}', {\boldsymbol{\beta}_2}', \cdots, {\boldsymbol{\beta}_G}').\end{aligned}$$

At each step, an update is simulated from a p-dimensional multivariate normal distribution,

$$\begin{aligned}{\boldsymbol{\beta}_g}'(\boldsymbol{\gamma}_g) \sim\ & q(\boldsymbol{\beta}_g(\boldsymbol{\gamma}_g)|\boldsymbol{\alpha}, \boldsymbol{\gamma}_g, b, \sigma^2, y) \\ \sim\ & N_{\sum_{j=0}^{p-1}\gamma_{g,j}} \left(V_1^{-1} \left[\frac{1}{\sigma^2} \sum_{i\in g} X_i^T(\boldsymbol{\gamma}_g)(\boldsymbol{y}_i - W_i\boldsymbol{\alpha} - Z_i\boldsymbol{b}_i) \right], V_1^{-1} \right),\end{aligned} \tag{4.7}$$

where $V_1 = \frac{1}{\sigma^2}\sum_{i\in g}(1+\frac{1}{n_i}) X_i^T(\boldsymbol{\gamma}_g) X_i(\boldsymbol{\gamma}_g)$.

Step 3. Consider updating $\boldsymbol{\alpha}$ using a Gibbs sampler. At each step, an update is simulated

from a p-dimensional multivariate normal distribution,

$$\begin{aligned}
\boldsymbol{\alpha}' \sim\ & q(\boldsymbol{\alpha}|\beta,\gamma,b,\sigma^2,y) \\
& \sim N_p\Bigg(V_2^{-1}\Bigg[\frac{1}{\sigma^2}\sum_{g=1}^{G}\Bigg(\sum_{i\in g}W_i^T(\boldsymbol{y}_i - X_i(\boldsymbol{\gamma}_g)\boldsymbol{\beta}_g(\boldsymbol{\gamma}_g) - Z_i\boldsymbol{b}_i) \\
& + (\sum_{i\in g}\frac{1}{n_i}X_i^T(\boldsymbol{\gamma}_g)W_i)^T(\sum_{i\in g}\frac{1}{n_i}X_i^T(\boldsymbol{\gamma}_g)X_i(\boldsymbol{\gamma}_g))^{-1}(\sum_{i\in g}\frac{1}{n_i}X_i^T(\boldsymbol{\gamma}_g)(\boldsymbol{y}_i - X_i(\boldsymbol{\gamma}_g)\boldsymbol{\beta}_g(\boldsymbol{\gamma}_g) - Z_i\boldsymbol{b}_i))\Bigg) \\
& \qquad + d_4\boldsymbol{d}_3\Bigg], V_2^{-1}\Bigg),
\end{aligned} \tag{4.8}$$

where

$$V_2 = \frac{1}{\sigma^2}\sum_{g=1}^{G}\Bigg[\sum_{i\in g}W_i^TW_i + \Bigg(\sum_{i\in g}\frac{1}{n_i}X_i^T(\boldsymbol{\gamma}_g)W_i\Bigg)^T\Bigg(\sum_{i\in g}\frac{1}{n_i}X_i^T(\boldsymbol{\gamma}_g)X_i(\boldsymbol{\gamma}_g)\Bigg)^{-1}\Bigg(\sum_{i\in g}\frac{1}{n_i}X_i^T(\boldsymbol{\gamma}_g)W_i\Bigg)\Bigg] + d_4.$$

Step 4. Consider updating σ^2 using a Gibbs sampler. At each step, an update is simulated from a Inverse-Gamma distribution, i.e. $(\sigma^2)' \sim q(\sigma^2|\boldsymbol{\alpha},\beta,\gamma,b,y)$.

$$\begin{aligned}
(\sigma^2)' \sim\ & q(\sigma^2|\boldsymbol{\alpha},\beta,\gamma,b,y) \\
& \sim \textit{Inv-Gamma}\Bigg(\frac{1}{2}(N + \sum_{g=1}^{G}\sum_{j=0}^{p-1}\gamma_{g,j}), \\
& \frac{1}{2}\sum_{g=1}^{G}\Bigg[\sum_{i\in g}(\boldsymbol{y}_i - W_i\boldsymbol{\alpha} - X_i(\boldsymbol{\gamma}_g)\boldsymbol{\beta}_g(\boldsymbol{\gamma}_g) - Z_i\boldsymbol{b}_i)^T(\boldsymbol{y}_i - W_i\boldsymbol{\alpha} - X_i(\boldsymbol{\gamma}_g)\boldsymbol{\beta}_g(\boldsymbol{\gamma}_g) - Z_i\boldsymbol{b}_i) \\
& \qquad + [\boldsymbol{\beta}_g - (\sum_{i\in g}\frac{1}{n_i}X_i^T(\boldsymbol{\gamma}_g)X_i(\boldsymbol{\gamma}_g))^{-1}(\sum_{i\in g}\frac{1}{n_i}X_i^T(\boldsymbol{\gamma}_g)\boldsymbol{\phi}_i)]^T(\sum_{i\in g}\frac{1}{n_i}X_i^T(\boldsymbol{\gamma}_g)X_i(\boldsymbol{\gamma}_g))^{-1} \\
& \qquad [\boldsymbol{\beta}_g - (\sum_{i\in g}\frac{1}{n_i}X_i^T(\boldsymbol{\gamma}_g)X_i(\boldsymbol{\gamma}_g))^{-1}(\sum_{i\in g}\frac{1}{n_i}X_i^T(\boldsymbol{\gamma}_g)\boldsymbol{\phi}_i)]\Bigg]\Bigg),
\end{aligned} \tag{4.9}$$

where $N = \sum_{g=1}^{G}\sum_{i\in g}n_i$. Note that $\textit{Inv-Gamma}(\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{-\alpha-1}\exp\left(\frac{-\beta}{x}\right)$, for $x \in (0,\infty)$, and $\alpha,\beta > 0$.

Step 5. Considering updating b using a Gibbs sampler. Schematically, the transition $b \to b'$

consists of n steps

$$\begin{aligned}(\boldsymbol{b}_1, \boldsymbol{b}_2, \cdots, \boldsymbol{b}_n) &\to (\boldsymbol{b}_1{}', \boldsymbol{b}_2, \cdots, \boldsymbol{b}_n) \\ &\to (\boldsymbol{b}_1{}', \boldsymbol{b}_2{}', \cdots, \boldsymbol{b}_n) \\ &\vdots \\ &\to (\boldsymbol{b}_1{}', \boldsymbol{b}_2{}', \cdots, \boldsymbol{b}_n{}').\end{aligned}$$

At each step, assuming the i-th subject is from the g-th group, an update is simulated from a p-dimensional multivariate normal distribution,

$$\begin{aligned}\boldsymbol{b}_i{}' \sim\ & q(\boldsymbol{b}_i|\boldsymbol{\alpha}, \boldsymbol{\beta}_g, \boldsymbol{\gamma}_g, \sigma^2, \lambda_D, \boldsymbol{y}_i) \\ \sim\ & N_p\Bigg(V_3^{-1}\frac{1}{\sigma^2}\Bigg[\frac{1}{n_i}Z_i^T X_i(\boldsymbol{\gamma}_g)\Big(\sum_{j\in g}\frac{1}{n_j}X_j^T(\boldsymbol{\gamma}_g)X_j(\boldsymbol{\gamma}_g)\Big)^{-1}\Big(\sum_{\substack{j\in g\\ j\neq i}}\frac{1}{n_j}X_j^T(\boldsymbol{\gamma}_g)\boldsymbol{\phi}_j + \frac{1}{n_i}X_i^T(\boldsymbol{\gamma}_g)\boldsymbol{\phi}_i\Big) \\ & + Z_i^T(\boldsymbol{y}_i - W_i\boldsymbol{\alpha} - (1+\frac{1}{n_i})X_i(\boldsymbol{\gamma}_g)\boldsymbol{\beta}_g(\boldsymbol{\gamma}_g))\Bigg], V_3^{-1}\Bigg),\end{aligned} \tag{4.10}$$

where $V_3 = \frac{1}{\sigma^2}Z_i^T Z_i + \lambda_D I + \frac{1}{\sigma^2}\frac{1}{n_i}Z_i^T X_i(\boldsymbol{\gamma}_g)(\sum_{j\in g}\frac{1}{n_j}X_j^T(\boldsymbol{\gamma}_g)X_j(\boldsymbol{\gamma}_g))^{-1}X_i^T(\boldsymbol{\gamma}_g)Z_i$.

Step 6. Consider updating λ_D using a Gibbs sampler. At each step, an update is simulated from a Gamma distribution,

$$\begin{aligned}\lambda_D{}' \sim\ & q(\lambda_D|b) \\ \sim\ & \Gamma(\frac{np}{2} + d_1,\ \frac{1}{2}\sum_{g=1}^{G}\sum_{i\in g}\boldsymbol{b}_i^T\boldsymbol{b}_i + d2).\end{aligned} \tag{4.11}$$

The posterior inference on model parameters can be estimated from the MCMC samples. Models with high posterior probability can be identified as those appearing most often in the MCMC output. One posterior quantity of interest is the marginal inclusion probability for each treatment effect, i.e. $1 - p(\boldsymbol{\gamma}_g = \mathbf{0}|y)$, $g = 1, \cdots, G$, which can be calculated using the proportion of draws in which $\boldsymbol{\gamma}_g$ is non-zero.

4.2.3 Stopping Criterion

Determining how long to run an MCMC simulation is critical to performing legitimate posterior inference. Premature termination often runs the risk of getting inaccurate

estimates. The relative standard deviation fixed-width stopping rule (FWSR) (see e.g. Flegal and Gong, 2015; Gong and Flegal, 2015) is implemented to terminate the MCMC simulation. It is a member of the FWSR family (for e.g. see Flegal and Gong, 2015; Flegal et al., 2008; Jones et al., 2006). The relative standard deviation FWSR is proved to be theoretically valid that: (1) it terminates a simulation w.p. 1, and (2) the resulting confidence interval achieves the nominal coverage probability. Moreover, it automates the stopping procedure for practitioners, and is shown to outperform convergence diagnostics using various numerical studies. Interested readers are directed to their papers for more details.

In short, the relative standard deviation FWSR terminates the simulation when the computational uncertainty is relatively small to the posterior uncertainty. Specifically, it controls the width of a confidence interval from a Markov chain central limit theorem (CLT) through a threshold ϵ and significant level δ. Gong and Flegal (2015) also establish a connection between the standard deviation FWSR and using effective sample size (ESS) as a stopping criteria, i.e. $K = 4z^2_{\delta/2}/\epsilon^2$, where K is the number of effective samples and $z_{\delta/2}$ is a critical value from the standard Normal distribution. Based on this connection, for instance, setting $\epsilon = 0.124$ and $\delta = 0.05$ in the relative standard deviation FWSR is equivalent to terminate the simulation when an ESS reaches $K = 1000$.

4.2.4 Simulation Study

We report the results of a simulation study undertaken to validate the model and estimation procedure. The simulated dataset consists of a control group and five treatment groups. The control group is simulated based on estimated parameters from a fitted linear mixed model on the experimental control group. That is, denote $Y_{i,t,99}$ as the weight of mouse $i \in \{1, \cdots, 297\}$ from the control group (Diet No.99) taken at time $t \in \{365, 395, 456, 517, 578, 639, 700, 760, 821, 882, 943, 1004, 1065, 1125, 1186\}$ corresponding to days on diet, we consider the following linear mixed model based on (4.2),

$$Y_{i,t,99} = \alpha_0 + \alpha_1 t + b_{0,i} + b_{1,i} t + \epsilon_{i,t}, \quad \epsilon_{i,t} \sim N(0, \sigma^2), \tag{4.12}$$

where α_0 and α_1 are the global intercept and slope, $b_{0,i}$ and $b_{1,i}$ are the subject specific random effects, where $\boldsymbol{b}_i = (b_{0,i}, b_{1,i})^T \sim N_2(\mathbf{0}, \lambda_D^{-1} I)$, and $\epsilon_{i,t}$ is the measurement error. Note that, as mentioned, the β's in (4.2) are set to zero for the control group to serve as

the baseline model.

The parameter estimation of (4.12) was carried out using the lmer() function in the R package `lme4` (Bates et al., 2012). Notice the time was rescaled using $t = (t-365)/365$ prior to model fitting. The maximum likelihood estimates (MLEs) are $\boldsymbol{\alpha} = (45.49, -5.75)^T$ and $\sigma^2 = 5.06$ and we set $\lambda_D^{-1} = 1.0$. Based on these parameter estimates, we simulated 297 subjects from (4.12) as the control group.

We then simulated five treatment groups, each with 36 subjects, by adding $\beta_{g,1}$'s to (4.12), while keeping other settings the same as for the simulated control group,

$$Y_{i,t,g} = \alpha_0 + \alpha_1 t + \beta_{g,1} t + b_{0,i} + b_{1,i} t + \epsilon_{i,t}, \quad g = 1, \cdots, 5, \tag{4.13}$$

where $\beta_{g,1} \in \{-2.0, -0.5, 0.0, 0.5, 2.0\}$ for each group. To be consistent with the experimental settings, we artificially differentiate the slope of each treatment group by $\alpha_1 + \beta_{g,1}$, but maintained the same global intercept α_0, since all mice are on the same diet at $t = 0$. Note that, we did not incorporate the "die-off" mechanism from the experiment into the simulation. Figure 4.3 shows the rescaled days on diet versus mean weight trajectories for this simulated dataset.

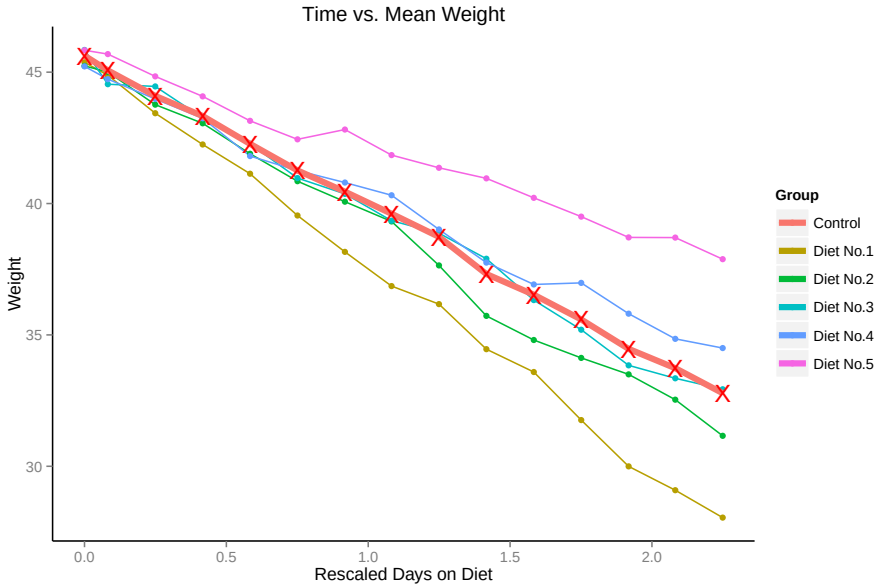

Figure 4.3: Combined plot of rescaled days on diet versus mean weight for 5 treatment groups and the control group (Diet No.99).

We followed the prior specification outlined in Section 4.2.1. The hyper-parameters d_1, d_2 were set to $d_1 = 0.001, d_2 = 0.001$ for the prior on λ_D to be vague. The hyper-parameters for $\boldsymbol{\alpha}|\boldsymbol{d}_3, d_4$ were set using estimates obtained from a fitted linear mixed model on the control group. The prior marginal inclusion probabilities for the treatment groups $\pi_g's$ were set to 0.5 for equal probability between inclusion and exclusion.

The component-wise Gibbs sampler was run as described in Section 4.2.2. The simulation was terminated by the relative standard deviation FWSR with the tuning parameters $\epsilon = 0.124$ and $\delta = 0.05$. It resulted in 16385 iterations with an effective sample size of at least 1000 for estimation of the posterior mean of all parameters. Figure 4.4 shows that the resulting MCMC outputs for variance of the random effects λ_D^{-1} and measurement errors σ^2 are close to the true values. Table 4.1 contains the posterior means (PM), 95% credible intervals and inclusion probabilities for the fixed effects coefficients, along with the LMM estimates from `lme4`.

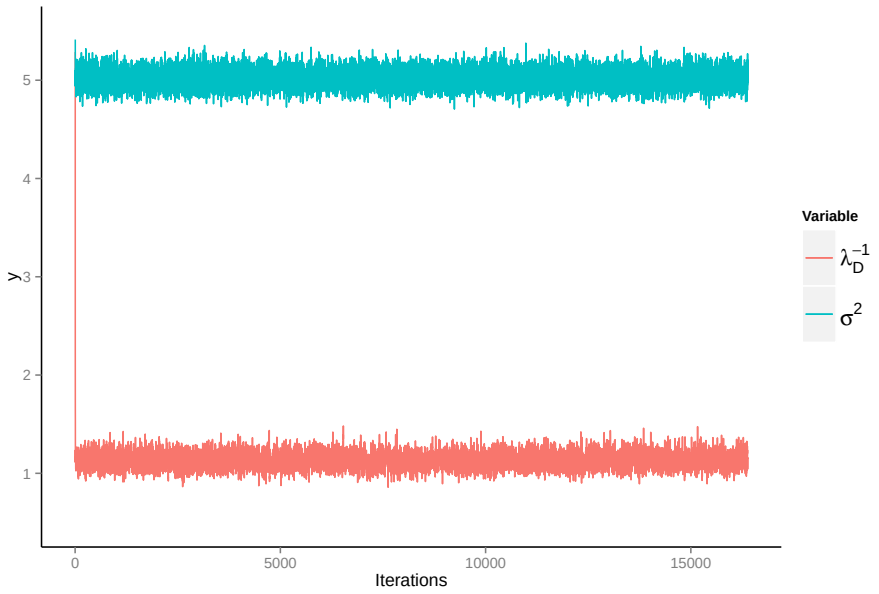

Figure 4.4: Gibbs sampler for variance terms λ_D^{-1} and σ^2 in the simulation study.

We compare our results to estimates from the LMM approach (see e.g. Fitzmaurice et al., 2004), as it is widely used to model such problems. Researchers often combine LMM with certain model selection criteria, e.g. Bayesian information criterion (BIC), to

Parameter	Truth	LMM	95% CI	PM	95% CI	$\Pr(\gamma_{g,j} = 1\|y)$
α_0	45.50	45.570	(45.431, 45.570)	45.591	(45.484, 45.700)	
α_1	-5.75	-5.708	(-5.852, -5.565)	-5.716	(-5.822, -5.612)	
$\beta_{1,1}$	-2.00	-2.130	(-2.546, -1.713)	-2.126	(-2.562, -1.685)	0.992(6.10e-5)
$\beta_{2,1}$	-0.50	-0.693	(-1.109, -0.276)	-0.698	(-1.126, -0.267)	0.983(7.44e-4)
$\beta_{3,1}$	0.00	-0.092	(-0.508, 0.325)	-0.093	(-0.518, 0.341)	0.442(3.88e-3)
$\beta_{4,1}$	0.50	0.708	(0.292, 1.125)	0.708	(0.283, 1.136)	0.987(5.74e-4)
$\beta_{5,1}$	2.00	2.266	(1.849, 2.683)	2.268	(1.830, 2.695)	0.992(6.10e-5)

Table 4.1: Fixed-effects estimates for the simulated dataset. Notice Column 4 contains the 95% confidence intervals from the LMM, Column 6 contains the 95% credible interval from the Bayesian model and Column 7 contains the marginal inclusion probability with standard error in the parenthesis.

determine which treatment are significantly differ from the control (see e.g. Spindler et al., 2013a). Despite that two approaches result in quite comparable parameter estimates, our approach introduces the probability of inclusion for each treatment group that is vital to straightforward interpretation and correct ranking of the true models, which remains challenging for the current frequentist method.

The sensitivity to the prior marginal inclusion probabilities was also evaluated by repeating the simulation with π_g's set to ranging from 0.3 to 0.7. We found no difference in model ranking, although the parameter estimates in Table 4.1 were slightly different. Other simulation settings showed comparable parameter estimations between our method and the LMM approach, and correctness in model ranking, although the results were not shown here.

4.3 Application

In this section, we use the methodology detailed in Section 4.2 to analyze the mouse weight data (see Section 1.1). Out of the 56 treatment groups in the original study, we limited our attention to 18 pre-screened treatments that the researchers are most interested in, as well as the control diet (Diet No.99). Therefore, one would expect most of these diets are significantly different from the control diet. For simplicity, we denote these 19 diets as $\mathcal{G} = \{21, 22, 23, 24, 27, 28, 29, 34, 35, 39, 42, 43, 44, 45, 48, 53, 55, 63, 99\}$. Similar to Section 4.2.4, the days on diet were rescaled prior to analysis. Figure 4.5 shows the rescaled

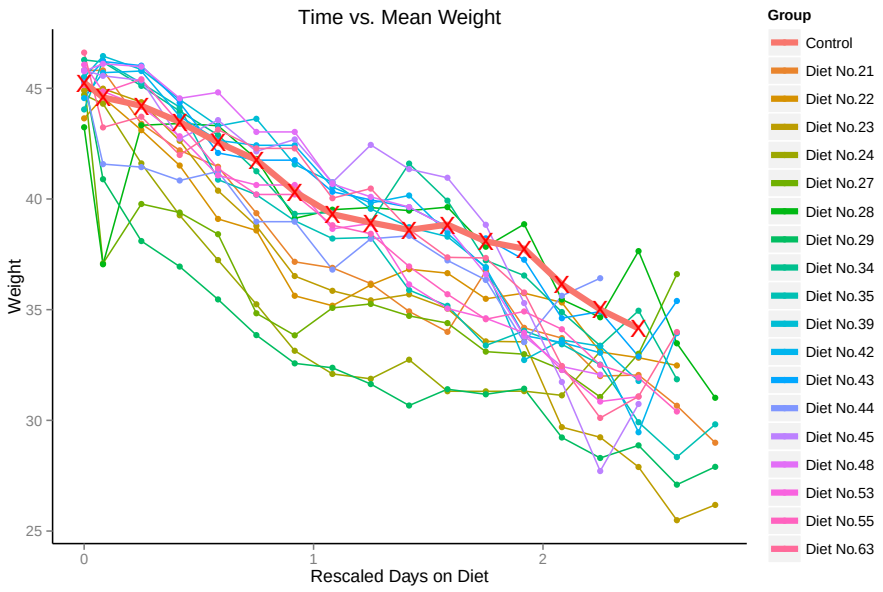

Figure 4.5: Combined plot of rescaled days on diet versus mean weight for 18 selected treatment diets and the control diet.

days on diet versus mean weight trajectories for the 18 diet groups.

Since every subjects start with the same control diet, it is reasonable to assume the same intercept for all groups. The individual weight trajectories suggest that, unlike the simulated dataset, a quadratic term is needed to characterize the trajectories. Specifically, we re-write the LMM from (4.3) as

$$Y_{i,t,g} = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \beta_{g,1}(\gamma_{g,1})t + \beta_{g,2}(\gamma_{g,2})t^2 + b_{0,i} + b_{1,i}t + b_{2,i}t^2 + \epsilon_{i,t}, \quad g \in \mathcal{G}. \quad (4.14)$$

Priors were specified as in the simulation study and $\boldsymbol{\gamma}_{99}$ was set to $\boldsymbol{\gamma}_{99} = (0, \cdots, 0)^T$. The component-wise Gibbs sampler was terminated by the relative standard deviation FWSR with $\epsilon = 0.124$ and $\delta = 0.05$, resulting in 115792 iterations with at least 1000 effective samples for estimation of posterior mean of parameters related to fixed-effects and variance components. Figure 4.6 shows the resulting MCMC outputs for variance of the random effects λ_D^{-1} and the measurement errors σ^2. The estimates indicate that, as expected, the variation among subjects outweighs it of the measurement errors. Table 4.2 presents the posterior means (PM), 95% credible intervals and inclusion probabilities for the fixed effects coefficients, along with the LMM estimates, for the experimental dataset. Figure 4.7 shows

the fitted weight trajectories and and Figure 4.8 illustrates pairwise comparisons between the control group and the 18 treatment groups. The results from the proposed method are comparable to the LMM results returned by the R package `lme4` in terms of point and interval estimates. However, the marginal inclusion probability provides a direct measure of the significance for each diet, which was unavailable in previous investigations using the LMM approach (see e.g. Spindler et al., 2014a,b, 2013a,b, 2014c).

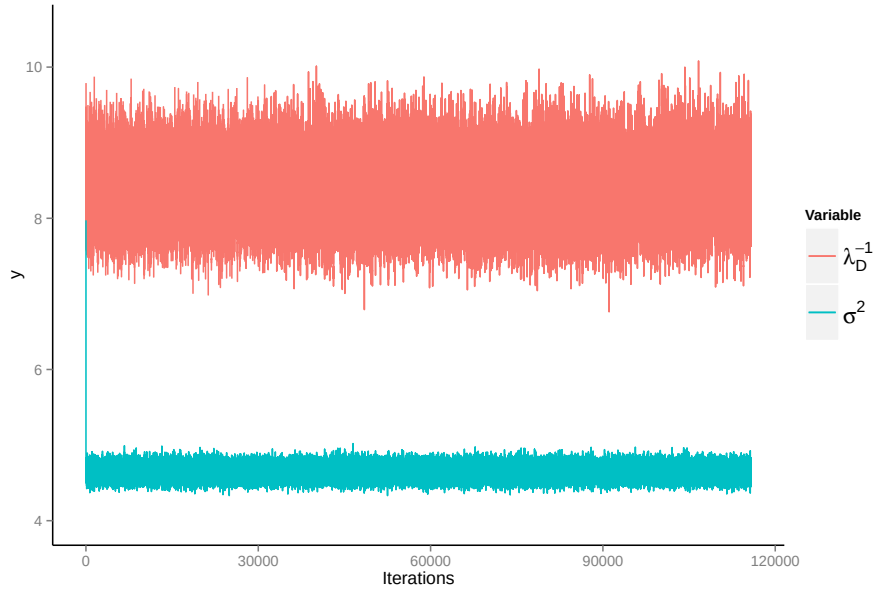

Figure 4.6: Gibbs sampler for variance terms λ_D^{-1} and σ^2 in the experimental application.

4.4 Discussion

This chapter proposes a novel method for Bayesian variable selection in LMM to compare multiple treatments with a control. It is a generalization of the SSVS approach, and it relies on a modification of the fractional prior proposed by Smith and Kohn (1997) and a component-wise Gibbs sampler. It provides practitioners with a framework to incorporate prior knowledge on different treatments and an intuitive evaluation of the significance of each treatment.

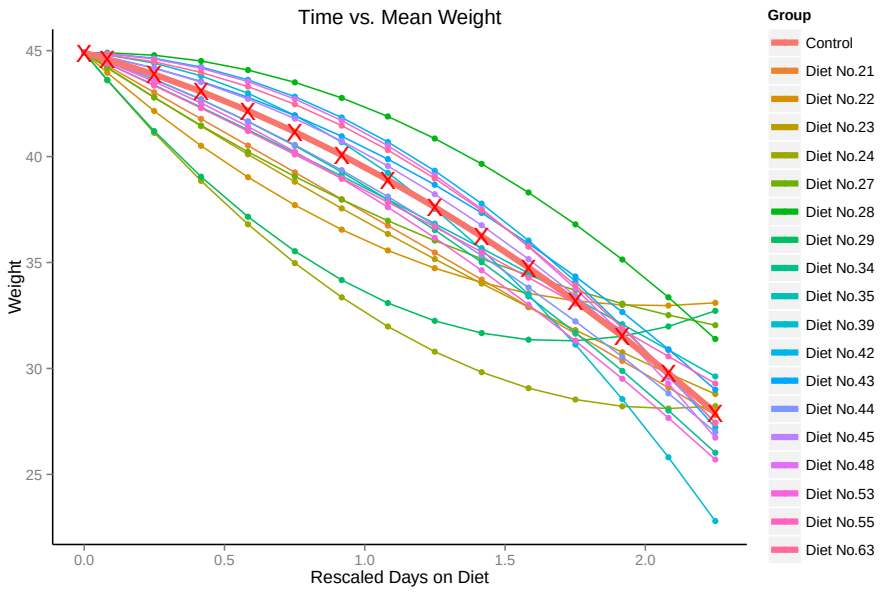

Figure 4.7: Estimated weight trajectories for 18 selected treatment groups and the control group based on the proposed model.

The proposed method is advantageous in that multiple treatments are compared with a control group simultaneously. In addition, the Bayesian framework introduces marginal posterior inclusion probabilities for each group, along with model-averaged coefficient estimates. Posterior model probabilities allow direct comparison among models, which is difficult to do using alternative frequentist approaches. In this chapter, our method is applied to a longitudinal study of a biochemical experiment, and is shown to perform well through simulated settings. Moreover, the method is quite general and have a wide range of potential applications in fields such as medicine and biology.

We emphasize on careful posterior inference with MCMC methodology. A major challenge for practitioners is determining how long to run an MCMC simulation. While some simulations are so complex that a fixed time approach is the only practical one, this is not so for most experiments. We advocate the use of relative standard deviation FWSR (Flegal and Gong, 2015; Gong and Flegal, 2015), since it is proved to be easy to use, theoretically valid and superior to using convergence diagnostics as a stopping criteria (Flegal et al., 2008; Jones et al., 2006).

4.5 Proofs and Calculations

Full conditional posterior distributions are derived from (4.5), for $i = 1, \cdots, n$, $g = 1, \cdots, G$, and $j = 0, \cdots, p-1$. To calculate the full conditional posterior $q(\gamma_{g,j}|\boldsymbol{\alpha}, \boldsymbol{\gamma}_{-(g,j)}, b, \sigma^2, y)$, we integrate out β in (4.5) as Smith and Kohn (1996)

$$\begin{aligned} q(\boldsymbol{\alpha}, \gamma, b, \sigma^2, \lambda_D|y) &= \int q(\boldsymbol{\alpha}, \beta, \gamma, b, \sigma^2, \lambda_D|y) d\beta \\ &\propto \left[\prod_{g=1}^{G}\left[\prod_{i\in g}\pi(\boldsymbol{b}_i|\lambda_D)\right]\pi(\boldsymbol{\gamma}_g)\right]\pi(\boldsymbol{\alpha})\pi(\lambda_D)\pi(\sigma^2) \\ &\quad \times \prod_{g=1}^{G}\int_{\boldsymbol{\beta}_g}\prod_{i\in g} p(\boldsymbol{y}_i|\boldsymbol{\alpha}, \boldsymbol{\beta}_g, \boldsymbol{\gamma}_g, \boldsymbol{b}_i, \lambda_D, \sigma^2)\pi(\boldsymbol{\beta}_g|\boldsymbol{\alpha}, \boldsymbol{\gamma}_g, b, \sigma^2) d\boldsymbol{\beta}_g \end{aligned} \tag{4.15}$$

To calculate (4.15), define $\boldsymbol{\phi}_i = \boldsymbol{y}_i - W_i\boldsymbol{\alpha} - Z_i\boldsymbol{b}_i$. For a given g, consider

$$\begin{aligned} &\int_{\boldsymbol{\beta}_g}\prod_{i\in g} p(\boldsymbol{y}_i|\boldsymbol{\alpha}, \boldsymbol{\beta}_g, \boldsymbol{\gamma}_g, \boldsymbol{b}_i, \lambda_D, \sigma^2)\pi(\boldsymbol{\beta}_g|\boldsymbol{\alpha}, \boldsymbol{\gamma}_g, b, \sigma^2) d\boldsymbol{\beta}_g \\ &\propto \int_{\boldsymbol{\beta}_g}\left[\prod_{i\in g}\sigma^{-n_i}\exp\{-\frac{1}{2\sigma^2}(\boldsymbol{\phi}_i - X_i(\boldsymbol{\gamma}_g)\boldsymbol{\beta}_g(\boldsymbol{\gamma}_g))^T(\boldsymbol{\phi}_i - X_i(\boldsymbol{\gamma}_g)\boldsymbol{\beta}_g(\boldsymbol{\gamma}_g))\}\right] \\ &\quad \times |\frac{1}{\sigma^2}\sum_{i\in g}\frac{1}{n_i}X_i^T(\boldsymbol{\gamma}_g)X_i(\boldsymbol{\gamma}_g)|^{\frac{1}{2}} \times (2\pi)^{-\frac{1}{2}\sum_{j=0}^{p-1}\gamma_{g,j}} \\ &\quad \times \exp\left\{-\frac{1}{2}\left[\boldsymbol{\beta}_g(\boldsymbol{\gamma}_g) - (\sum_{i\in g}\frac{1}{n_i}X_i^T(\boldsymbol{\gamma}_g)X_i(\boldsymbol{\gamma}_g))^{-1}(\sum_{i\in g}\frac{1}{n_i}X_i^T(\boldsymbol{\gamma}_g)\boldsymbol{\phi}_i)\right]^T\left(\frac{1}{\sigma^2}\sum_{i\in g}X_i^T(\boldsymbol{\gamma}_g)X_i(\boldsymbol{\gamma}_g)\right)\right. \\ &\qquad \left.\left[\boldsymbol{\beta}_g(\boldsymbol{\gamma}_g) - (\sum_{i\in g}\frac{1}{n_i}X_i^T(\boldsymbol{\gamma}_g)X_i(\boldsymbol{\gamma}_g))^{-1}(\sum_{i\in g}\frac{1}{n_i}X_i^T(\boldsymbol{\gamma}_g)\boldsymbol{\phi}_i)\right]\right\} d\boldsymbol{\beta}_g \\ &= \sigma^{-\sum_{i\in g}n_i}\left(\frac{|\sum_{i\in g}\frac{1}{n_i}X_i^T(\boldsymbol{\gamma}_g)X_i(\boldsymbol{\gamma}_g)|}{|\sum_{i\in g}(1+\frac{1}{n_i})X_i^T(\boldsymbol{\gamma}_g)X_i(\boldsymbol{\gamma}_g)|}\right)^{\frac{1}{2}} \\ &\quad \times \exp\left\{-\frac{1}{2\sigma^2}\sum_{g=1}^{G}\left[\sum_{i\in g}\boldsymbol{\phi}_i^T\boldsymbol{\phi}_i + (\sum_{i\in g}\frac{1}{n_i}X_i^T(\boldsymbol{\gamma}_g)\boldsymbol{\phi}_i)^T(\sum_{i\in g}\frac{1}{n_i}X_i^T(\boldsymbol{\gamma}_g)X_i(\boldsymbol{\gamma}_g))^{-1}(\sum_{i\in g}\frac{1}{n_i}X_i^T(\boldsymbol{\gamma}_g)\boldsymbol{\phi}_i)\right.\right. \\ &\qquad \left.\left. - (\sum_{i\in g}(1+\frac{1}{n_i})X_i^T(\boldsymbol{\gamma}_g)\boldsymbol{\phi}_i)^T(\sum_{i\in g}(1+\frac{1}{n_i})X_i^T(\boldsymbol{\gamma}_g)X_i(\boldsymbol{\gamma}_g))^{-1}(\sum_{i\in g}(1+\frac{1}{n_i})X_i^T(\boldsymbol{\gamma}_g)\boldsymbol{\phi}_i)\right]\right\} \end{aligned}$$

Therefore, (4.15) is further simplified

$$
\begin{aligned}
&= \lambda_D^{np/2} \exp\{-\frac{\lambda_D}{2} \sum_{g=1}^{G} \sum_{i \in g} \boldsymbol{b}_i^T \boldsymbol{b}_i\} \prod_{g=1}^{G} \prod_{j=1}^{p} \pi_g^{\gamma_{g,j}} (1-\pi_g)^{1-\gamma_{g,j}} \exp\{-\frac{1}{2}(\boldsymbol{\alpha}-\boldsymbol{d}_3)^T d_4 (\boldsymbol{\alpha}-\boldsymbol{d}_3)\} \\
&\quad \times \lambda_D^{d_1-1} \exp\{-d_2 \lambda_D\} \prod_{g=1}^{G} \sigma^{-\sum_{i \in g} n_i} \left(\frac{|\sum_{i \in g} \frac{1}{n_i} X_i^T(\boldsymbol{\gamma}_g) X_i(\boldsymbol{\gamma}_g)|}{|\sum_{i \in g} (1+\frac{1}{n_i}) X_i^T(\boldsymbol{\gamma}_g) X_i(\boldsymbol{\gamma}_g)|} \right)^{\frac{1}{2}} \\
&\quad \times \exp\left\{ -\frac{1}{2\sigma^2} \sum_{g=1}^{G} \left[\sum_{i \in g} \boldsymbol{\phi}_i^T \boldsymbol{\phi}_i + (\sum_{i \in g} \frac{1}{n_i} X_i^T(\boldsymbol{\gamma}_g) \boldsymbol{\phi}_i)^T (\sum_{i \in g} \frac{1}{n_i} X_i^T(\boldsymbol{\gamma}_g) X_i(\boldsymbol{\gamma}_g))^{-1} (\sum_{i \in g} \frac{1}{n_i} X_i^T(\boldsymbol{\gamma}_g) \boldsymbol{\phi}_i) \right.\right. \\
&\qquad \left.\left. - (\sum_{i \in g} (1+\frac{1}{n_i}) X_i^T(\boldsymbol{\gamma}_g) \boldsymbol{\phi}_i)^T (\sum_{i \in g} (1+\frac{1}{n_i}) X_i^T(\boldsymbol{\gamma}_g) X_i(\boldsymbol{\gamma}_g))^{-1} (\sum_{i \in g} (1+\frac{1}{n_i}) X_i^T(\boldsymbol{\gamma}_g) \boldsymbol{\phi}_i) \right] \right\}
\end{aligned}
\tag{4.16}
$$

Based on (4.16), the full posterior distribution is characterized by

$$
\begin{aligned}
&q(\gamma_{g,j} | \boldsymbol{\alpha}, \boldsymbol{\gamma}_{-(g,j)}, b, \sigma^2, y) \propto \pi_g^{\gamma_{g,j}} (1-\pi_g)^{1-\gamma_{g,j}} \left(\frac{|\sum_{i \in g} \frac{1}{n_i} X_i^T(\boldsymbol{\gamma}_g) X_i(\boldsymbol{\gamma}_g)|}{|\sum_{i \in g} (1+\frac{1}{n_i}) X_i^T(\boldsymbol{\gamma}_g) X_i(\boldsymbol{\gamma}_g)} \right)^{\frac{1}{2}} \\
&\times \exp\left\{ -\frac{1}{2\sigma^2} \left[\sum_{i \in g} \boldsymbol{\phi}_i^T \boldsymbol{\phi}_i + (\sum_{i \in g} \frac{1}{n_i} X_i^T(\boldsymbol{\gamma}_g) \boldsymbol{\phi}_i)^T (\sum_{i \in g} \frac{1}{n_i} X_i^T(\boldsymbol{\gamma}_g) X_i(\boldsymbol{\gamma}_g))^{-1} (\sum_{i \in g} \frac{1}{n_i} X_i^T(\boldsymbol{\gamma}_g) \boldsymbol{\phi}_i) \right.\right. \\
&\qquad \left.\left. - (\sum_{i \in g} (1+\frac{1}{n_i}) X_i^T(\boldsymbol{\gamma}_g) \boldsymbol{\phi}_i)^T (\sum_{i \in g} (1+\frac{1}{n_i}) X_i^T(\boldsymbol{\gamma}_g) X_i(\boldsymbol{\gamma}_g))^{-1} (\sum_{i \in g} (1+\frac{1}{n_i}) X_i^T(\boldsymbol{\gamma}_g) \boldsymbol{\phi}_i) \right] \right\},
\end{aligned}
\tag{4.17}
$$

where $\boldsymbol{\gamma}_{-(g,j)} = (\gamma_{g,0}, \cdots, \gamma_{g,j-1}, \gamma_{g,j+1}, \cdots, \gamma_{g,p-1})^T$.

Parameter	LMM	95% CI	PM	95% CI	$\Pr(\gamma_{g,j} = 1\|y)$
α_0	44.879	(44.638, 45.120)	44.903	(44.720, 45.086)	
α_1	-4.124	(-4.864, -3.384)	-3.687	(-4.134, -3.237)	
α_2	-0.992	(-1.444, -0.540)	-1.718	(-2.076, -1.369)	
$\beta_{21,1}$	-3.671	(-6.637, -0.705)	-3.800	(-6.202, -1.432)	0.991(2.13e-4)
$\beta_{21,2}$	1.244	(-0.553, 3.041)	1.667	(-0.151, 3.475)	0.809(1.15e-3)
$\beta_{22,1}$	-7.894	(-10.828, -4.959)	-8.069	(-10.353, -5.789)	0.996(1.22e-5)
$\beta_{22,2}$	4.429	(2.644, 6.214)	4.611	(2.780, 6.461)	0.996(8.63e-6)
$\beta_{23,1}$	-4.467	(-7.282, -1.652)	-4.902	(-6.995, -2.810)	0.996(1.73e-5)
$\beta_{23,2}$	1.819	(0.203, 3.435)	2.352	(0.638, 4.046)	0.970(4.75e-4)
$\beta_{24,1}$	-12.373	(-15.317, -9.428)	-12.453	(-14.795, -10.103)	0.996(8.64e-6)
$\beta_{24,2}$	5.349	(3.533, 7.165)	5.596	(3.752, 7.444)	0.996(8.64-e6)
$\beta_{27,1}$	-6.643	(-8.694, -4.592)	-5.135	(-6.671, -3.570)	0.996(8.64e-6)
$\beta_{27,2}$	3.173	(1.948, 4.397)	3.098	(1.872, 4.344)	0.996(8.64e-6)
$\beta_{28,1}$	2.717	(0.585, 4.849)	3.899	(2.227, 5.611)	0.997(1.50e-5)
$\beta_{28,2}$	-0.969	(-2.273, 0.335)	-1.047	(-2.423, 0.316)	0.747(1.27e-3)
$\beta_{29,1}$	-13.462	(-15.574, -11.350)	-12.333	(-13.946, -10.704)	0.996(8.64e-6)
$\beta_{29,2}$	6.499	(5.195, 7.803)	6.432	(5.097, 7.751)	0.996(8.64e-6)
$\beta_{34,1}$	-0.768	(-2.898, 1.363)	-0.892	(-2.441, 0.638)	0.663(1.39e-3)
$\beta_{34,2}$	0.351	(-1.002, 1.704)	0.022	(-1.315, 1.459)	0.592(1.44e-3)
$\beta_{35,1}$	-1.552	(-3.607, 0.503)	-2.355	(-3.911, -0.752)	0.983(3.33e-4)
$\beta_{35,2}$	0.563	(-0.670, 1.796)	1.384	(0.140, 2.657)	0.920(7.83e-4)
$\beta_{39,1}$	3.183	(0.955, 5.410)	2.699	(0.829, 4.532)	0.975(4.29e-4)
$\beta_{39,2}$	-2.422	(-3.877, -0.967)	-2.212	(-3.707, -0.659)	0.976(4.16e-4)
$\beta_{42,1}$	3.625	(1.560, 5.689)	3.478	(1.874, 5.054)	0.996(2.86e-5)
$\beta_{42,2}$	-2.047	(-3.298, -0.795)	-1.686	(-2.953, -0.423)	0.963(5.28e-4)
$\beta_{43,1}$	1.627	(-0.442, 3.697)	1.298	(-0.201, 2.902)	0.815(1.13e-3)
$\beta_{43,2}$	-0.659	(-1.910, 0.591)	-0.364	(-1.705, 0.929)	0.584(1.45e-3)
$\beta_{44,1}$	-2.230	(-4.532, 0.068)	-1.028	(-2.777, 0.626)	0.675(1.37e-3)
$\beta_{44,2}$	0.853	(-0.626, 2.331)	0.274	(-1.195, 1.889)	0.549(1.46e-3)
$\beta_{45,1}$	1.094	(-1.122, 3.310)	1.357	(-0.351, 3.196)	0.772(1.23e-3)
$\beta_{45,2}$	-0.160	(-1.631, 1.310)	-0.696	(-2.308, 0.875)	0.633(1.41e-3)
$\beta_{48,1}$	4.118	(2.032, 6.204)	3.361	(1.695, 5.014)	0.996(6.29e-5)
$\beta_{48,2}$	-2.268	(-3.541, -0.995)	-1.728	(-3.020, -0.439)	0.960(5.48e-4)
$\beta_{53,1}$	-1.138	(-3.242, 0.966)	-1.381	(-2.955, 0.185)	0.813(1.14e-3)
$\beta_{53,2}$	0.047	(-1.343, 1.249)	0.176	(-1.181, 1.574)	0.530(1.47e-3)
$\beta_{55,1}$	-1.714	(-3.800, 0.372)	-2.442	(-4.072, -0.812)	0.983(3.27e-4)
$\beta_{55,2}$	0.695	(-0.584, 1.975)	1.356	(0.061, 2.658)	0.894(8.92e-4)
$\beta_{63,1}$	-2.483	(-4.590, -0.376)	2.712	(1.052, 4.376)	0.991(2.07e-4)
$\beta_{63,2}$	-1.367	(-2.678, -0.056)	-1.305	(-2.580, 0.019)	0.863(1.00e-3)

Table 4.2: Fixed-effects estimates for the experimental dataset. Notice Column 4 contains the 95% confidence intervals from the LMM, Column 6 contains the 95% credible interval from the Bayesian model and Column 7 contains the marginal inclusion probability with standard error in the parenthesis.

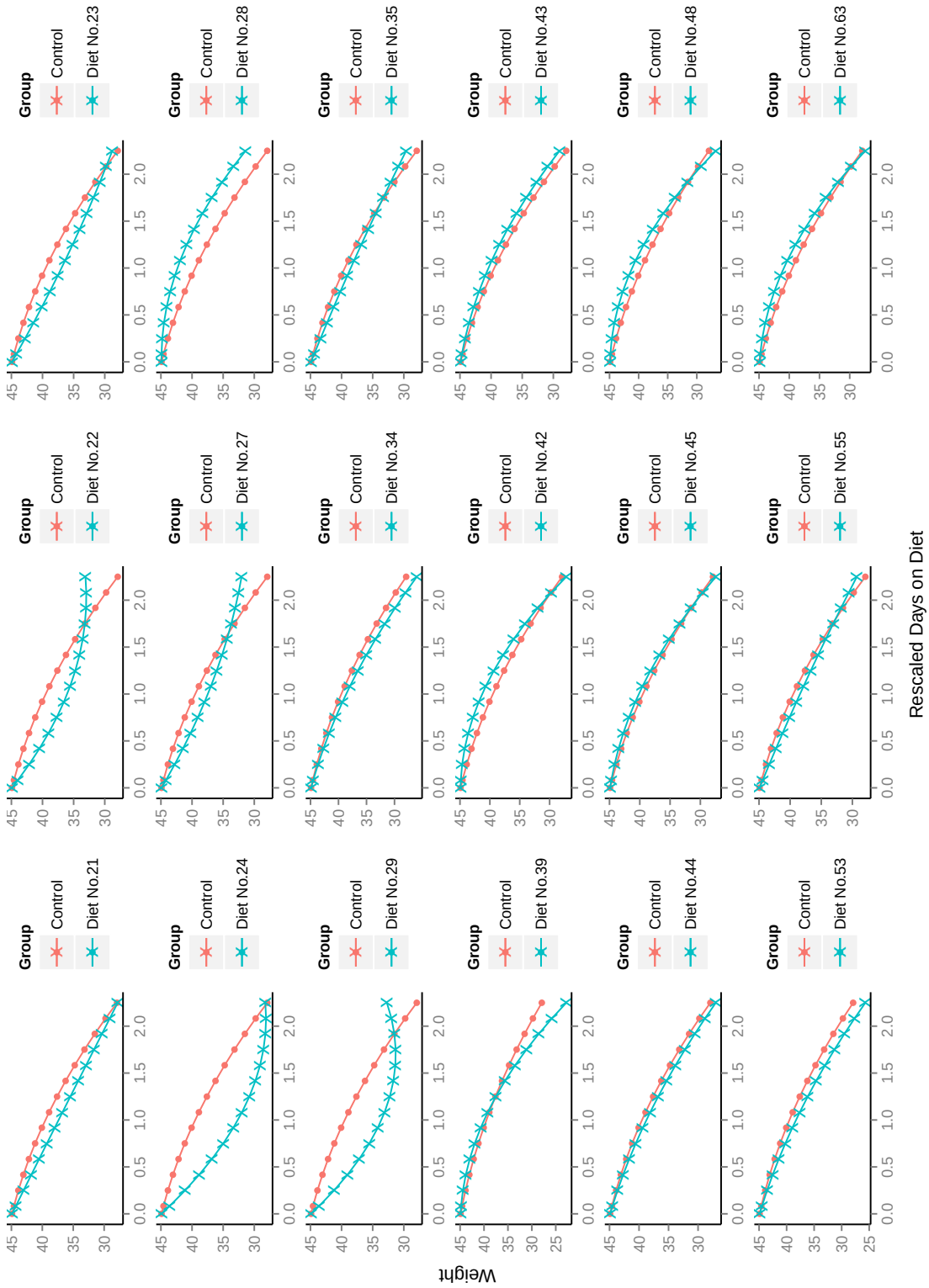

Figure 4.8: Pairwise comparisons between 18 selected treatment groups and the control group based on estimates from the proposed model.

Bibliography

Atkinson, Q. D., Gray, R. D., and Drummond, A. J. (2008). mtDNA variation predicts population size in humans and reveals a major Southern Asian chapter in human prehistory. *Molecular Biology and Evolution*, 25(2):468–474.

Banerjee, S., Gelfand, A. E., and Carlin, B. P. (2004). *Hierarchical modeling and analysis for spatial data*. CRC Press.

Bates, D., Maechler, M., and Bolker, B. (2012). lme4: Linear mixed-effects models using s4 classes.

Bednorz, W. and Latuszyński, K. (2007). A few remarks on 'Fixed-width output analysis for Markov chain Monte Carlo' by Jones et al. *Journal of the American Statistical Association*, 102:1485–1486.

Best, N., Cowles, M., and Vines, S. (1995). Coda manual version 0.30. *MRC Biostatistics Unit, Cambridge, UK*, 46:2020–2027.

Biesel, H. (1977). Recursive calculation of the standard deviation with increased accuracy. *Chromatographia*, 10(4):173–175.

Billingsley, P. (1999). *Convergence of Probability Measures*. Wiley, New York.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., and White, J.-S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution*, 24(3):127–135.

Boone, E. L., Merrick, J. R., and Krachey, M. J. (2014). A hellinger distance approach to mcmc diagnostics. *Journal of Statistical Computation and Simulation*, 84(4):833–849.

Brooks, S., Gelman, A., Jones, G., and Meng, X. (2010). *Handbook of Markov chain Monte Carlo: Methods and Applications*. Chapman & Hall.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7:434–455.

Brooks, S. P. and Roberts, G. O. (1999). On quantile estimation and Markov chain Monte Carlo convergence. *Biometrika*, 86:710–717.

Burnham, K. P. and Anderson, D. R. (2002). *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media.

Caffo, B. S., Jank, W., and Jones, G. L. (2005). Ascent-based Monte Carlo EM. *Journal of the Royal Statistical Society,* Series B, 67:235–251.

Carpenter, P. A., Just, M. A., Keller, T. A., Eddy, W. F., and Thulborn, K. R. (1999). Time course of fMRI-activation in language and spatial networks during sentence comprehension. *Neuroimage*, 10:216–224.

Ceperley, D., Chen, Y., Craiu, R. V., Meng, X.-L., Mira, A., and Rosenthal, J. (2012). Challenges and advances in high dimensional and high complexity monte carlo computation and theory. In *Proceedings of the Workshop at the Banff International Research Station for Mathematical Innovation Discovery*.

Chan, K. S. and Geyer, C. J. (1994). Comment on "Markov chains for exploring posterior distributions". *The Annals of Statistics*, 22:1747–1758.

Chen, Z. and Dunson, D. B. (2003). Random effects selection in linear mixed models. *Biometrics*, 59(4):762–769.

Clyde, M. and George, E. I. (2004). Model uncertainty. *Statistical Science*, pages 81–94.

Cowles, M. K. and Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91:883–904.

Cowles, M. K., Roberts, G. O., and Rosenthal, J. S. (1999). Possible biases induced by MCMC convergence diagnostics. *Journal of Statistical Computing and Simulation*, 64:87–104.

Doss, C., Flegal, J. M., Jones, G. L., and Neath, R. C. (2014). Markov chain Monte Carlo estimation of quantiles. *Electronic Journal of Statistics*, 8(2):2448–2478.

Douc, R., Fort, G., Moulines, E., and Soulier, P. (2004). Practical drift conditions for subgeometric rates of convergence. *The Annals of Applied Probability*, 14:1353–1377.

Drummond, A. J., Ho, S. Y., Phillips, M. J., and Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. *PLoS biology*, 4(5):e88.

Elith, J., Leathwick, J., and Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, 77(4):802–813.

Finley, A. O. and Banerjee, S. (2013). spBayes: Univariate and multivariate spatial modeling R package version 0.3-7. http://CRAN.R-project.org/package=spBayes.

Finley, A. O., Banerjee, S., and Gelfand, A. E. (2012). Bayesian dynamic modeling for large space-time datasets using gaussian predictive processes. *Journal of Geographical Systems*, 14(1):29–47.

Fitzmaurice, G., Laird, N., and Ware, J. (2004). *Applied longitudinal analysis*. Wiley series in probability and statistics. Wiley-Interscience.

Flegal, J. M. (2012). Applicability of subsampling bootstrap methods in Markov chain Monte Carlo. In Wozniakowski, H. and Plaskota, L., editors, *Monte Carlo and Quasi-Monte Carlo Methods 2010*, volume 23, pages 363–372. Springer Proceedings in Mathematics & Statistics.

Flegal, J. M. and Gong, L. (2015). Relative fixed-width stopping rules for Markov chain Monte Carlo simulations. *Statistica Sinica*, 25:655–676.

Flegal, J. M., Haran, M., and Jones, G. L. (2008). Markov chain Monte Carlo: Can we trust the third significant figure? *Statistical Science*, 23:250–260.

Flegal, J. M. and Hughes, J. (2012). mcmcse: Monte Carlo standard errors for MCMC R package version 1.0-1. http://cran.r-project.org/web/packages/mcmcse/index.html.

Flegal, J. M. and Jones, G. L. (2010). Batch means and spectral variance estimators in Markov chain Monte Carlo. *The Annals of Statistics*, 38:1034–1070.

Flegal, J. M. and Jones, G. L. (2011). Implementing Markov chain Monte Carlo: Estimating with confidence. In Brooks, S., Gelman, A., Jones, G., and Meng, X., editors, *Handbook of Markov chain Monte Carlo*, pages 175–197. Chapman & Hall/CRC Press.

Fort, G. and Moulines, E. (2000). V-subgeometric ergodicity for a Hastings-Metropolis algorithm. *Statistics and Probability Letters*, 49:401–410.

Fort, G. and Moulines, E. (2003). Polynomial ergodicity of Markov transition kernels. *Stochastic Processes and their Applications*, 103:57–99.

Fristedt, B. and Gray, L. F. (1997). *A Modern Approach to Probability Theory*. Birkhauser Verlag.

Friston, K., Ashburner, J., Frith, C. D., Poline, J.-B., Heather, J. D., Frackowiak, R. S., et al. (1995). Spatial registration and normalization of images. *Human brain mapping*, 3(3):165–189.

Friston, K., Fletcher, P., Josephs, O., Holmes, A., Rugg, M., and Turner, R. (1998). Event-related fMRI: characterizing differential responses. *Neuroimage*, 7(1):30–40.

Gelfand, A. E., Banerjee, S., and Gamerman, D. (2005). Spatial process modelling for univariate and multivariate dynamic spatial data. *Environmetrics*, 16(5):465–479.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85:398–409.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman & Hall/CRC, Boca Raton, second edition.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, 7:457–472.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6):721–741.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica sinica*, 7(2):339–373.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments (with discussion). In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., editors, *Bayesian Statistics 4. Proceedings of the Fourth Valencia International Meeting*, pages 169–188. Clarendon Press.

Geweke, J. et al. (1996). Variable selection and model comparison in regression. *Bayesian Statistics*, 5:609–620.

Geyer, C. J. (2011). Introduction to Markov chain Monte Carlo. In *Handbook of Markov Chain Monte Carlo*. CRC, London.

Geyer, C. J. and Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent. *Journal of the Royal Statistical Society,* Series B, 54:657–683.

Glover, G. H. (1999). Deconvolution of impulse response in event-related bold fMRI. *Neuroimage*, 9(4):416–429.

Glynn, P. and Whitt, W. (1988). Ordinary CLT and WLLN versions of $l = \lambda w$. *Mathematics of Operations Research*, pages 674–692.

Glynn, P. W. and Whitt, W. (1992). The asymptotic validity of sequential stopping rules for stochastic simulations. *The Annals of Applied Probability*, 2:180–198.

Gong, L. and Flegal, J. M. (2015). A practical sequential stopping rule for high-dimensional Markov chain Monte Carlo. *Journal of Computational and Graphical Statistics*, (just-accepted):00–00.

Gössl, C., Auer, D. P., and Fahrmeir, L. (2001). Bayesian spatiotemporal inference in functional magnetic resonance imaging. *Biometrics*, 57(2):554–562.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1).

Hijmans, R., Phillips, S., Leathwick, J., and Elith, J. (2010). dismo: species distribution modeling. R package version 0.5-4.

Hobert, J. P. (2011). The data augmentation algorithm: Theory and Methodology. In Brooks, S., Gelman, A., Jones, G., and Meng, X.-L., editors, *Handbook of Markov Chain Monte Carlo*. Chapman & Hall/CRC Press, London.

Hobert, J. P., Jones, G. L., Presnell, B., and Rosenthal, J. S. (2002). On the applicability of regenerative simulation in Markov chain Monte Carlo. *Biometrika*, 89:731–743.

Holmes, C., Denison, D., and Mallick, B. (2002). Bayesian model order determination and basis selection for seemingly unrelated regressions. *Journal of Computational and Graphical Statistics*, 11:533s551.

Huerta, G., Sansó, B., and Stroud, J. R. (2004). A spatiotemporal model for Mexico City ozone levels. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 53(2):231–248.

Ibragimov, I. A. (1962). Some limit theorems for stationary processes. *Theory of Probability and Its Applications*, 7:349–382.

Jarner, S. F. and Hansen, E. (2000). Geometric ergodicity of Metropolis algorithms. *Stochastic Processes and Their Applications*, 85:341–361.

Jarner, S. F. and Roberts, G. O. (2002). Polynomial convergence rates of Markov chains. *Annals of Applied Probability*, 12:224–247.

Jarner, S. F. and Roberts, G. O. (2007). Convergence of heavy-tailed Monte Carlo Markov chain algorithms. *Scandinvian Journal of Statistics*, 24:101–121.

Jarner, S. F. and Tweedie, R. L. (2003). Necessary conditions for geometric and polynomial ergodicity of random-walk-type Markov chains. *Bernoulli*, 9:559–578.

Jeske, D. R., Flegal, J. M., and Spindler, S. R. (2014). Minimum size survival analysis sampling plans for comparing multiple treatment groups to a single control group. *Communications in Statistics-Theory and Methods*, 43(13):2689–2701.

Johnson, A. A. and Jones, G. L. (2010). Gibbs sampling for a Bayesian hierarchical general linear model. *Electronic Journal of Statistics*, 4:313–333.

Johnson, A. A., Jones, G. L., Neath, R. C., et al. (2013). Component-wise Markov chain Monte Carlo: Uniform and geometric ergodicity under mixing and composition. *Statistical Science*, 28(3):360–375.

Jones, G. L. (2004). On the Markov chain central limit theorem. *Probability Surveys*, 1:299–320.

Jones, G. L., Haran, M., Caffo, B. S., and Neath, R. (2006). Fixed-width output analysis for Markov chain Monte Carlo. *Journal of the American Statistical Association*, 101:1537–1547.

Jones, G. L. and Hobert, J. P. (2001). Honest exploration of intractable probability distributions via Markov chain Monte Carlo. *Statistical Science*, 16:312–334.

Kass, R. E., Carlin, B. P., Gelman, A., and Neal, R. M. (1998). Markov chain Monte Carlo in practice: A roundtable discussion. *The American Statistician*, 52(2):93–100.

Keller, T., Just, M., and Stenger, V. (2001). Reading span and the time-course of cortical activation in sentence-picture verification. In *Annual Convention of the Psychonomic Society*.

Kim, T. Y. and Lee, S. (2005). Kernel density estimator for strong mixing processes. *Journal of Statistical Planning and Inference*, 133(2):273–284.

Kinney, S. K. and Dunson, D. B. (2007). Fixed and random effects selection in linear and logistic models. *Biometrics*, 63(3):690–698.

Kish, L. (1965). *Survey Sampling*. John Wiley and Sons.

Kohn, R., Smith, M., and Chan, D. (2001). Nonparametric regression using linear combinations of basis functions. *Statistics and Computing*, 11(4):313–322.

Kuo, L. and Mallick, B. (1998). Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 65–81.

Leathwick, J., Elith, J., Chadderton, W., Rowe, D., and Hastie, T. (2008). Dispersal, disturbance and the contrasting biogeographies of New Zealands diadromous and non-diadromous fish species. *Journal of Biogeography*, 35(8):1481–1497.

Lee, K.-J., Jones, G. L., Caffo, B. S., and Bassett, S. S. (2014). Spatial Bayesian variable selection models on functional magnetic resonance imaging time-series data. *Bayesian Analysis*, 9(3):699–732.

Lindquist, M. A. (2008). The statistical analysis of fMRI data. *Statistical Science*, 23(4):439–464.

Lindquist, M. A., Meng Loh, J., Atlas, L. Y., and Wager, T. D. (2009). Modeling the hemodynamic response function in fMRI: efficiency, bias and mis-modeling. *Neuroimage*, 45(1):S187–S198.

Liu, J. S., Chen, R., and Wong, W. H. (1998). Rejection control and sequential importance sampling. *Journal of the American Statistical Association*, 93(443):1022–1031.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall, London, second edition.

Mengersen, K. and Tweedie, R. L. (1996). Rates of convergence of the Hastings and Metropolis algorithms. *The Annals of Statistics*, 24:101–121.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equations of state calculations by fast computing machine. *Journal of Chemical Physics*, 21.

Meyn, S., Tweedie, R., and Glynn, P. (2009). *Markov chains and stochastic stability*, volume 2. Cambridge University Press Cambridge.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032.

Moller, J. and Waagepetersen, R. P. (2003). *Statistical inference and simulation for spatial point processes*. CRC Press.

Mykland, P., Tierney, L., and Yu, B. (1995). Regeneration in Markov chain samplers. *Journal of the American Statistical Association*, 90:233–241.

O'Hara, R. B., Sillanpää, M. J., et al. (2009). A review of Bayesian variable selection methods: what, how and which. *Bayesian Analysis*, 4(1):85–117.

Oodaira, H. and Yoshihara, K.-i. (1972). Functional central limit theorems for strictly stationary processes satisfying the strong mixing condition. *Kodai Mathematical Seminar Reports*, 24:259–269.

Pinheiro, J. and Bates, D. (2006). *Mixed-effects models in S and S-PLUS*. Springer Science & Business Media.

Raftery, A. E. and Lewis, S. M. (1992a). Comment on "The Gibbs sampler and Markov chain Monte Carlo". *Statistical Science*, 7:493–497.

Raftery, A. E. and Lewis, S. M. (1992b). How many iterations in the Gibbs sampler? In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., editors, *Bayesian Statistics 4. Proceedings of the Fourth Valencia International Meeting*, pages 763–773. Clarendon Press.

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer, New York, second edition.

Roberts, G. O. and Rosenthal, J. S. (2004). General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1:20–71.

Roberts, G. O. and Tweedie, R. L. (1996). Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, 83:95–110.

Sen, P. K. (1972). On the Bahadur representation of sample quantiles for sequences of ϕ-mixing random variables. *Journal of Multivariate Analysis*, 2:77–95.

Smith, B. J. (2005). Bayesian output analysis program (boa) version 1.1 users manual. *University of Iowa, Iowa City, IA*.

Smith, M. and Fahrmeir, L. (2007). Spatial Bayesian variable selection with application to functional magnetic resonance imaging. *Journal of the American Statistical Association*, 102(478):417–431.

Smith, M. and Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *Journal of Econometrics*, 75(2):317–343.

Smith, M. and Kohn, R. (1997). A Bayesian approach to nonparametric bivariate regression. *Journal of the American Statistical Association*, 92(440):1522–1535.

Spindler, S. R., Mote, P. L., and Flegal, J. M. (2014a). Dietary supplementation with lovaza and krill oil shortens the life span of long-lived f1 mice. *Age*, 36(3):1345–1352.

Spindler, S. R., Mote, P. L., and Flegal, J. M. (2014b). Lifespan effects of simple and complex nutraceutical combinations fed isocalorically to mice. *Age*, 36(2):705–718.

Spindler, S. R., Mote, P. L., Flegal, J. M., and Teter, B. (2013a). Influence on longevity of blueberry, cinnamon, green and black tea, pomegranate, sesame, curcumin, morin, pycnogenol, quercetin, and taxifolin fed iso-calorically to long-lived, f1 hybrid mice. *Rejuvenation research*, 16(2):143–151.

Spindler, S. R., Mote, P. L., Li, R., Dhahbi, J. M., Yamakawa, A., Flegal, J. M., Jeske, D. R., and Lublin, A. L. (2013b). β1-adrenergic receptor blockade extends the life span of drosophila and long-lived mice. *Age*, 35(6):2099–2109.

Spindler, S. R., Mote, P. L., Lublin, A. L., Flegal, J. M., Dhahbi, J. M., and Li, R. (2014c). Nordihydroguaiaretic acid extends the lifespan of drosophila and mice, increases mortality-related tumors and hemorrhagic diathesis, and alters energy homeostasis in mice. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, page glu190.

Stephens, P. A., Buskirk, S. W., Hayward, G. D., and Martinez Del Rio, C. (2005). Information theory and hypothesis testing: a call for pluralism. *Journal of Applied Ecology*, 42(1):4–12.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *The Annals of Statistics*, 22:1701–1762.

Wang, X. and Mitchell, T. (2002). Detecting cognitive states using machine learning. Technical report, CMU CALD Technical Report for Summer Work.

Woolrich, M. W., Jenkinson, M., Brady, J. M., and Smith, S. M. (2004). Fully Bayesian spatio-temporal modeling of fMRI data. *Medical Imaging, IEEE Transactions on*, 23(2):213–231.

Yu, B. (1993). Density estimation in the L^∞ norm for dependent data with applications to the Gibbs sampler. *The Annals of Statistics*, 21:711–735.

Zellner, A. (1986). On assessing prior distributions and bayesian regression analysis with g-prior distributions. *Bayesian inference and decision techniques: Essays in Honor of Bruno De Finetti*, 6:233–243.