# Output Analysis for Markov Chain Monte Carlo

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Dootika Vats

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Galin L. Jones, Adviser

February 2017

# Acknowledgements

I would like to thank my advisor Galin Jones for his continued support, advice, and guidance through the years. In spite of being given every reason to, he never lost his patience and continued to help make me a better researcher. I could never thank him enough.

I am grateful to my final exam committee members Dennis Cook, Adam Rothman, and Jim Hodges for their help and guidance through this process. I would also like to thank my preliminary exam committee members Charlie Geyer and John Hughes for their guidance and help with coding. I owe a debt to the excellent faculty at the School of Statistics for ensuring I had a statistical foundation before I explored the world of research. I owe an enormous debt to the staff for guiding me through paperwork and always being there to help out; thank you Kate and Taryn. I am also grateful to the Louise T. Dosdall Foundation and the School of Statistics alumni for supporting my dissertation research.

My time in the department was made exponentially better by the continued support and help of my fellow students and friends; thank you Aaron, Adam, Brad, Christina, Dan, Haema, Lindsey, Megan, Sakshi, Sijia, and Yang. I would also like to thank my friends back home in India for whom the time difference was only a hypothetical problem.

I will always be grateful to my brother for introducing me to statistics and to my sister-in-law whose passion for her work inspires me everyday; thank you for your support and encouragement. And finally, I owe every ounce of my past, present, and future success to my parents. They instilled in me the importance of hard work and education. I am an academic today because of them.

# Abstract

Markov chain Monte Carlo (MCMC) is a sampling method used to estimate expectations with respect to a target distribution. An important question is when should sampling stop so that we have good estimates of these expectations? The key to answering this question lies in assessing the Monte Carlo error through a multivariate Markov chain central limit theorem (CLT). The multivariate nature of this Monte Carlo error largely has been ignored in the MCMC literature. This dissertation discusses the drawbacks of the current univariate methods of terminating simulation and introduces a multivariate framework for terminating simulation. Theoretical properties of the procedures are established.

A multivariate effective sample size is defined and estimated using strongly consistent estimators of the covariance matrix in the Markov chain CLT, a property that is shown for the multivariate batch means estimator and the multivariate spectral variance estimator. A critical aspect of this procedure is that a lower bound on the number of effective samples required for a pre-specified level of precision can be determined *a priori*. This lower bound depends on the problem only in the dimension of the expectation being estimated, and not on the underlying stochastic process. This result is obtained by drawing a connection between terminating the simulation via effective sample size and terminating it using a relative standard deviation fixed-volume sequential stopping rule. The finite sample properties of the proposed methods are demonstrated in a variety of examples.

The proposed method requires the existence of a Markov chain CLT, establishing which requires bounding the rate of convergence of the Markov chains. We establish a geometric rate of convergence for a class of Bayesian penalized regression models. We

also present examples showing how it may be easier to establish rates of convergence for linchpin variable samplers than for its competitors.

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

Markov chain Monte Carlo (MCMC) algorithms are used to estimate expectations with respect to a distribution when obtaining independent samples is difficult. Typically, interest is in estimating a vector of quantities. However, analysis of MCMC output routinely focuses on inference about complicated joint distributions only through their marginals. This, despite the fact that the assumption of independence across components holds rarely in settings where MCMC is relevant. Thus standard univariate convergence diagnostics, sequential stopping rules for termination, effective sample size definitions, and confidence intervals all lead to an incomplete understanding of the estimation process. In this dissertation, we overcome the drawbacks of univariate analysis by developing a methodological framework for multivariate analysis of MCMC output.

This chapter introduces the problem of estimation in MCMC and discusses the state of the art methods for output analysis. As motivation, we present the following Bayesian logistic regression model.

**Example 1.1 (Bayesian Logistic Regression)**
For $i = 1, \ldots, K$, let $Y_i$ be a binary response variable and $X_i = (x_{i1}, x_{i2}, \ldots, x_{i5})$ be the observed vector of predictors for the $i$th observation. Assume $\tau^2 > 0$ is known

and let $I_5$ be the $5 \times 5$ identity matrix. A Bayesian logistic regression model is

$$Y_i \mid X_i, \beta \overset{ind}{\sim} \text{Bernoulli}\left(\frac{1}{1+e^{-X_i\beta}}\right), \quad \text{and} \quad \beta \sim N_5(0, \tau^2 I_5)\,. \tag{1.1}$$

This simple hierarchical model results in an intractable posterior distribution on $\mathbb{R}^5$. Let $y_1, y_2, \ldots, y_K$ be the observed realizations of the response. The posterior probability density function is,

$$\begin{aligned} f(\beta \mid y, X) &\propto f(\beta) \prod_{i=1}^{K} f(y_i \mid X_i, \beta) \\ &\propto \exp\left\{-\frac{\beta^T\beta}{2\tau^2}\right\} \prod_{i=1}^{K} \left(\frac{1}{1+e^{-X_i\beta}}\right)^{y_i} \left(\frac{e^{-X_i\beta}}{1+e^{-X_i\beta}}\right)^{1-y_i}. \end{aligned} \tag{1.2}$$

In (1.2) the proportionality sign indicates that the normalizing constant for the density $f(\beta \mid y, X)$ is unknown and intractable. This is an example of a scenario where MCMC methods are used to draw samples from $F$ and make inference on the regression coefficients. The posterior mean of $\beta$ might be a quantity of interest. □

In general, $F$ is a distribution with support $\mathcal{X}$, equipped with a countably generated $\sigma$-field $\mathcal{B}(\mathcal{X})$, and $g : \mathcal{X} \to \mathbb{R}^p$ is an $F$-integrable function such that $\theta := \mathrm{E}_F g$ is of interest. The dimension of $\theta$, $p$, indicates the number of quantities of interest and may be different from the dimension of $\mathcal{X}$.

Calculating $\theta$ can be difficult outside of well behaved distributions. As a result, MCMC methods are often used to estimate $\theta$. In MCMC, a Markov chain $\{X_t\}$ is constructed such that $F$ is its invariant distribution. If the Markov chain is aperiodic, irreducible, and Harris recurrent (see Section 2.1 for definitions), then the sample mean of the observed chain is a strongly consistent estimator of $\theta$. That is, as $n \to \infty$

$$\theta_n := \frac{1}{n}\sum_{t=1}^{n} g(X_t) \overset{a.s.}{\to} \theta\,, \tag{1.3}$$

where $\xrightarrow{a.s.}$ denotes almost sure convergence. The samples $X_1, X_2, \ldots, X_n$ thus obtained are correlated and are not exact draws from $F$. Even so, due to (1.3), estimation remains reliable since for a sufficiently large $n$, the estimates obtained will be close to the truth.

**Example 1.2 (Example 1.1 continued)**

We will use the Bayesian logistic regression model to analyze the `logit` dataset in the `mcmc` R package. We set $\tau^2 = 1$ for this data. The goal is to estimate the posterior mean of $\beta$, $\mathrm{E}_F\beta$. Thus $g$ here is the identity function mapping to $\mathbb{R}^5$, and $p = 5$. Since $f(\beta \mid y, X)$ is intractable,

$$\theta = \int_{\mathbb{R}^5} \beta\, f(\beta \mid y, X)\, d\beta$$

is also intractable and MCMC methods may be used to estimate $\theta$. We implement a random walk Metropolis-Hastings algorithm with a multivariate normal proposal distribution $N_5(\,\cdot\,, 0.35^2 I_5)$ where the 0.35 scaling ensures an optimal acceptance probability (see Roberts et al. (1997)). The starting value for $\beta$ is a random draw from the prior distribution. The complete algorithm is presented below.

---

Draw $\beta^{(0)} \sim N_5(0, I_5)$. Given $\beta^{(k-1)}$

1. Draw $\beta^* \sim N_5(\beta^{(k-1)}, 0.35^2 I_5)$ and $u \sim \text{Uniform}(0, 1)$.
2. If $u \leq \min\left\{1, f(\beta^* \mid y, X)/f(\beta^{(k-1)} \mid y, X)\right\}$ then set $\beta^{(k)} = \beta^*$
3. Otherwise, set $\beta^{(k)} = \beta^{(k-1)}$.
4. Set $k = k + 1$ and repeat until $k = n$.

---

It is well known that this sampler is Harris ergodic and thus Monte Carlo averages are strongly consistent. We obtain $10^5$ Monte Carlo samples and in Figure 1.1 plot the running average for the intercept $\beta_0$. As the Monte Carlo sample size increases, the

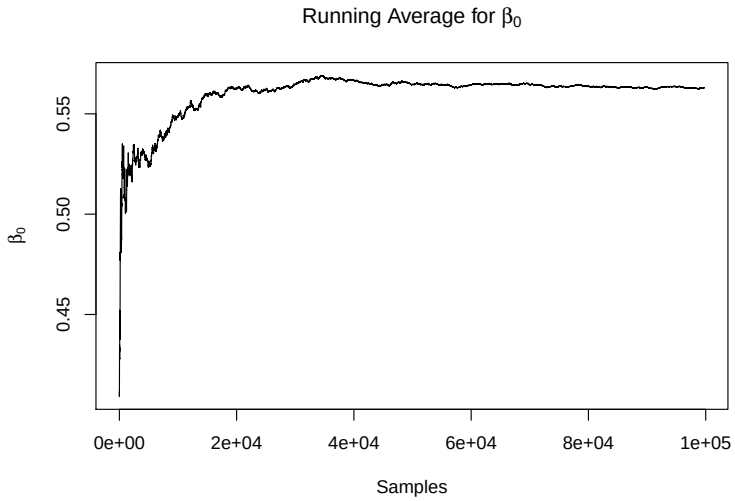



Figure 1.1: Running estimate for $\mathrm{E}_F\beta_0$ for a Monte Carlo sample of size $10^5$.

estimate for $\mathrm{E}_F\beta_0$ converges. However, the choice of Monte Carlo size $10^5$ was made arbitrarily and it is unclear if more samples are required to ensure good estimation. That is, there is no assessment of the error in our estimation of $\mathrm{E}_F\beta_0$. □

Finite sampling leads to an unknown *Monte Carlo error*, $\theta_n - \theta$. Estimating this Monte Carlo error is essential to assessing the quality of estimation for $\theta$. Under certain conditions (see Section 2.1) an approximate sampling distribution for the Monte Carlo error is available via a Markov chain central limit theorem (CLT). That is, there exists a $p \times p$ positive definite matrix, $\Sigma$, such that as $n \to \infty$,

$$\sqrt{n}(\theta_n - \theta) \xrightarrow{d} N_p(0, \Sigma) \,, \tag{1.4}$$

where $\xrightarrow{d}$ denotes convergence in distribution. Thus the CLT describes the asymptotic behavior of the Monte Carlo error and the strong law for $\theta_n$ ensures that large $n$ leads to a small Monte Carlo error. But how large is large enough?

Univariate methods have motivated most of the work in answering this question.

That is, instead of studying the joint asymptotic distribution as in (1.4), focus is on the marginal asymptotic distributions for individual components of $\theta_n - \theta$. Although useful in understanding marginal Monte Carlo error, univariate methods ignore the dependence between the components of $\theta_n$. We present a multivariate framework for assessing the Monte Carlo error and develop termination rules that learn from this multivariate structure. We first present the current state-of-the-art univariate methods used for determining Monte Carlo error and simulation termination.

## 1.1 Univariate Termination Rules

Let $\theta_{n,i}$ and $\theta_i$ be the $i$th components of $\theta_n$ and $\theta$, respectively and let $\sigma_i^2$ be the $i$th diagonal element of $\Sigma$. Note that for $i = 1, \ldots, p$, $\theta_i = \mathrm{E}_F g_i$ where $g = (g_1, g_2, \ldots, g_p)$. A univariate Markov chain CLT holds if for each $i = 1, \ldots, p$, as $n \to \infty$

$$\sqrt{n}(\theta_{n,i} - \theta_i) \xrightarrow{d} N(0, \sigma_i^2) \; . \tag{1.5}$$

Assessing the quality of estimation of $\theta_i$ requires estimating $\sigma_i^2$. This is a challenging problem since $\sigma_i^2 \neq \mathrm{Var}_F \, g_i(X_1)$, and in fact due to the serial correlation in the Markov chain,

$$\sigma_i^2 = \mathrm{Var}_F \, g_i(X_1) + 2 \sum_{k=1}^{\infty} \mathrm{Cov}_F \left( g_i(X_1), g_i(X_{1+k}) \right) \; . \tag{1.6}$$

Significant effort has gone into estimating $\sigma_i^2$. Geyer (1992) proposed the conservative initial sequence estimator. Jones et al. (2006) prove conditions under which the batch means estimator of $\sigma_i^2$ is strongly consistent and Flegal and Jones (2010) provide conditions under which spectral variance estimators are strongly consistent. Flegal and Jones (2010) also showed mean square consistency for the batch means

and spectral variance estimators.

Many output analysis tools that rely on (1.5) have been developed in MCMC (see Atchadé (2011), Atchadé (2016), Flegal and Jones (2010), Flegal and Gong (2015), Gelman and Rubin (1992), Gong and Flegal (2016), and Jones et al. (2006)). We specifically focus on two of these methods: terminating via effective sample size and fixed-width sequential stopping rule.

### 1.1.1 Univariate Effective Sample Size

Effective sample size (ESS) for estimating $\mathrm{E}_F g_i$ is the number of equivalent independent and identically distributed (i.i.d) samples required to attain the same standard error as the correlated sample. It is standard to stop simulation when the number of effective samples for each component reaches a pre-specified lower bound (see Atkinson et al. (2008), Drummond et al. (2006), Giordano et al. (2015), and Kruschke (2014) for a few examples).

Before defining effective sample size formally, we present some preliminary definitions. The autocorrelation for the $i$th component at lag $k$ is defined as

$$\rho_i(k) = \frac{\mathrm{Cov}_F(g_i(X_1), g_i(X_{1+k}))}{\mathrm{Var}_F(g_i(X_1))}\,.$$

Notice that by (1.6),

$$\sigma_i^2 = \mathrm{Var}_F(g_i(X_1)) \left(1 + 2\sum_{k=1}^{\infty} \rho_i(k)\right)\,. \tag{1.7}$$

Thus, higher autocorrelations lead to a larger asymptotic variance. Figure 1.2 shows the estimated autocorrelation function (ACF) plot for $\beta_0$ in the Bayesian logistic regression model. The significant lag autocorrelations contribute to the size of $\sigma_i^2$.

Let $\Lambda = \mathrm{Var}_F(g(X_1))$ and $\lambda_i^2 = \mathrm{Var}_F(g_i(X_1))$. Gong and Flegal (2016) define ESS

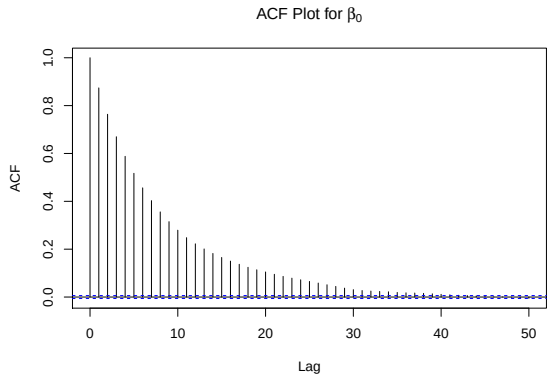



Figure 1.2: ACF plot for $\beta_0$ from a Monte Carlo sample size of $10^5$.

| $\text{ESS}_0$ | $\text{ESS}_1$ | $\text{ESS}_2$ | $\text{ESS}_3$ | $\text{ESS}_4$ |
|---|---|---|---|---|
| 6972 | 4623 | 6009 | 6391 | 4543 |

Table 1.1: Univariate effective sample size for estimating $\text{E}_F\beta$. Monte Carlo sample size is $10^5$.

for the $i$th component of the process as

$$\text{ESS}_i = \frac{n}{1 + 2\sum_{k=1}^{\infty} \rho_i(k)} = n\frac{\lambda_i^2}{\sigma_i^2}\,.$$

When the samples are i.i.d, $\text{ESS}_i = n$ since $\lambda_i^2 = \sigma_i^2$ and when there is positive correlation in the Markov chain, $\text{ESS}_i < n$. Using strongly consistent estimators of $\sigma_i^2$ and $\lambda_i^2$, Gong and Flegal (2016) estimate $\text{ESS}_i$ consistently and demonstrate that terminating when $\text{ESS}_i$ reaches a pre-specified lower bound is theoretically justified.

Due to the univariate construction, a separate $\text{ESS}_i$ is calculated for each $i = 1, \ldots, p$. Table 1.1 shows the estimated effective sample size for each of the five components of $\text{E}_F\beta$. Conservative termination dictates terminating when the smallest estimate among all $\text{ESS}_4$ (in this case $\text{ESS}_4$) is larger than a pre-specified lower bound. Thus termination is dictated by the slowest mixing component.

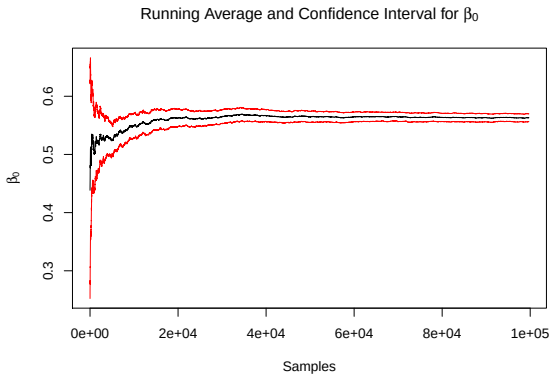



Figure 1.3: The running average and confidence interval for the estimate of $\mathrm{E}_F\beta_0$ over $10^5$ Monte Carlo samples.

### 1.1.2 Fixed-width Termination Rules

Jones et al. (2006) laid the foundation for termination based on quality of estimation rather than convergence of the Markov chain. From the univariate CLT in (1.5), the approximate distribution of the Monte Carlo error is

$$\theta_{n,i} - \theta_i \approx N\left(0, \frac{\sigma_i^2}{n}\right),$$

where $\sigma_i^2/n$ is the Monte Carlo standard error. For the Bayesian logistic regression example, we estimate $\sigma_i^2/n$ using the strongly consistent univariate batch means estimator described in Jones et al. (2006), and the estimate is used to create confidence intervals for $\beta_0$. Figure 1.3 shows the running average and confidence intervals around those averages for $\beta_0$. Notice, as the Monte Carlo sample size increases, the size of the confidence intervals decreases, indicating convergence of the estimate of $\sigma_i^2$. Jones et al. (2006) use the diminishing size of the confidence interval as motivation for their termination rule.

To determine the number of Monte Carlo samples needed until termination, Jones et al. (2006) implemented the *fixed-width sequential stopping rule* where simulation is terminated the first time the width of the confidence interval for each component

is small. Let $\sigma^2_{n,i}$ be a strongly consistent estimator of $\sigma^2_i$ and $t_*$ be an appropriate $t$-distribution quantile. Then for a desired tolerance of $\epsilon_i$ for component $i$, the rule terminates simulation the first time after some $n^*$ iterations, for all components

$$t_* \frac{\sigma_{n,i}}{\sqrt{n}} + n^{-1} \leq \epsilon_i \ .$$

Estimation in this way is reliable in the sense that if the procedure is repeated again, the estimates will not be vastly different (Flegal et al., 2008).

Since the termination procedure is univariate, a separate termination criterion is set for each component. Common practice is to terminate simulation when *all* components satisfy a termination rule. Due to multiple testing, a Bonferroni correction is often used. To create $100(1-\alpha)\%$ univariate confidence intervals, the fixed-width rule terminates simulation at the random time,

$$\inf \left\{ n > 0 : 2t_* \frac{\sigma_{n,i}}{\sqrt{n}} + \epsilon_i I(n < n^*) + \frac{1}{n} \leq \epsilon_i \quad \text{for all } i = 1, \ldots, p \right\} \ ,$$

where for uncorrected intervals $t_* = t_{1-\alpha/2,*}$ and for Bonferroni corrected intervals $t_* = t_{1-\alpha/2p,*}$.

A separate tolerance level $\epsilon_i$ is required for each component, which can be challenging for large $p$. Flegal and Gong (2015) present the *relative standard deviation fixed-width sequential stopping rule* that overcomes this problem by terminating simulation relative to the estimated standard deviation for $g_i$ under $F$. Figure 1.4 motivates the relative standard deviation fixed-width stopping rule. The plot on the left shows the density of a normal distribution with standard deviation 5 and the right plot has the density of a normal distribution with standard deviation 2. The gray region is the area within one standard deviation of the mean, and the black region is the area within an $\epsilon$th fraction of the standard deviation of the mean; this is the desired width of the confidence interval. Thus, the desired width of the confidence

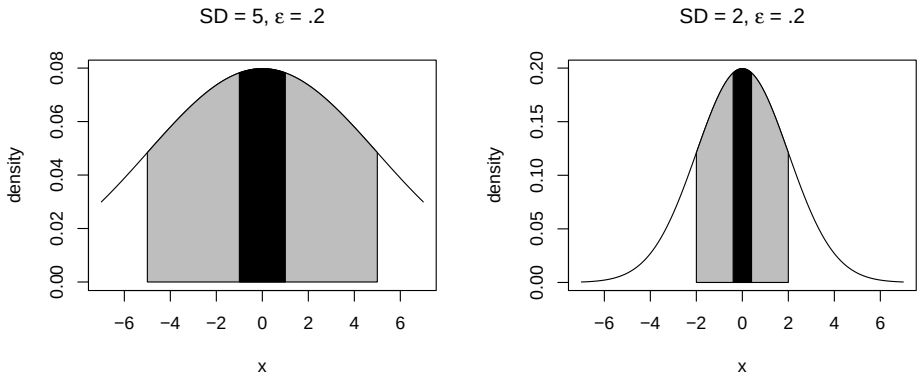



Figure 1.4: This figure motivates the relative standard deviation fixed-width stopping rule where $\epsilon$ is set to .2. The plot on the left shows the density of a normal distribution with standard deviation 5 and the left plots the density of a normal distribution with standard deviation 2. The gray region shades the area within one standard deviation of the mean, and the black region shades the area within an $\epsilon$th fraction of the standard devision of the mean.

interval adapts to the underlying variability in the distribution.

Let $\lambda_{n,i}$ be a strongly consistent estimator of $\lambda_i$, for example, $\lambda_{n,i}$ may be the sample standard deviation for $g_i(X)$. The relative standard deviation fixed-width sequential stopping rule terminates at the random time,

$$\inf\left\{n > 0 : \frac{1}{\lambda_{n,i}}\left(2t_*\frac{\sigma_{n,i}}{\sqrt{n}} + \epsilon_i\lambda_{n,i}I(n < n^*) + \frac{1}{n}\right) \leq \epsilon_i \quad \text{for all } i = 1, \ldots, p\right\},$$

where for uncorrected intervals $t_* = t_{1-\alpha/2,*}$ and for Bonferroni corrected intervals $t_* = t_{1-\alpha/2p,*}$. Flegal and Gong (2015) showed that this rule improved over the termination rule of Jones et al. (2006) and is easier to use for large $p$ problems. Implementing this rule is still challenging since

(a) when $p$ is even moderately large, the Bonferroni corrected intervals are large, leading to delayed termination; and

(b) simulation stops when each component satisfies the termination criterion; there-

fore, termination is governed by the slowest mixing component.

## 1.2 Multivariate Termination

The drawbacks of both the univariate termination methods presented in the previous section originate from ignoring the multivariate nature of the estimation problem. Recall the multivariate CLT presented in (1.4). Here

$$\Sigma = \mathrm{Var}_F\, g(X_1) + \sum_{k=1}^{\infty} \mathrm{Cov}_F\left(g(X_1), g(X_{1+k})\right) + \sum_{k=1}^{\infty} \mathrm{Cov}_F\left(g(X_{1+k}), g(X_1)\right) \ .$$

Although the structure of $\Sigma$ looks similar to (1.6), the details are more complicated. All three terms in the above expression are matrices where the first represents the covariance structure of the target distribution and the latter two represent the covariance structure due to the serial correlation in the Markov chain. Using the univariate methods presented in the previous section is akin to assuming that both $\mathrm{Var}_F\, g(X_1)$ and $\sum_{k=1}^{\infty} \mathrm{Cov}_F\left(g(X_1), g(X_{1+k})\right)$ are diagonal matrices. That is, the target distribution has uncorrelated components for $g$, and the components of the Markov chain are uncorrelated to each other. Outside of trivial examples, these assumptions are rarely satisfied when MCMC is relevant.

A helpful tool in understanding the entries of $\Sigma$ is the cross correlation function (CCF) plot. For entry $(i, j)$ of $\Sigma$, the CCF plot shows the correlation at lag $k$ between $g_i(X_1)$ and $g_j(X_{1+k})$. That is, the cross correlation $\rho_{i,j}(k)$ is

$$\rho_{i,j}(k) = \frac{\mathrm{Cov}_F\left(g_i(X_1), g_j(X_{1+k})\right)}{\sqrt{\mathrm{Var}_F\, g_i(X_1)\ \mathrm{Var}_F\, g_j(X_1)}}\,. \tag{1.8}$$

When $i = j$, this is the autocorrelation at lag $k$ and for $k = 0$, $\rho_{i,j}(0)$ is the correlation between the $i$th and the $j$th components in the target distribution. That is, $\rho_{i,j}(0)$

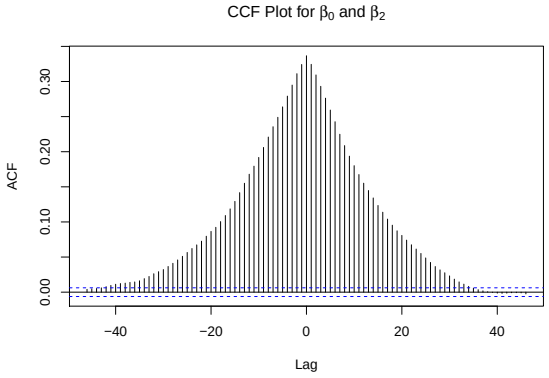



Figure 1.5: The cross correlation plot between $\beta_0$ and $\beta_2$ for the Bayesian logistic regression model from a Monte Carlo sample size of $10^5$.

is the $(i, j)$th entry of $\Lambda = \mathrm{Var}_F(g(X_1))$. Outside of trivial cases, $\rho_{i,j}(k)$ is non-zero and in fact can be large for smaller lags. For example, Figure 1.5 shows the CCF plot for $\beta_0$ and $\beta_2$ in the Bayesian logistic regression example. Notice how there is significant cross correlation above lag 30. The estimated correlation in the posterior for $\beta_0$ and $\beta_2$ is around 0.30. This inherent correlation in the posterior distribution is also ignored by univariate methods.

Thus even simple MCMC problems produce complex dependence structures within and across components of the samples. Ignoring this structure leads to an incomplete understanding of the estimation process. Not only do we gain more information about the Monte Carlo error using multivariate methods, we also avoid using conservative Bonferroni methods.

Assume for now that $\Sigma$ can be estimated consistently by $\Sigma_n$. In Chapter 4 we will discuss procedures for estimating $\Sigma$. Then as $n \to \infty$

$$(\theta_n - \theta)^T \Sigma_n^{-1} (\theta_n - \theta) \xrightarrow{d} \text{Hotelling's } T^2_{p,q},$$

where $q$ is determined by the choice of $\Sigma_n$. The above asymptotic distribution allows construction of large sample confidence ellipsoids around $\theta_n$. For example, Figure 1.6 shows the 90% joint confidence ellipse for $\beta_0$ and $\beta_2$ in the Bayesian logistic regression

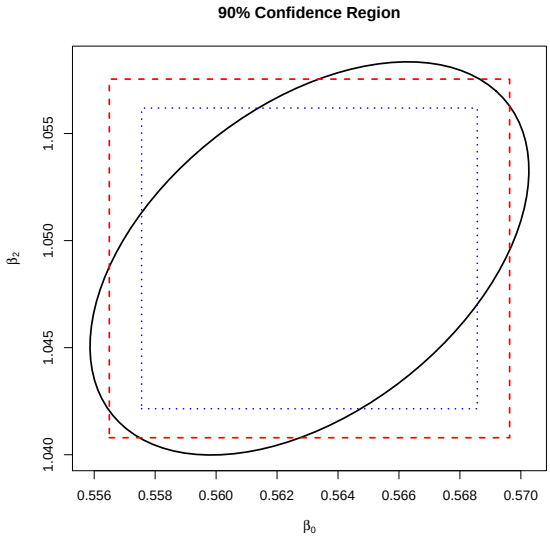



Figure 1.6: 90% confidence regions for $\beta_0$ and $\beta_2$. The solid ellipse is constructed using a multivariate estimator for $\Sigma$. The larger dashed box is constructed using Bonferroni corrected univariate methods and the smaller dotted box is constructed using uncorrected univariate methods.

example along with univariate confidence boxes constructed with and without a Bonferonni correction for the Bayesian logistic regression model. Note that the ellipse is oriented along non-standard axes indicating the presence of significant non-zero entries in $\Sigma$ and thus accounting for cross correlation in the estimation process.

With such confidence ellipsoids in mind, the drawbacks of the fixed-width sequential stopping procedure are overcome by the proposed *relative standard deviation fixed-volume sequential stopping rule*. This rule differs from the Jones et al. (2006) procedure in two fundamental ways:

(a) it is motivated by the multivariate CLT in (1.4) and not by the univariate CLT in (1.5), and

(b) it terminates simulation not by the absolute size of the confidence region, but by its size relative to the variability of $g$ under the target distribution.

Let $\Lambda_n$ be the sample covariance matrix, $|\cdot|$ denote determinant, and $\epsilon > 0$ be the tolerance level. The relative standard deviation fixed-volume sequential stopping rule terminates the first time, after some $n^*$ (user-specified) iterations that,

$$\text{Volume of Confidence Region}^{1/p} + n^{-1} < \epsilon |\Lambda_n|^{1/2p} \, . \tag{1.9}$$

The user-specified $n^*$ ensures that simulation does not terminate due to early bad estimates. The simulation terminates when the volume of the confidence region is small compared to the estimated variability of $g$ under the target distribution. When $p = 1$, this rule is equivalent to the relative standard deviation fixed-width sequential stopping rule of Flegal and Gong (2015).

Instead of the univariate effective sample size framework, we focus on a multivariate study of effective sample size since univariate treatment of ESS ignores cross-correlations across components, thus painting an inaccurate picture. A multivariate approach to ESS has not been studied in the literature. Define

$$\text{ESS} = n \left( \frac{|\Lambda|}{|\Sigma|} \right)^{1/p}.$$

When there is no correlation in the Markov chain, $\Sigma = \Lambda$ and ESS is equal to the number of Monte Carlo samples. In Chapter 3 we show that terminating according to the relative standard deviation fixed-volume sequential stopping rule is asymptotically equivalent to terminating when the estimated ESS satisfies

$$\widehat{\text{ESS}} \geq W_{p,\alpha,\epsilon},$$

where $W_{p,\alpha,\epsilon}$ can be calculated *a priori* and is a function only of the dimension of the estimation problem, the level of confidence of the confidence regions, and the relative precision desired. Thus, not only do we show that terminating via ESS is a rigorous procedure, we also provide theoretically valid, practical lower bounds on the number of effective samples required.

In Chapter 4 we present the multivariate batch means and the multivariate spectral variance estimators for $\Sigma$ and provide conditions for their strong consistency. Estimating $\Sigma$ consistently has not received much attention. Seila (1982) proposed a consistent estimator of $\Sigma$ using regenerative properties of Markov chains. Since identifying regenerations is often challenging, the estimator is difficult to use. More recently, Dai and Jones (2016) presented a conservative estimator of $\Sigma$ by introducing a multivariate initial sequence estimator.

In Chapter 5 we apply our multivariate methods to a wide array of examples. Each example serves a different purpose in illustrating the advantages of multivariate output analysis over univariate analysis. The first example is a vector autoregressive (VAR) process of order 1. This example is unique in that the true covariance matrix $\Sigma$ is known and by changing certain parameters of the model, the underlying stochastic process can be made to mix arbitrarily slow or fast. We use these properties of the VAR example to examine and compare the performance of our estimators of $\Sigma$. Next we continue with our example of the Bayesian logistic regression model and implement our stopping rules.

Our third example is the Bayesian lasso model where the underlying Markov chain is described by a Gibbs sampler. Although the Markov chain mixes fairly well in this example, $p$ is quite large. All examples thus far are known to satisfy the conditions required for our theoretical results. The next example is a Bayesian dynamic spatio-temporal model where it is unknown if the required conditions are satisfied. This is also an instance where the target distribution is heavily correlated and thus the Markov chain is fairly slow mixing.

Throughout the examples presented in this dissertation, the multivariate stopping rules terminate earlier than univariate methods because

(a) termination is dictated by the joint behavior of the components of the Markov chain and not by the component that mixes the slowest,

(b) using the inherent multivariate nature of the problem and acknowledging cross correlations leads to a more realistic understanding of the estimation process, and

(c) avoiding corrections for multiple testing gives considerably smaller confidence regions even in moderate $p$ problems.

# Chapter 2

# Markov Chains and the Strong Invariance Principle

MCMC is used to generate samples from a target distribution $F$ defined on $\mathcal{X}$. Typically, for a function $g : \mathcal{X} \to \mathbb{R}^p$, interest is in estimating $\theta = \mathrm{E}_F g$. When $\mathcal{X}$ is high dimensional, i.i.d sampling is often either impossible or inefficient. Instead, MCMC methods simulate a Markov chain such that the desired target distribution is its stationary distribution. This is typically done using Metropolis-Hastings algorithms, Gibbs sampling or a combination of both. These methods simulate a Markov chain $X = \{X_t\}$ such that $X_\infty$ is an exact draw from $F$. Thus, the sample is neither independent nor identically distributed. With the goal of estimating $\theta$, statistical analysis of the output data, $\{X_t\}$, is referred to as output analysis. Naturally, output analysis for MCMC relies heavily on Markov chain theory. Specifically, the rate at which the Markov chain converges to the stationary distribution often dictates the quality of the estimates of $\theta$ for a given Monte Carlo sample size.

In this chapter we present relevant definitions and Markov chain properties. We focus on the conditions required for the Markov chain CLT to hold. In addition, we discuss the conditions on the Markov chain and $g$ that guarantee the existence of a strong invariance principle.

## 2.1 Markov Chain Theory

Let $X = \{X_t\}$ be a time-homogeneous discrete-time Markov chain defined on the state space $\mathcal{X}$. Recall that $\mathcal{X}$ is equipped with a countably generated $\sigma$-field, $\mathcal{B}(\mathcal{X})$. The Markov chain $X$ is defined by its transition kernel, $P : (\mathcal{X}, \mathcal{B}(\mathcal{X})) \to [0, 1]$. For $x \in \mathcal{X}$ and $A \in \mathcal{B}(\mathcal{X})$, $P(x, A)$ is defined as

$$P(x, A) = \Pr(X_{t+1} \in A \mid X_t = x) .$$

$P(x, \cdot)$ defines a probability measure on $\mathcal{B}(\mathcal{X})$ whereas $P(\cdot, A)$ defines a measurable function on $\mathcal{X}$. The $n$-step transition kernel $P^n(x, \cdot)$ is defined as

$$P^n(x, A) = \Pr(X_{t+n} \in A \mid X_t = x) \quad \text{for } x \in \mathcal{X} \text{ and } A \in \mathcal{B}(\mathcal{X}) .$$

If for all $x, y \in \mathcal{X}$

$$F(dy) = \int_{\mathcal{X}} P(x, dy) F(dx) , \tag{2.1}$$

then $F$ is called the *invariant* or *stationary* distribution and the Markov chain $X$ is called $F$-invariant. Equation (2.1) ensures that for an initial value drawn from $F$, the Markov chain leaves the distribution of the next value unchanged. Thus if $X_1 \sim F$, then the Markov chain produces exact correlated draws from $F$.

In situations where MCMC is relevant, it is generally not possible to produce $X_1 \sim F$. A natural question is, under what conditions can the Markov chain converge to $F$ and how is convergence assessed? Before we present the conditions, we introduce some preliminary definitions.

**Definition 2.1** A Markov chain is *$F$-irreducible* is for any $x \in \mathcal{X}$ and any set $A \in$

$\mathcal{B}(\mathcal{X})$ with $F(A) > 0$, there exists a finite $n$ such that $P^n(x, A) > 0$. □

Irreducibility ensures that with positive probability the Markov chain can eventually reach any set with positive $F$-probability from any state in the state space.

**Definition 2.2** A Markov chain is *periodic* if there exists $d \geq 2$ and disjoint sets $A_1, A_2, \ldots, A_d \subseteq \mathcal{X}$ with $P(x, A_{i+1}) = 1$ for all $x \in A_i$ for $i = 1, \ldots, d-1$ and $P(x, A_1) = 1$ for all $x \in A_d$. Otherwise, the Markov chain is *aperiodic*. □

If the Markov chain is periodic, then we can partition the state space into sets such that once in one of the sets, the Markov chain systematically cycles through the sets. Thus aperiodicity ensures that the Markov chain does not get stuck in such a cycle.

**Definition 2.3** An $F$-invariant Markov chain is *Harris recurrent* if for all $A \in \mathcal{B}(\mathcal{X})$ with $F(A) > 0$, and for all $x \in \mathcal{X}$, $\Pr(\exists n \in \mathbb{N} : X_n \in A \mid X_1 = x) = 1$. If $F$ is a probability distribution then the Markov chain is *positive recurrent*. □

Harris recurrence is a stronger property than irreducibility. In addition, it also implies that starting at any point in the state space the Markov chain will visit any $F$-positive set infinitely often. When a Markov chain is $F$-irreducible, aperiodic and positive recurrent, we call it a *Harris ergodic* Markov chain. We now define a notion of distance between two measures.

**Definition 2.4** For two probability measures $\nu_1$ and $\nu_2$ on $\mathcal{B}(\mathcal{X})$, the *total variation distance* between $\nu_1$ and $\nu_2$ is defined as,

$$\|\nu_1(\cdot) - \nu_2(\cdot)\|_{TV} = \sup_{A \in \mathcal{B}(\mathcal{X})} |\nu_1(A) - \nu_2(A)| . \qquad \square$$

**Theorem 2.1** Let the Markov chain with transition kernel $P$ be $F$-invariant and Harris ergodic. Then as $n \to \infty$

$$\|P^n(x,\cdot) - F(\cdot)\|_{TV} \to 0 \text{ for all } x \in \mathcal{X}\,. \qquad \square$$

The above theorem is proved in Athreya et al. (1996) (Theorem 1). Athreya et al. (1996) also show that under the same conditions, the strong law of large numbers for the Monte Carlo estimator also holds. However, for a Markov chain CLT to hold, the Markov chain has to converge "fast enough".

**Definition 2.5** Let $X$ be an $F$-invariant Harris ergodic Markov chain. If there exists $M : \mathcal{X} \to \mathbb{R}^+$ and $\psi : \mathbb{N} \to \mathbb{R}^+$ such that for all $n \in \mathbb{N}$ and for all $x \in \mathcal{X}$,

$$\|P^n(x,\cdot) - F(\cdot)\|_{TV} \leq M(x)\psi(n)\,, \tag{2.2}$$

then,

(a) if $\psi(n) = n^{-m}$ for some $m > 0$, $X$ is *polynomially ergodic* of order $m$.

(b) if $\psi(n) = t^n$ for some $0 \leq t < 1$, $X$ is *geometrically ergodic*.

(c) if $\sup_{x\in\mathcal{X}} M(x) < \infty$ and $\psi(n) = t^n$ for some $0 \leq t < 1$, $X$ is *uniformly ergodic*. $\square$

Uniform ergodicity of the Markov chain implies geometric ergodicity, which implies polynomial ergodicity of the Markov chain. With the assumption of higher finite moments for $g$, a Markov chain CLT holds if the chain is uniformly ergodic, geometrically ergodic, or polynomially ergodic.

Recall that $g = (g_1, g_2, \ldots, g_p)$; and $\theta_i$ and $\theta_{n,i}$ denotes the $i$th component of $\theta$ and $\theta_n$ respectively.

**Theorem 2.2** Let $X$ be an $F$-invariant Harris ergodic Markov chain. Suppose at least one of the following holds for $g_i$:

(a) (Jones, 2004) $X$ is polynomially ergodic of order $k$, $\mathrm{E}_F|M(x)| < \infty$ and $\mathrm{E}_F|g_i(x)|^{2+\delta} < \infty$ for some $\delta$ such that $k\delta > 2 + \delta$,

(b) (Jones, 2004) $X$ is polynomially ergodic of order $k > 1$, $\mathrm{E}_F|M(x)| < \infty$ and $\sup_{x \in X} |g_i(x)| < \infty$,

(c) (Chan and Geyer, 1994) $X$ is geometrically ergodic and $\mathrm{E}_F|g_i(x)|^{2+\delta} < \infty$ for some $\delta > 0$,

(d) (Ibragimov and Linnik, 1971) $X$ is uniformly ergodic and $\mathrm{E}_F[g_i(x)^2] < \infty$,

then, for any initial distribution, as $n \to \infty$,

$$\sqrt{n}(\theta_{n,i} - \theta_i) \xrightarrow{d} N(0, \sigma_i^2)\,.$$

□

By the Cramér-Wold theorem, a multivariate CLT holds under the same conditions as Theorem 2.2. That is, there exists a $p \times p$ positive definite matrix $\Sigma$ such that as $n \to \infty$,

$$\sqrt{n}(\theta_n - \theta) \xrightarrow{d} N_p(0, \Sigma)\,.$$

Here $\Sigma = \lim_{n\to\infty} n \operatorname{Var}_F(\theta_n)$. Using mixing properties of Harris ergodic Markov chains,

$$\begin{aligned}
&n \operatorname{Var}_F(\theta_n) \\
&= n \operatorname{Var}_F\left(\frac{1}{n}\sum_{t=1}^{n} g(X_t)\right) \\
&= \frac{1}{n}\sum_{t=1}^{n}\sum_{l=1}^{n} \operatorname{Cov}_F\left(g(X_t), g(X_l)\right) \\
&= \frac{1}{n}\left(\sum_{t=1}^{n} \operatorname{Var}_F(g(X_t)) + \sum_{t<l} \operatorname{Cov}_F(g(X_t), g(X_l)) + \sum_{t<l} \operatorname{Cov}_F(g(X_l), g(X_t))\right)
\end{aligned}$$

$$= \mathrm{Var}_F(g(X_1)) + \sum_{k=1}^{n-1} \left(\frac{n-k}{n}\right) \mathrm{Cov}_F(g(X_1), g(X_{1+k})) + \sum_{k=1}^{n-1} \left(\frac{n-k}{n}\right) \mathrm{Cov}_F(g(X_{1+k}), g(X_1))$$

Thus,

$$\Sigma = \mathrm{Var}_F\, g(X_1) + \sum_{k=1}^{\infty} \mathrm{Cov}_F(g(X_1), g(X_{1+k})) + \sum_{k=1}^{\infty} \mathrm{Cov}_F(g(X_{1+k}), g(X_1))\,. \tag{2.3}$$

One of our contributions is demonstrating strong consistency for two classes of estimators of $\Sigma$. Our main assumption on the Markov chain is the existence of a *strong invariance principle*. In fact, our theoretical results for estimators of $\Sigma$ hold outside the context of Markov chains for processes that satisfy a strong invariance principle. In the next section we discuss a wide variety of processes for which a strong invariance principle holds and the exact conditions for Markov chains that yield the strong invariance principle.

## 2.2 Strong Invariance Principle

Although our focus is on Markov chains, our theoretical results in this chapter hold for stochastic processes satisfying a strong invariance principle, which we now describe. Let $\|\cdot\|$ denote the Euclidean norm. Let $\{B(t), t \geq 0\}$ be a $p$-dimensional standard Brownian motion. A strong invariance principle (SIP) holds for $\theta_n$ if there exists a $p \times p$ lower triangular matrix $L$, a nonnegative increasing function $\psi$ on the positive integers, a finite random variable $D$, and a sufficiently rich probability space such

that, with probability 1,

$$\|n(\theta_n - \theta) - LB(n)\| < D\psi(n) \text{ as } n \to \infty. \tag{2.4}$$

Intuitively, (2.4) means that the centered and appropriately scaled partial sum process is similar to a scaled Brownian motion. Dividing (2.4) by $n$ throughout we get, with probability 1

$$\left\|(\theta_n - \theta) - L\frac{B(n)}{n}\right\| = O\left(\frac{\psi(n)}{n}\right).$$

By the strong law of large numbers for the classical setting, as $n \to \infty$, $B(n)/n \to 0$ with probability 1 and thus if $\psi(n)/n \to 0$, the strong law for $\theta_n$ holds.

Similarly, dividing (2.4) by $\sqrt{n}$,

$$\left\|\sqrt{n}(\theta_n - \theta) - L\frac{B(n)}{\sqrt{n}}\right\| = O\left(\frac{\psi(n)}{\sqrt{n}}\right).$$

If $\psi(n)/\sqrt{n} \to 0$, then since $B(n)/\sqrt{n}$ is a $p$-dimensional standard normal distribution we arrive at the central limit theorem and $\Sigma = LL^T$. The strong invariance principle also implies a functional CLT for $\theta_n$.

The existence of an SIP has attracted much research interest. Consider the univariate case when $p = 1$. For i.i.d processes, the first result of this kind is due to Strassen (1964) who showed that (2.4) holds with $\psi(n) = \sqrt{n \log\log n}$. Komlós et al. (1975) found that if $\mathrm{E}_F|g|^{2+\delta} < \infty$, then (2.4) holds with $\psi(n) = n^{1/2-\lambda}$ for $\lambda > 0$ (often called the KMT bound). Komlós et al. (1975) also showed that if $g$ has all moments in a neighborhood of 0, then $\psi(n) = \log n$. The results of Komlós et al. (1975) are the strongest to date in the i.i.d setting. The main reference for a univariate strong invariance principle for dependent sequences is Philipp and Stout (1975) who prove bounds similar to that of Komlós et al. (1975) for a variety of weakly

dependent processes including $\phi$-mixing, regenerative, and strongly mixing processes. Also, see Wu (2007) for a univariate strong invariance principle for certain classes of dependent processes.

Many of the univariate SIPs have been extended to the multivariate setting. For independent processes, Berkes and Philipp (1979), Einmahl (1989), and Zaitsev (1998) extend the results of Komlós et al. (1975). For correlated processes, Eberlein (1986) showed the existence of a strong invariance principle for Martingale sequences and Horvath (1984) proved the KMT bound for multivariate extended renewal processes. For $\phi$-mixing, strongly mixing, and absolutely regular processes, Kuelbs and Philipp (1980) and Dehling and Philipp (1982) extended the Philipp and Stout (1975) results to the multivariate case.

As mentioned before, our strong consistency results for estimators of $\Sigma$ hold under the assumption of a strong invariance principle with $\psi(n)$ satisfying certain conditions. These conditions depend on the choice of estimator used for $\Sigma$. Next, we give an overview of the conditions under which a univariate strong invariance principle holds for Markov chains and establish conditions under which a multivariate strong invariance principle holds.

### 2.2.1 Univariate SIP for Markov Chains

The existence of a univariate strong invariance principle for uniformly ergodic and geometrically ergodic Markov chains was discussed by Jones et al. (2006) and Bednorz and Łatuszyński (2007). Before we present their results, we introduce the concept of minorization. A one-step minorization condition holds if there exists a function $s : \mathcal{X} \to [0, 1]$ with $\mathrm{E}_F s > 0$ and a probability measure $Q$ such that for all $x \in \mathcal{X}$ and $A \in \mathcal{B}(\mathcal{X})$

$$P(x, A) \geq s(x)Q(A)\,. \tag{2.5}$$

As explained in Jones et al. (2006), for finite state spaces (2.5) holds by fixing $x_* \in \mathcal{X}$ and then setting $s(x) = I(x = x_*)$. Then for $x \neq x_*$, $P(x, A) \geq 0$ remains true and for $x = x_*$, $P(x, A) = s(x)P(x_*, A)$. For general state spaces, establishing a minorization is more difficult. However, Mykland et al. (1995) describes a recipe for establishing (2.5) that is often useful.

**Theorem 2.3** Let $X$ be a Harris ergodic $F$-invariant Markov chain and $g : \mathcal{X} \to \mathbb{R}$.

1. (Jones et al., 2006) If $X$ is uniformly ergodic and $\mathrm{E}_F|g|^{2+\delta} < \infty$ for some $\delta > 0$, then (2.4) holds with $\psi(n) = n^{1/2-\alpha}$ where $\alpha > \delta/(24 + 12\delta)$.

2. (Jones et al., 2006; Bednorz and Łatuszyński, 2007) If $X$ is geometrically ergodic, (2.5) holds, and $\mathrm{E}_F|g|^{2+\delta+\epsilon} > 0$ for some $\delta > 0$ and $\epsilon > 0$, then (2.4) holds with $\psi(n) = n^{\alpha} \log n$ where $\alpha > 1/(2 + \delta)$. □

Notice how the above results require more than a finite second moment, even though a finite second moment might guarantee the existence of a CLT. A more recent result for bounded functions improves on the previous result.

**Theorem 2.4 (Merlevède and Rio (2015))** Let $X$ be a Harris ergodic $F$-invariant Markov chain and let $g : \mathcal{X} \to \mathbb{R}$ be such that $|g(x)| < R$ for some $R > 0$. If $X$ is geometrically ergodic and (2.5) holds, then (2.4) holds with $\psi(n) = \log n$. □

The rate $\psi(n) = \log n$ is the best possible. When $|g(x)| < R$, then $g$ has all moments in a neighborhood of zero. Recall that the rate $\psi(n) = \log n$ is also obtained for i.i.d. processes by Komlós et al. (1975). Thus for bounded functions, almost nothing is lost when an i.i.d process is replaced with a geometrically ergodic chain with a minorization. Also notice from Theorem 2.2 that for bounded functions the CLT holds for polynomially ergodic Markov chains. It is natural to wonder whether

the result of Merlevède and Rio (2015) can be obtained for polynomially ergodic Markov chains.

### 2.2.2 Multivariate SIP for Markov Chains

Let $S = \{S_t\}_{t\geq 1}$ be a strictly stationary stochastic process on a probability space $(\Omega, \mathcal{F}, \mathsf{P})$ and set $\mathcal{F}_k^l = \sigma(S_k, \ldots, S_l)$. Define the $\alpha$-mixing coefficients for $n = 1, 2, 3, \ldots$ as

$$\alpha(n) = \sup_{k\geq 1} \sup_{A\in\mathcal{F}_1^k,\, B\in\mathcal{F}_{k+n}^\infty} |\mathsf{P}(A\cap B) - \mathsf{P}(A)\mathsf{P}(B)| \,.$$

The process $S$ is said to be *strongly mixing* if $\alpha(n) \to 0$ as $n \to \infty$. It is easy to see that Harris ergodic Markov chains are strongly mixing; see, for example, Jones (2004). We will use the following result from Kuelbs and Philipp (1980) to establish the conditions under which a strong invariance principle holds for Harris ergodic Markov chains.

**Theorem 2.5 (Kuelbs and Philipp (1980))** Let $g(S_1), g(S_2), \ldots$ be an $\mathbb{R}^p$-valued stationary process such that $\mathrm{E}_F\|g\|^{2+\delta} < \infty$ for some $0 < \delta \leq 1$. Let $\alpha_g(n)$ be the mixing coefficients of the process $\{g(S_t)\}$ and suppose, as $n \to \infty$,

$$\alpha_g(n) = O\left(n^{-(1+\epsilon)(1+2/\delta)}\right) \quad \text{for } \epsilon > 0.$$

Then a strong invariance principle holds as in (2.4) with $\psi(n) = n^{1/2-\lambda}$ for some $\lambda > 0$ depending on $\epsilon$, $\delta$, and $p$ only. □

**Corollary 2.1** Let $\mathrm{E}_F\|g\|^{2+\delta} < \infty$ for some $\delta > 0$. If $X$ is a polynomially ergodic Markov chain of order $m \geq (1+\epsilon_1)(1+2/\delta)$ for some $\epsilon_1 > 0$, then (2.4) holds for any initial distribution with $\psi(n) = n^{1/2-\lambda}$ for some $\lambda > 0$. □

**Proof**

Let $\alpha$ be the mixing coefficient for the Markov chain $X = \{X_t\}$ and $\alpha_g$ be the mixing coefficient for the mapped process $\{g(X_t)\}$. Then by elementary properties of sigma-algebras (cf. Chow and Teicher, 1978, p. 16), $\alpha_g(n) \leq \alpha(n)$ for all $n$.

Note that since $X$ is polynomially ergodic of order $m$,

$$\begin{aligned}
&\|P^n(x,\cdot) - F(\cdot)\|_{TV} \leq M(x)n^{-m} \\
\Rightarrow &\sup_{A\in\mathcal{B}(\mathcal{X})} |P^n(x,A) - F(A)| \leq M(x)n^{-m}
\end{aligned}$$

Thus, for all $A \in \mathcal{B}(\mathcal{X})$ and arbitrary $k \in \mathbb{N}, B \in \mathcal{B}(\mathcal{X})$

$$\begin{aligned}
&|P^n(x,A) - F(A)| \leq M(x)n^{-m} \\
\Rightarrow &\int_B |P^n(x,A) - F(A)|F(dx) \leq \int_B M(x)n^{-m}F(dx) \\
\Rightarrow &\left|\int_B P^n(x,A)F(dx) - F(A)F(dx)\right| \leq \int_B M(x)n^{-m}F(dx) \\
\Rightarrow &|\Pr(X_{n+k} \in A \text{ and } X_k \in B) - F(A)F(B)| \leq \mathrm{E}_F M n^{-m} \\
\Rightarrow &\sup_{k\geq 1} \sup_{B\in\mathcal{F}_1^k, A\in\mathcal{F}_{k+n}^\infty} |\Pr(X_{n+k} \in A \text{ and } X_k \in B) - F(A)F(B)| \leq \mathrm{E}_F M n^{-m} \\
\Rightarrow &\alpha(n) \leq \mathrm{E}_F M n^{-m}\,.
\end{aligned}$$

Thus, $\alpha(n) \leq \mathrm{E}_F M n^{-m}$ for all $n$ and hence if $m \geq (1+\epsilon_1)(1+2/\delta)$, then $\alpha_g(n) \leq \mathrm{E}_F M n^{-m} = O(n^{-(1+\epsilon_1)(1+2/\delta)})$. The result follows from Theorem 2.5 and thus the strong invariance principle, as stated, holds at stationarity. A standard Markov chain argument (see, e.g., Proposition 17.1.6 in Meyn and Tweedie (2009)) shows that if the result holds for any initial distribution, then it holds for every initial distribution. We present the proof below for completeness.

Let

$$g_\infty(x) = \Pr\left(\sum_{t=1}^{n} g(X_t) - n\theta - LB(n) = O\left(n^{1/2-\lambda}\right) \Big| X_1 = x\right).$$

Then if $X_1 \sim F, \int g_\infty dF = 1$. We will show that $g_\infty$ is a harmonic function, which together with Theorem 17.1.5 of Meyn and Tweedie (2009) would imply that $g_\infty$ is a constant function.

$$\begin{aligned}
&Pg_\infty(x) \\
&= \int_{\mathcal{X}} P(x, dy) g_\infty(y) \\
&= \mathrm{E}\left[\Pr\left(\sum_{t=1}^{n} g(X_t) - n\theta - LB(n) = O\left(n^{1/2-\lambda}\right) \;\Big|\; X_2 = y\right) \;\Big|\; X_1 = x\right]
\end{aligned}$$

By the Markov property

$$\begin{aligned}
&= \mathrm{E}\left[\Pr\left(\sum_{t=1}^{n} g(X_t) - n\theta - LB(n) = O\left(n^{1/2-\lambda}\right) \Big| X_2 = y, X_1 = x\right) \Big| X_1 = x\right] \\
&= \Pr\left(\sum_{t=1}^{n} g(X_t) - n\theta - LB(n) = O\left(n^{1/2-\lambda}\right) \Big| X_1 = x\right) \\
&= g_\infty(x).
\end{aligned}$$

Thus $g_\infty$ is a harmonic function, and $g_\infty(x) = 1$ for all $x$. Hence stationarity is not a necessary condition for the strong invariance principle to hold. □

**Remark 2.1** This is the first direct presentation of the existence of a strong invariance principle (univariate or multivariate) for polynomially ergodic Markov chains. Thus we weaken the conditions even for the univariate case by only requiring polynomial ergodicity and not requiring the minorization condition as in (2.5). □

**Remark 2.2** Kuelbs and Philipp (1980) show that $\lambda$ only depends on $p$, $\epsilon$ and $\delta$, but the exact relationship remains an open problem. For slowly mixing Markov chains $\lambda$ is closer to 0 while for fast mixing chains $\lambda$ is closer to 1/2 (Damerdji, 1991). □

**Remark 2.3** A set $C \in \mathcal{B}(\mathcal{X})$ is called a *small set* if there exists $\xi > 0$ and a probability measure $\mu$ such that for all $x \in C$ and some $k \in \mathbb{N}$,

$$P^k(x, \cdot) \geq \xi\mu(\cdot)\,.$$

The constant $\xi$ measures the rate at which the effect of the initial value int he Markov chain is lost. Polynomial ergodicity is often proved by establishing the following drift condition. For a function $V : \mathcal{X} \to [1, \infty)$ there exists $d > 0, b < \infty$, and $0 \leq \tau < 1$ such that for $x \in \mathcal{X}$

$$\mathrm{E}[V(X_{n+1})|X_n = x] - V(x) \leq -d[V(x)]^\tau + bI(x \in C)\,, \tag{2.6}$$

where $C$ is a small set and the expectation is with respect to $P(x, \cdot)$. In order to verify that $\mathrm{E}_F M < \infty$, it is sufficient to show that $\mathrm{E}_F V < \infty$ by Theorem 14.3.7 in Meyn and Tweedie (2009). □

# Chapter 3

# Multivariate Analysis

This chapter introduces our multivariate methods for terminating simulation in more detail. Throughout this chapter we assume that $\Sigma_n$ is a strongly consistent estimator of $\Sigma$. In Chapter 4 we will present two such estimators.

## 3.1 Termination Rules

Let $T^2_{1-\alpha,p,q}$ denote the $(1-\alpha)$ quantile of a Hotelling's T-squared distribution with dimensionality parameter $p$ and degrees of freedom $q$ (the $\alpha$ here is the usual confidence level and different from the $\alpha$ in the previous chapter). Recall that due to the Markov chain CLT in (1.4), as $n \to \infty$

$$n(\theta_n - \theta)^T \Sigma_n^{-1} (\theta_n - \theta) \xrightarrow{d} T^2_{p,q}\,,$$

where $q$ is determined by the choice of estimator for $\Sigma_n$. A $100(1-\alpha)\%$ confidence region for $\theta$ is the set

$$C_\alpha(n) = \left\{\theta \in \mathbb{R}^p : n(\theta_n - \theta)^T \Sigma_n^{-1} (\theta_n - \theta) < T^2_{1-\alpha,p,q}\right\}\,.$$

Then $C_\alpha(n)$ forms an ellipsoid in $p$ dimensions oriented along the directions of the eigenvectors of $\Sigma_n$. The eigenvalues of $\Sigma_n$ dictate the length of the directional axes

for the ellipsoid. For large samples, $C_\alpha(n)$ is the smallest volume confidence region around $\theta_n$. Recall that $|\cdot|$ denotes determinant and let $\Gamma(\cdot)$ denote the Gamma function. The volume of the confidence region is

$$\text{Vol}(C_\alpha(n)) = \frac{2\pi^{p/2}}{p\Gamma(p/2)} \left( \frac{T^2_{1-\alpha,p,q}}{n} \right)^{p/2} |\Sigma_n|^{1/2} . \tag{3.1}$$

Note that $q$ is often increasing in the Monte Carlo size $n$ and thus $T^2_{1-\alpha,p,q} \to \chi^2_{1-\alpha,p}$ as $n \to \infty$, where $\chi^2_{1-\alpha,p}$ is the $(1-\alpha)$th quantile of the $\chi^2$ distribution with $p$ degrees of freedom. Also since $\Sigma_n$ is strongly consistent for $\Sigma$ and $|\cdot|$ is a continuous function, by the continuous mapping theorem,

$$\begin{aligned}
\lim_{n\to\infty} \text{Vol}(C_\alpha(n)) &= \lim_{n\to\infty} \frac{2\pi^{p/2}}{p\Gamma(p/2)} \left( \frac{T^2_{1-\alpha,p,q}}{n} \right)^{p/2} |\Sigma_n|^{1/2} \\
&= \frac{2\pi^{p/2}}{p\Gamma(p/2)} |\Sigma|^{1/2} \lim_{n\to\infty} \left( \frac{T^2_{1-\alpha,p,q}}{n} \right)^{p/2} \\
&= \frac{2\pi^{p/2}}{p\Gamma(p/2)} |\Sigma|^{1/2} (\chi^2_{1-\alpha,p})^{p/2} \lim_{n\to\infty} \left( \frac{1}{n} \right)^{p/2} \\
&= 0 \text{ with probability } 1 .
\end{aligned}$$

Hence, as more samples are obtained, the volume of the confidence interval decreases. Note that the decrease may not be monotonic due to the randomness of $\Sigma_n$.

Let $s(n)$ be a positive and decreasing function on $\{1, 2, \ldots, \}$ and $\epsilon > 0$ be the tolerance level. Glynn and Whitt (1992) present the *fixed-volume sequential stopping rule* which terminates the simulation at the random time

$$T(\epsilon) = \inf \left\{ n \geq 0 : \text{Vol}(C_\alpha(n))^{1/p} + s(n) \leq \epsilon \right\} . \tag{3.2}$$

Glynn and Whitt (1992) provide conditions so that terminating at $T(\epsilon)$ yields confi-

dence regions that are asymptotically valid in the sense that,

$$\Pr\left[\theta \in C_\alpha(T(\epsilon))\right] \to 1-\alpha \text{ as } \epsilon \to 0\,.$$

The conditions required by Glynn and Whitt (1992) for asymptotic validity are (i) a functional CLT holds for the stochastic process, (ii) $\Sigma_n$ is strongly consistent for $\Sigma$, and (iii) $s(n) = o(n^{-1/2})$. In particular for $n^* > 0$, they let $s(n) = \epsilon I(n < n^*) + n^{-1}$ which ensures simulation does not terminate before $n^*$ iterations due to initial bad estimates of $\Sigma$.

The sequential stopping rule (3.2) can be difficult to implement in practice since the choice of $\epsilon$ depends on the units of $\theta$, and has to be chosen for every application. In addition, if the components of $\theta$ are in different units, then $\epsilon$ lacks interpretability. We consider an alternative to (3.2) which can be used more naturally and which we will show connects nicely to the idea of effective sample size.

Recall that $\|\cdot\|$ denotes the Euclidean norm. Let $K(g(X), p) > 0$ be an attribute of the estimation process and suppose $K_n(g(X), p) > 0$ is an estimator of $K(g(X), p)$; for example, take $K(g(X), p) = \|\theta\|$ and $K_n(g(X), p) = \|\theta_n\|$. Set $s(n) = \epsilon K_n(g(X), p) I(n < n^*) + n^{-1}$. Then the relative fixed-volume sequential stopping rule terminates simulation at the random time,

$$T^*(\epsilon) = \inf\left\{n \geq 0 : \mathrm{Vol}(C_\alpha(n))^{1/p} + s(n) \leq \epsilon K_n(g(X), p)\right\}\,.$$

We call $K(g(X), p)$ the *relative metric*. Simulation terminates the first time after $n^*$ iterations, the volume of the confidence region is an $\epsilon$th fraction of the relative metric.

**Remark 3.1** The above rule is a more general version of the *relative precision fixed-volume sequential stopping rule* of Glynn and Whitt (1992). Specifically, they use $K(g(X), p) = \|\theta\|$ and $K_n(g(X), p) = \|\theta_n\|$. Thus simulation terminates relative to

the size of the Monte Carlo estimates. However, if $\|\theta\|$ is zero, this method cannot be used. □

For the relative fixed-volume sequential stopping rule to be effective, $T^*(\epsilon)$ should increase as $\epsilon \to 0$. That is, more desired precision would require delayed random time termination. The following result establishes this property and asymptotic validity of the termination rule. The proof is provided in Appendix A.

**Theorem 3.1** Let $g : \mathcal{X} \to \mathbb{R}^p$ be such that $\mathrm{E}_F\|g\|^{2+\delta} < \infty$ for some $\delta > 0$ and let $X$ be an $F$-invariant polynomially ergodic Markov chain of order $m > (1+\epsilon_1)(1+2/\delta)$ for some $\epsilon_1 > 0$. If $K_n(g(X), p) \to K(g(X), p)$ with probability 1 and $\Sigma_n \to \Sigma$ with probability 1 as $n \to \infty$, then as $\epsilon \to 0$, $T^*(\epsilon) \to \infty$ and $\Pr\left[\theta \in C_\alpha(T^*(\epsilon))\right] \to 1-\alpha$.□

**Remark 3.2** Glynn and Whitt (1992) required the existence of a functional CLT for asymptotic validity results. As discussed in Chapter 2, for Markov chains a functional CLT is implied by the strong invariance principle. In Chapter 4 we will require a strong invariance principle to hold for $\Sigma_n$ to be strongly consistent for $\Sigma$. Thus in Theorem 3.1, by using Corollary 2.1, we encapsulate these conditions under the framework for polynomial ergodicity and moment conditions. □

**Remark 3.3** Theorem 3.1 holds when $K(g(X), p) = K_n(g(X), p) = 1$. This choice of the relative metric leads to the fixed-volume sequential stopping rule of Glynn and Whitt (1992). It is also a multivariate generalization of the procedure developed by Jones et al. (2006). □

To implement the relative stopping rule, careful choice of the relative metric $K(g(X), p)$ is required. As discussed earlier, one can choose $K(g(X), p) = \|\theta\|$, if small confidence regions relative to the magnitude of the estimate are desired. How-

ever, we present a default relative metric that is intuitive and connects nicely to the idea of effective sample size. Let

$$K(g(X), p) = |\Lambda|^{1/2p} = |\operatorname{Var}_F g(X_1)|^{1/2p} .$$

We estimate $\Lambda$ by the sample covariance matrix. That is,

$$\Lambda_n := \frac{1}{n} \sum_{t=1}^{n} \left(g(X_t) - \theta\right) \left(g(X_t) - \theta\right)^T .$$

By the strong law for Harris ergodic Markov chains, $\Lambda_n \to \Lambda$ with probability 1 as $n \to \infty$. Since $|\cdot|$ is a continuous function, $|\Lambda_n|^{1/2p} \to |\Lambda|^{1/2p}$ with probability 1 as $n \to \infty$. As long as $n$ is larger than $p$, $\Lambda_n$ is positive definite, and thus $|\Lambda_n|^{1/2p} > 0$. Thus we set

$$K_n(g(X), p) = |\Lambda_n|^{1/2p} .$$

The reason for choosing the relative metric to be $|\Lambda|^{1/2p}$ is two-fold. First, it seems reasonable that when comparing the volume of an ellipsoid to an attribute of the estimation process, both quantities should be in the same scale and units. By (3.1), the term $\operatorname{Vol}(C_\alpha(n))^{1/p}$ is proportional to $|\Sigma_n|^{1/2p}$. Both $\Lambda_n$ and $\Sigma_n$ estimate variability associated with $g$ under $F$; $\Lambda_n$ estimates variability when the samples are i.i.d and $\Sigma_n$ estimates variability when the samples are obtained through MCMC.

Second, our choice of relative metric leads to a nice interpretation of the stopping rule. That is, $T^*(\epsilon)$ is the first time the uncertainty in estimation (measured via the volume of the confidence region) is an $\epsilon$th fraction of the uncertainty in the target distribution for $g$. If the uncertainty for $g$ under $F$ is large, then we allow for Monte Carlo variation to also be large, and if the variability for $g$ under $F$ is small, then tighter Monte Carlo estimates are desired. The *relative standard deviation fixed-*

*volume sequential stopping rule* is formalized as terminating at random time

$$T_{SD}(\epsilon) = \inf\left\{n \geq 0 : \mathrm{Vol}(C_\alpha(n))^{1/p} + \epsilon|\Lambda_n|^{1/2p} I(n < n^*) + n^{-1} \leq \epsilon|\Lambda_n|^{1/2p}\right\} . \quad (3.3)$$

## 3.2 Effective Sample Size

Recall the notation used in previous chapters: $\Sigma$ is the asymptotic covariance matrix in the Markov chain CLT with diagonals $\sigma_i^2$; $\Lambda$ is the covariance matrix for $g$ under the target distribution with diagonals $\lambda_i^2$.

As discussed in Chapter 1, a common way of terminating the simulation in MCMC is by using effective sample size (ESS). The ESS of a sample for estimating $\theta$ is the number of i.i.d samples with the same standard error as this sample.

For example, suppose interest is in estimating $\theta_1 = \mathrm{E}_F g_1$ from an MCMC sample of size $n$. The estimate used is $\theta_{n,1}$. Also suppose that i.i.d. samples of size $E_*$ can be drawn from $F$ and $\theta$ is then estimated using the mean of the sample, $\bar{\theta}$. Then the effective sample size for the MCMC sample is $E_*$ such that

$$\mathrm{Var}_F\, \theta_{n,1} \approx \frac{\lambda_1^2}{E_*} \Rightarrow \frac{\sigma_1^2}{n} \approx \frac{\lambda_1^2}{E_*} .$$

This definition has been formalized and presented by many authors. Kass et al. (1998), Liu (2008), and Robert and Casella (2013) define ESS for the $i$th component of the process as

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

where recall that $\rho_i(k)$ is the lag $k$ autocorrelation for the $i$th component of $g(X)$. It is challenging to estimate $\rho_i(k)$ consistently, especially for larger $k$. Alternatively,

Gong and Flegal (2016) rewrite the above definition as,

$$\text{ESS}_i = n\frac{\lambda_i^2}{\sigma_i^2},$$

Using this formulation, a consistent estimator of $\text{ESS}_i$ is obtained by using strongly consistent estimators of $\lambda_i^2$ and $\sigma_i^2$ via the sample variance ($\lambda_{n,i}^2$) and univariate batch means or spectral variance estimators ($\sigma_{n,i}^2$), respectively. The `R` package `coda` estimates $\text{ESS}_i$ by using the sample variance to estimate $\lambda_i^2$ but estimates $\sigma_i^2$ by determining the spectral density at frequency zero of an approximating autoregressive process. Thus, it does not estimate $\sigma_i^2$ directly. The `R` package `mcmcse` uses the univariate batch means method to estimate $\sigma_i^2$. The estimate of $\text{ESS}_i$ is then

$$\widehat{\text{ESS}}_i = n\frac{\lambda_{n,i}^2}{\sigma_{n,i}^2},$$

Then $\widehat{\text{ESS}}_i$ is a strongly consistent estimator of the univariate $\text{ESS}_i$. Users calculate $\widehat{\text{ESS}}_i$ for each of the $p$ components and set an ad-hoc lower bound for a sufficient effective sample size. When the smallest estimated effective sample size among all $p$ estimates is larger than the ad-hoc lower bound, it is deduced that enough Monte Carlo samples have been obtained. Gong and Flegal (2016) addressed the issue of the ad-hoc lower bound by presenting a theoretically valid lower bound when $p = 1$. However, the following two significant challenges remain:

1. An effective sample size needs to be calculated for each of the $p$ components and the smallest among them is chosen. This process is arduous and termination is delayed.

2. Effective sample size is a property not only of the correlated sample, but also (and more importantly) of the expectation being estimated. Using univariate effective sample size for the $i$th component implies interest is in $g_i$, thus ignoring all other $g_j$, $j \neq i$. Since interest is in estimating the whole vector $g$, univariate

estimation of effective sample size ignores all cross correlations between components of $\theta_n$, as discussed in Chapter 1.

A multivariate definition of effective sample size does not exist. We begin by defining this quantity and introduce strongly consistent estimators of it.

Instead of using the diagonals of $\Lambda$ and $\Sigma$ to define ESS, we use the matrices themselves. Let $S_p^+$ denote the set of all $p \times p$ positive definite matrices. Scalar quantification of the matrices requires a mapping $S_p^+ \to \mathbb{R}_+$ that captures the variability described by the covariance matrix. Wilks (1932) used the determinant as a univariate measure of spread for a multivariate distribution, and called the determinant of a covariance matrix of a distribution its *generalized variance*. Wilks (1932) recommended the use of the $p$th root of the generalized variance. This was formalized by SenGupta (1987) as the *standardized generalized variance*, to compare variability over different dimensions. We define

$$\text{ESS} = n \left( \frac{|\Lambda|}{|\Sigma|} \right)^{1/p}. \tag{3.4}$$

When $p = 1$, the ESS reduces to the form of univariate ESS presented above. When independent samples are obtained from $F$, the ESS is exactly $n$, as expected. Recall that $\Lambda_n$ is the sample covariance matrix of $\{g(X_t)\}$ and let $\Sigma_n$ be a strongly consistent estimator of $\Sigma$. Then a strongly consistent estimator of ESS is

$$\widehat{\text{ESS}} = n \left( \frac{|\Lambda_n|}{|\Sigma_n|} \right)^{1/p}.$$

Another interesting interpretation of ESS is that it is determined by the ratio of the geometric means of the eigenvalues of $\Lambda$ and $\Sigma$. The eigenvalues of a covariance matrix determine the amount of variability in the direction of the corresponding eigenvector. If the off diagonals of $\Sigma$ and $\Lambda$ are zero, then the multivariate effective sample size is

the geometric mean of all the $p$ univariate effective sample sizes, since

$$\mathrm{ESS} = n \prod_{i=1}^{p} \left( \frac{\lambda_i^2}{\sigma_i^2} \right)^{1/p} = \prod_{i=1}^{p} (\mathrm{ESS}_i)^{1/p} .$$

### 3.2.1 Relation to Termination Rules

Our choice of the relative metric in Section 3.1 helps us arrive at a lower bound on the number of effective samples required. Rearranging the defining inequality in (3.3) yields that when $n \geq n^*$

$$\begin{aligned}
\epsilon |\Lambda_n|^{1/2p} &\geq \mathrm{Vol}(C_\alpha(n))^{1/p} + n^{-1} \\
&= \left( \frac{2\pi^{p/2}}{p\Gamma(p/2)} \left( \frac{T^2_{1-\alpha,p,q}}{n} \right)^{p/2} |\Sigma_n|^{1/2} \right)^{1/p} + n^{-1} \\
&= \left( \frac{2\pi^{p/2}}{p\Gamma(p/2)} \right)^{1/p} \left( \frac{T^2_{1-\alpha,p,q}}{n} \right)^{1/2} |\Sigma_n|^{1/2p} + n^{-1} \\
\Rightarrow \sqrt{n} \frac{|\Lambda_n|^{1/2p}}{|\Sigma_n|^{1/2p}} &\geq \frac{\sqrt{n}}{\epsilon} \left( \frac{2\pi^{p/2}}{p\Gamma(p/2)} \right)^{1/p} \left( \frac{T^2_{1-\alpha,p,q}}{n} \right)^{1/2} + \frac{|\Sigma_n|^{-1/2p}}{\epsilon\sqrt{n}} \\
\Rightarrow \widehat{\mathrm{ESS}} &\geq \left[ \left( \frac{2\pi^{p/2}}{p\Gamma(p/2)} \right)^{1/p} \left( T^2_{1-\alpha,p,q} \right)^{1/2} + \frac{|\Sigma_n|^{-1/2p}}{n^{1/2}} \right]^2 \frac{1}{\epsilon^2} .
\end{aligned}$$

Thus, the relative standard deviation fixed-volume sequential stopping rule is equivalent to terminating the first time $\widehat{\mathrm{ESS}}$ is larger than a lower bound. This lower bound is difficult to determine before starting the simulation. However, as $n \to \infty$, $T^2_{p,q}$ converges in distribution to a $\chi^2_p$ and $\Sigma_n$ converges of $\Sigma$ with probability 1, leading to the following approximation

$$\widehat{\mathrm{ESS}} \geq \frac{2^{2/p}\pi}{(p\Gamma(p/2))^{2/p}} \frac{\chi^2_{1-\alpha,p}}{\epsilon^2} . \tag{3.5}$$

Due to (3.5), one can *a priori* determine the number of effective samples required for the choice of $\epsilon$ and $\alpha$. That is, the number of effective samples required depends on

the relative tolerance level $\epsilon$ and the confidence level determined by $\alpha$. The lower bound is not affected by the Markov chain or the target distribution but the observed ESS is. Slow converging Markov chains will have high correlations, and thus smaller $\widehat{\text{ESS}}$, taking longer to reach the desired lower bound. As $p \to \infty$,

$$\frac{2^{2/p}\pi}{(p\Gamma(p/2))^{2/p}} \frac{\chi^2_{1-\alpha,p}}{\epsilon^2} \to \frac{2\pi e}{\epsilon^2}\,.$$

Thus for large $p$, the lower bound is mainly determined by the choice of $\epsilon$. The choice of $\epsilon$ should be made keeping in mind that the samples obtained are only *approximately* from $F$. On the other hand, for a fixed $\alpha$, having obtained $W$ effective samples, the user can use the lower bound to understand the level of precision ($\epsilon$) in their estimation. In this way, (3.5) can be used to make informed decisions regarding termination.

**Example 3.1**

Suppose $p = 5$ (as in the Bayesian logistic regression setting of Chapter 1) and that we want a precision of $\epsilon = .05$ (so the Monte Carlo error is 5% of the uncertainty in the target distribution) for a 95% confidence region. This requires $\widehat{\text{ESS}} \geq 8605$. On the other hand, if we simulate until $\widehat{\text{ESS}} = 10000$, we obtain a precision of $\epsilon = .0464$.□

# Chapter 4

# Estimating Monte Carlo Standard Error

Multivariate analysis of the output generated from MCMC requires estimating the covariance matrix in the Markov chain CLT, $\Sigma$. In this chapter we present two estimators of $\Sigma$ and provide conditions for strong consistency.

Recall that $X = \{X_t\}$ denotes the Markov chain with invariant distribution $F$ having support $\mathcal{X}$ equipped with a countably generated $\sigma$-field. In addition, $g$ is an $F$-integrable function such that $g : \mathcal{X} \to \mathbb{R}^p$, and interest is in estimating $\theta = \mathrm{E}_F g$. The estimator of choice is

$$\theta_n = \frac{1}{n} \sum_{t=1}^{n} g(X_t) \, .$$

Also recall the Markov chain is polynomially ergodic of order $m$ where $m > 0$ if there exists $M : \mathcal{X} \to \mathbb{R}^+$ with $E_F M < \infty$ such that

$$\| P^n(x, \cdot) - F(\cdot) \|_{TV} \leq M(x) \, n^{-m} \, .$$

## 4.1 Multivariate Spectral Variance Estimator

Let $Y_t = g(X_t) - \theta$ for $t \in \mathbb{N}$ and define the lag $s$, $s \geq 0$, autocovariance matrix as

$$\gamma(s) = \gamma(-s)^T = \mathrm{E}_F \left[ Y_t \, Y_{t+s}^T \right] .$$

Define $I_s$ as $I_s = \{1, \ldots, (n-s)\}$ for $s \geq 0$ and as $I_s = \{(1-s), \ldots, n\}$ for $s < 0$. Let $\bar{Y}_n = n^{-1} \sum_{t=1}^{n} Y_t$ and define the lag $s$ sample autocovariance as

$$\gamma_n(s) = \frac{1}{n} \sum_{t \in I_s} (Y_t - \bar{Y}_n)(Y_{t+s} - \bar{Y}_n)^T . \tag{4.1}$$

From the structure of $\Sigma$ in (2.3), it is known that

$$\Sigma = \sum_{s=-\infty}^{\infty} \gamma(s) .$$

Replacing $\gamma(s)$ with $\gamma_n(s)$ leads to a strongly consistent estimator. However, this estimator has poor finite sample properties (see Anderson (1971)) since $\gamma_n(s)$ is a poor estimator for large $s$. The multivariate spectral variance (mSV) estimator is defined as a weighted and truncated sum of the lag $s$ sample autocovariances,

$$\Sigma_{SV} = \sum_{s=-(b_n-1)}^{b_n-1} w_n(s) \gamma_n(s), \tag{4.2}$$

where $w_n(\cdot)$ is the *lag window* and $b_n$ is the *truncation point*. The lag window has to satisfy the following additional conditions.

**Condition 4.1** The lag window $w_n(\cdot)$ is an even function defined on $\mathbb{Z}$ such that

(a) $|w_n(s)| \leq 1$ for all $n$ and $s$,

(b) $w_n(0) = 1$ for all $n$, and

(c) $w_n(s) = 0$ for all $|s| \geq b_n$. □

Anderson (1971) gives a list of lag windows that satisfy Condition 4.1. We will consider some of these later.

Under Condition 4.1 and using the fact that $\gamma_n(0)$ is symmetric, $\Sigma_{SV}$ is symmetric.

$$\begin{aligned}
\Sigma_{SV} &= \sum_{s=-(b_n-1)}^{b_n-1} w_n(s)\gamma_n(s) \\
&= w_n(0)\gamma_n(0) + \sum_{s=1}^{b_n-1} w_n(s)\gamma_n(s) + \sum_{s=1}^{b_n-1} w_n(-s)\gamma_n(-s) \\
&= w_n(0)\gamma_n(0) + \sum_{s=1}^{b_n-1} w_n(s)\left[\gamma_n(s) + \gamma_n(s)^T\right] \\
&= w_n(0)\gamma_n(0)^T + \sum_{s=1}^{b_n-1} w_n(s)\left[\gamma_n(s)^T + \gamma_n(s)\right] \\
&= \Sigma_{SV}^T \,.
\end{aligned}$$

For the univariate setting, where $\Sigma$ is a scalar, the spectral variance estimator has been well studied. Damerdji (1991) showed strong consistency of the estimator for general stochastic processes. Their estimator was adapted to the context of MCMC and the conditions weakened by Flegal and Jones (2010). Atchadé (2011) also proved strong consistency of the univariate estimator for adaptive MCMC samplers.

mSV estimators have also been studied in the time series literature. They are often used for heteroscedastic and autocorrelation consistent (HAC) estimation of covariance matrices which, for example, arise in the study of generalized method of moments and autoregressive processes with heteroscedastic errors. See Andrews (1991) for motivating examples. In the context of HAC estimation, De Jong (2000) obtained conditions under which the class of mSV estimators is strongly consistent. However, these conditions are restrictive in the context of MCMC. In particular, his Assumption 2 (De Jong, 2000, page 264) will not be satisfied in many typical MCMC applications. Additionally, we require weaker mixing conditions on the underlying

stochastic process.

To prove strong consistency of $\Sigma_{SV}$ we require the existence of a strong invariance principle for both $g(X_t)$ and $h(X_t) = [g(X_t) - \theta]^2$, where the square is taken element-wise. That is, in addition to (2.4), we assume that there exists a finite $p$-vector $\theta_h$, a $p \times p$ lower triangular matrix $L_h$, an increasing function $\psi_h$ on the integers, a finite random variable $D_h$, and a sufficiently rich probability space such that, with probability 1,

$$\left\| \sum_{t=1}^{n} h(X_t) - n\theta_h - L_h B(n) \right\| < D_h \, \psi_h(n) \, . \tag{4.3}$$

The following Conditions 4.2 and 4.3 are technical conditions ensuring that $b_n$ grows at the right rate compared to $n$.

**Condition 4.2** Let $b_n$ be an integer sequence such that $b_n \to \infty$ and $n/b_n \to \infty$ as $n \to \infty$ where $b_n$ and $n/b_n$ are non-decreasing. □

**Condition 4.3** Let $b_n$ be an integer sequence such that

(a) there exists a constant $c \geq 1$ such that $\sum_n (b_n/n)^c < \infty$,

(b) $b_n n^{-1} \log n \to 0$ as $n \to \infty$,

(c) $b_n^{-1} \log n = O(1)$, and

(d) $n > 2b_n$. □

If $b_n = \lfloor n^\nu \rfloor$, where $0 < \nu < 1$, then Condition 4.3 is satisfied if $n > 2^{1/(1-\nu)}$.

Define

$$\Delta_1 w_n(k) = w_n(k-1) - w_n(k)$$

and

$$\Delta_2 w_n(k) = w_n(k-1) - 2w_n(k) + w_n(k+1) \; .$$

**Condition 4.4** Let $b_n$ be an integer sequence, $w_n$ be the lag window, and $\psi(n)$ and $\psi_h(n)$ be positive functions on the integers such that,

(a) $b_n n^{-1} \sum_{k=1}^{b_n} k|\Delta_1 w_n(k)| \to 0$ as $n \to \infty$,

(b) $b_n \psi(n)^2 \log n \left( \sum_{k=1}^{b_n} |\Delta_2 w_n(k)| \right)^2 \to 0$ as $n \to \infty$,

(c) $\psi(n)^2 \sum_{k=1}^{b_n} |\Delta_2 w_n(k)| \to 0$ as $n \to \infty$,

(d) $b_n^{-1} \psi_h(n) \to 0$ as $n \to \infty$, and

(e) $b_n^{-1} \psi(n) \to 0$ as $n \to \infty$. □

Condition 4.4a connects the truncation point $b_n$ to the lag window $w_n$. Later we will present examples of lag windows that satisfy this condition. The functions $\psi(n)$ and $\psi_h(n)$ in Conditions 4.4b, 4.4c, 4.4d, and 4.4e correspond to the functions described in (2.4) and (4.3) and thus these four conditions connect the truncation point $b_n$, the lag window $w_n$, and the correlation of the process, measured indirectly by $\psi(n)$ and $\psi_h(n)$. In Lemma 4.1 below we present sufficient conditions for Conditions 4.4a, 4.4b, and 4.4c.

**Lemma 4.1** Reparameterize $w_n$ such that $w_n$ is defined on $[0, 1]$ and $w_n(0) = 1$ and $w_n(1) = 0$. Further assume that $w_n$ is twice continuously differentiable and that there exists finite constants $D_1$ and $D_2$ such that $|w_n'(x)| \leq D_1$ and $|w_n''(x)| < D_2$. Then as $n \to \infty$,

1. Condition 4.4a holds if $b_n^2 n^{-1} \to 0$,

2. Conditions 4.4b and 4.4c holds if $b_n^{-1}\psi(n)^2 \log n \to 0$. □

The following theorem demonstrates strong consistency of $\Sigma_{SV}$ for processes that satisfy a strong invariance principle.

**Theorem 4.1** Suppose the strong invariance principles (2.4) and (4.3) hold. If Conditions 4.1, 4.2, 4.3, and 4.4 hold, then $\Sigma_{SV} \to \Sigma$, with probability 1, as $n \to \infty$. □

Theorem 4.1 holds for all processes that satisfy the strong invariance principles as stated. Next using Corollary 2.1, we present strong consistency of $\Sigma_{SV}$ when the underlying process is a Harris ergodic Markov chain.

**Theorem 4.2** Suppose $\mathrm{E}_F\|g\|^{4+\delta} < \infty$ for some $\delta > 0$. Let $X$ be a polynomially ergodic Markov chain of order $m \geq (1+\epsilon_1)(1+2/\delta)$ for some $\epsilon_1 > 0$. Then (2.4) and (4.3) hold with

$$\psi(n) = \psi_h(n) = n^{1/2-\lambda},$$

for some $\lambda > 0$ that depends on $p$, $\epsilon$, and $\delta$. If Conditions 4.1, 4.2, 4.3, and 4.4 hold, then $\Sigma_{SV} \to \Sigma$, with probability 1, as $n \to \infty$. □

**Remark 4.1** When $p = 1$, the mSV estimator reduces to the spectral variance estimator (SV) considered by Atchadé (2011), Damerdji (1991), and Flegal and Jones (2010). In this case our result requires weaker conditions. First notice that Flegal and Jones (2010) required weaker conditions than Damerdji (1991). Thus we only need to compare Theorem 4.2 to the results in Atchadé (2011) and Flegal and Jones (2010), both of whom required the Markov chains to be geometrically ergodic and to satisfy a one-step minorization condition. Thus Theorem 4.2 substantially weakens

the conditions on the underlying Markov chain, while extending the results to the $p \geq 1$ setting. □

**Remark 4.2** It is common to use $b_n = \lfloor n^\nu \rfloor$ in which case Conditions 4.4a, 4.4b, and 4.4c hold, if we choose $0 < \nu < 1/2$ such that $n^{-\nu}\psi(n)^2 \log n \to 0$ as $n \to \infty$. □

**Remark 4.3** We now consider some examples of lag windows that satisfy Condition 4.1 and consider whether Conditions 4.4a, 4.4b, and 4.4c hold.

1. *Simple Truncation*: $w_n(k) = I(|k| < b_n)$. Using this window the estimator obtained is truncated at $b_n$ but weighted identically. In this case, $\Delta_2 w_n(k) = 0$ for $k = 1, \ldots, b_n - 2$, $\Delta_2 w_n(b_n - 1) = -1$ and $\Delta_2 w_n(b_n) = 1$. It is easy to see that Condition 4.4c is not satisfied.

2. *Blackman-Tukey*: $w_n(k) = [1 - 2a + 2a\cos(\pi|k|/b_n)]\, I(|k| < b_n)$ where $a > 0$. This is a generalization for the *Tukey-Hanning* window where $a = 1/4$. For fixed $a$, the Blackman-Tukey window satisfies the conditions of Lemma 4.1, thus Conditions 4.4a, 4.4b, and 4.4c hold if $b_n^2 n^{-1} \to 0$ and $b_n^{-1}\psi(n)^2 \log n \to 0$ as $n \to \infty$.

3. *Parzen*: $w_n(k) = [1 - |k|^q / b_n^q]\, I(|k| < b_n)$ for $q \in \mathbb{Z}^+$. When $q = 1$ this is the *modified Bartlett* window. It is easy to show that the Parzen window satisfies the conditions for Lemma 4.1, and thus Conditions 4.4a, 4.4b, and 4.4c hold if $b_n^2 n^{-1} \to 0$ and $b_n^{-1}\psi(n)^2 \log n \to 0$ as $n \to \infty$.

4. *Scale-parameter modified Bartlett*: $w_n(k) = [1 - \eta|k|/b_n]\, I(|k| < b_n)$ where $\eta$ is a positive constant not equal to 1. Then $\Delta_1 w_n(k) = \eta b_n^{-1}$ for $k = 1, 2, \ldots, b_n - 1$ and $\Delta_1 w_n(b_n) = 1 - \eta + \eta b_n^{-1}$ so that Condition 4.4a is satisfied when $b_n^2 n^{-1} \to 0$ as $n \to \infty$. Also, $\Delta_2 w_n(k) = 0$ for $k = 1, 2, \ldots, b_n - 2$, $\Delta_2 w_n(b_n - 1) = \eta - 1$ and

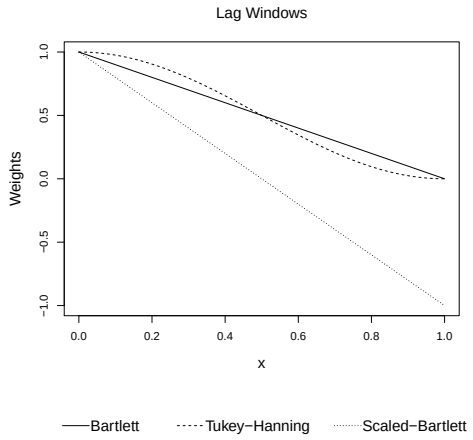



Figure 4.1: Plot of three lag windows, modified Bartlett(Bartlett), Tukey-Hanning and the scale-parameterBartlett with scale parameter 2 (Scaled-Bartlett).

$\Delta_2 w_n(b_n) = 1 - \eta + \eta b_n^{-1}$. We conclude that $\sum_{k=1}^{b_n} |\Delta_2 w_n(k)|$ does not converge to 0 and hence Condition 4.4c is not satisfied. □

Figure 4.1 provides a graph of three lag windows specifically, the modified Bartlett, Tukey-Hanning, and scale-parameter modified Bartlett windows. It is evident that the modified Bartlett and Tukey-Hanning windows are similar and the scale-parameter modified Bartlett window weighs the lags more severely.

## 4.2 Multivariate Batch Means Estimator

Let $n = a_n b_n$, where $a_n$ denotes the number of batches and $b_n$ is the batch size. For $k = 0, \ldots, a_n - 1$, define $\bar{g}_k := b_n^{-1} \sum_{t=1}^{b_n} g(X_{kb_n+t})$. Then $\bar{g}_k$ is the mean vector for batch $k$ and the mBM estimator of $\Sigma$ is given by

$$\Sigma_{BM} = \frac{b_n}{a_n - 1} \sum_{k=0}^{a_n-1} (\bar{g}_k - \theta_n)(\bar{g}_k - \theta_n)^T . \tag{4.4}$$

Since $\Sigma$ is non-singular $\Sigma_{BM}$ should be non-singular, which requires $a_n > p$.

When $g(X_t)$ is univariate, the batch means estimator has been well studied for MCMC problems (Jones et al., 2006; Flegal and Jones, 2010) and for steady state simulations (Damerdji, 1994; Glynn and Iglehart, 1990; Glynn and Whitt, 1991). Glynn and Whitt (1991) showed that the batch means estimator cannot be consistent for fixed batch size, $b_n$. Damerdji (1994, 1995), Jones et al. (2006) and Flegal and Jones (2010) established its asymptotic properties including strong consistency and mean square consistency when *both* the batch size and number of batches increases with $n$.

The multivariate extension as in (4.4) was first introduced by Chen and Seila (1987). For steady-state simulation output Charnes (1995) and Muñoz and Glynn (2001) studied confidence regions for $\theta$ based on the mBM, however, the theoretical properties of mBM remain unexplored.

In Theorem 4.3, we present conditions for strong consistency of $\Sigma_{BM}$ in estimating $\Sigma$ for MCMC. Aside from the existence of a strong invariance principle, we require the following additional conditions on the batch size $b_n$.

**Condition 4.5** The batch size $b_n$ satisfies the following conditions,

(a) the batch size $b_n$ is an integer sequence such that $b_n \to \infty$ and $n/b_n \to \infty$ as $n \to \infty$ where, $b_n$ and $n/b_n$ are monotonically increasing,

(b) there exists a constant $c \geq 1$ such that $\sum_n \left(b_n n^{-1}\right)^c < \infty$. □

**Theorem 4.3** Let $g$ be such that $E_F\|g\|^{2+\delta} < \infty$ for some $\delta > 0$. Let $X$ be an $F$-invariant polynomially ergodic Markov chain of order $m > (1+\epsilon_1)(1+2/\delta)$ for some $\epsilon_1 > 0$. Then (2.4) holds with $\psi(n) = n^{1/2-\lambda}$ for some $\lambda > 0$. If Condition 4.5 holds and $b_n^{-1/2}(\log n)^{1/2} n^{1/2-\lambda} \to 0$ as $n \to \infty$, then $\Sigma_{BM} \to \Sigma$, with probability 1, as $n \to \infty$. □

**Remark 4.4** The theorem holds more generally outside the context of Markov chains for processes that satisfy (2.4). This includes independent processes (Berkes and Philipp, 1979; Einmahl, 1989; Zaitsev, 1998), Martingale sequences (Eberlein, 1986), renewal processes (Horvath, 1984) and $\phi$-mixing and strongly mixing processes (Kuelbs and Philipp, 1980; Dehling and Philipp, 1982). The general statement of the theorem is provided in Appendix B.3. □

**Remark 4.5** It is natural to consider $b_n = \lfloor n^\nu \rfloor$ for $0 < \nu < 1$. Then $\nu > 1 - 2\lambda$ is required to satisfy $b_n^{-1/2}(\log n)^{1/2} n^{1/2-\lambda} \to 0$ as $n \to \infty$. Hence, for fast mixing processes smaller batch sizes suffice and slow mixing processes require larger batch sizes. This reinforces our intuition that higher correlation calls for larger batch sizes. Calibrating $\nu$ in $b_n = \lfloor n^\nu \rfloor$ is essential to ensuring the mBM estimates perform well in finite samples. Using mean square consistency of univariate batch means estimators, Flegal and Jones (2010) concluded that an asymptotically optimal batch size is proportional to $\lfloor n^{1/3} \rfloor$ with an unknown proportionality constant. □

**Remark 4.6** For $p = 1$, Jones et al. (2006) proved strong consistency of the batch means estimator under the stronger assumption of geometric ergodicity and a one-step minorization, which we do not make. Thus, in Theorem 4.3 while extending the result of strong consistency to $p \geq 1$, we also weaken the conditions for the univariate case. □

## 4.3 Strong Consistency of Eigenvalues

**Theorem 4.4** Let $\widehat{\Sigma}$ be any strongly consistent estimator of $\Sigma$ and let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$ be the eigenvalues of $\Sigma$. Let $\hat{\lambda}_1, \ldots, \hat{\lambda}_p$ be the $p$ eigenvalues of $\widehat{\Sigma}$ such that $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p$, then $\hat{\lambda}_k \to \lambda_k$, with probability 1, as $n \to \infty$ for all

$1 \leq k \leq p$. □

**Proof**

Let $\|\cdot\|_F$ denote the Frobenius norm. By Weyl's inequality (Franklin, 2012), for $\epsilon_1 > 0$, if $\|\widehat{\Sigma} - \Sigma\|_F \leq \epsilon_1$, then for all $1 \leq k \leq p$, $|\hat{\lambda}_k - \lambda_k| \leq \epsilon_1$, which gives the desired result. □

**Remark 4.7** Theorem 4.4 immediately implies that under the conditions of Theorems 4.1, 4.2, and 4.3, the sample eigenvalues of the mSV and mBM estimators are consistent for the population eigenvalues. □

Sample eigenvalues can play an important role in multivariate analyses. For example, the length of any axis of the confidence region constructed from an estimator of $\Sigma$ is determined by the magnitude of the relevant estimated sample eigenvalues. Thus the largest eigenvalue is associated with the axis having the largest estimated Monte Carlo error. This also suggests that dimension reduction methods could be useful in assessing the reliability of the simulation effort.

## 4.4 mBM Versus mSV

We have provided two classes of estimators in this chapter, the multivariate batch means (mBM) estimator and the multivariate spectral variance (mSV) estimator. Both classes of estimators were shown to be strongly consistent under conditions on the Markov chain and certain moment conditions.

Both mSV and mBM estimators require only polynomial ergodicity, however the mSV estimator requires $4+\delta$ finite moments of $g$ under $F$, whereas the mBM estimator only requires $2+\delta$ finite moments for some $\delta > 0$. Thus the mBM estimator requires a weaker moment condition. However, for the univariate case Flegal and Jones (2010)

showed that the SV estimator is more statistically efficient than the BM estimator and thus it is likely that that result is also true for the multivariate case. We do not pursue this here as it is outside the scope and requires much additional work.

Computationally, the mBM estimator is significantly faster than the mSV estimator. We investigate finite sample properties and time comparisons by using a vector autoregressive process of order 1.

### 4.4.1 Vector Autoregressive Process

Consider the vector autoregressive process of order 1 (VAR(1)). For $t = 1, 2, \ldots,$

$$Y_t = \Phi Y_{t-1} + \epsilon_t,$$

where $Y_t \in \mathbb{R}^p$, $\Phi$ is a $p \times p$ matrix, $\epsilon_t \overset{iid}{\sim} N_p(0, \Omega)$, and $\Omega$ is a $p \times p$ positive definite matrix. The matrix $\Phi$ determines the nature of the autocorrelation. This Markov chain has invariant distribution $F = N_p(0, V)$ where $vec(V) = (I_{p^2} - \Phi \otimes \Phi)^{-1} vec(\Omega)$, $\otimes$ denotes the Kronecker product, and is geometrically ergodic when the spectral radius of $\Phi(\phi_{\max})$ is less than 1 (Tjøstheim, 1990).

Consider the goal of estimating the mean of $F$, $\mathrm{E}_F Y = 0$ with the Monte Carlo estimator. Note that,

$$\begin{aligned}
Y_s &= \Phi Y_{s-1} + \epsilon_s \\
&= \Phi(\Phi Y_{s-2} + \epsilon_{s-1}) + \epsilon_s \\
&= \Phi^2 Y_{s-2} + \Phi\epsilon_{s-1} + \epsilon_s \\
&= \Phi^2(\Phi Y_{s-3} + \epsilon_{s-2}) + \Phi\epsilon_{s-1} + \epsilon_s \\
&= \Phi^3 Y_{s-3} + \Phi^2\epsilon_{s-2} + \Phi\epsilon_{s-1} + \epsilon_s \\
&\vdots \\
&= \Phi^s Y_0 + \Phi^{s-1}\epsilon_1 + \Phi^{s-2}\epsilon_2 + \cdots + \Phi^2\epsilon_{s-2} + \Phi\epsilon_{s-1} + \epsilon_s \ .
\end{aligned}$$

Thus

$$\begin{aligned}
\gamma(s) &= \mathrm{Cov}_F(Y_0, Y_s) \\
&= \mathrm{Cov}_F(Y_0,\ \Phi^s Y_0 + \Phi^{s-1}\epsilon_1 + \Phi^{s-2}\epsilon_2 + \cdots + \Phi^2\epsilon_{s-2} + \Phi\epsilon_{s-1} + \epsilon_s) \\
&= \mathrm{Cov}_F(Y_0, \Phi^s Y_0) \\
&= \Phi^s \,\mathrm{Cov}_F(Y_0, Y_0) \\
&= \Phi^s V\,.
\end{aligned}$$

Similarly $\gamma(-s) = V(\Phi^T)^s$. Since $F$ has a moment generating function, a CLT holds with

$$\begin{aligned}
\Sigma &= \sum_{s=-\infty}^{\infty} \gamma(s) \\
&= \sum_{s=0}^{\infty} \gamma(s) + \sum_{s=-\infty}^{0} \gamma(s) - V \\
&= \sum_{s=0}^{\infty} \Phi^s V + \sum_{s=-\infty}^{0} V(\Phi^T)^s - V \\
&= (I_p - \Phi)^{-1} V + V(I_p - \Phi)^{-1} - V\,. \qquad (4.5)
\end{aligned}$$

We will investigate the finite sample properties of the mSV and mBM estimators of $\Sigma$ by comparing six different estimators:

- mBM with $b_n = \lfloor n^{1/3} \rfloor$.
- mBM with $b_n = \lfloor n^{1/2} \rfloor$.
- mSV: Bartlett lag window with $b_n = \lfloor n^{1/3} \rfloor$.
- mSV: Bartlett lag window with $b_n = \lfloor n^{1/2} \rfloor$.
- mSV: Tukey-Hanning lag window with $b_n = \lfloor n^{1/3} \rfloor$.
- mSV: Tukey-Hanning lag window with $b_n = \lfloor n^{1/2} \rfloor$.

| Setting | $p$ | $\Omega$ | Range of eigen($\Phi$) |
|---|---|---|---|
| 1 | 2 | $I_p$ | $[.01, .20)$ |
| 2 | 2 | $I_p$ | $[.40, .60)$ |
| 3 | 2 | $I_p$ | $[.70, .90)$ |
| 4 | 10 | $I_p$ | $[.01, .20)$ |
| 5 | 10 | $I_p$ | $[.40, .60)$ |
| 6 | 10 | $I_p$ | $[.70, .90)$ |
| 7 | 50 | $I_p$ | $[.01, .20)$ |
| 8 | 50 | $I_p$ | $[.40, .60)$ |
| 9 | 50 | $I_p$ | $[.70, .90)$ |

Table 4.1: Simulation settings 1 through 9. The eigenvalues of $\Phi$ are spaced equally in each interval.

We set

$$\Omega_1 = I_p \text{ and } \Omega_2 = \text{AR}(.5),$$

where AR(.5) is the first order autoregressive covariance matrix with correlation $\rho = 0.5$. For each $\Omega$, we generate 9 simulation settings based on the choice of $\Phi$ and $p$. The results for both choices of $\Omega$ were similar and thus we only show results for $\Omega_1 = I_p$. The settings are presented in Table 4.1. For Settings 1, 4, and 7, $\phi_{\max} = .2$, Settings 2, 5, and 8, $\phi_{\max} = .6$ and Settings 3, 6, and 9, $\phi_{\max} = .9$. Thus, these three sets of settings yield processes with different mixing rates.

For each setting, we do the following in each of 100 independent replications. We observe the process for a Monte Carlo sample size of $10^5$, and at samples $\{10^4, 5 \times 10^4, 10^5\}$ calculate the estimate of $\Sigma$ using the six estimators presented earlier. The error in estimation is determined by calculating the average relative difference in Frobenius norm, i.e. if $\widehat{\Sigma}$ is one of the six estimators of $\Sigma$,

$$\text{Error} = \|\widehat{\Sigma} - \Sigma\|_F / \|\Sigma\|_F \,.$$

Figure 4.2 shows the error in the estimation of $\Sigma$ versus the Monte Carlo sample size for each of the nine settings. The dark circles are for all estimators with $b_n = \lfloor n^{1/2} \rfloor$ and the hollow circles are for all estimators with $b_n = \lfloor n^{1/3} \rfloor$. The key feature to note is that generally, the mSV estimators perform better than the mBM estimators, with the Tukey lag window being the best. This behavior is expected from what is known in the univariate case. Also interesting is that there is a clear separation between the two $b_n$, with $b_n = \lfloor n^{1/3} \rfloor$ being the best for when $\phi_{\max}$ is not large and $b_n = \lfloor n^{1/2} \rfloor$ being the best for when $\phi_{\max}$ is large. Thus, it seems that tuning $b_n$ is more important than the choice of the estimator. The effect of $p$ seems minimal.

Next, in Figure 4.3 and Figure 4.4 we plot the density of the estimator of the largest eigenvalue of the six estimators and compare it to the truth. The estimates are calculated from a Monte Carlo sample size of $10^5$. The main point to note again is that larger batch sizes lead to better estimation when there is relatively high correlation in the process. It is also apparent for large $p$ than $10^5$ Monte Carlo samples might not be enough to obtain good estimates.

Finally, we compare the performance of the estimators with regard to the computing time. Table 4.2 shows the average time required to calculate the six estimators for setting 7. There is no doubt that the mBM estimator is significantly faster to compute. In addition, better estimation at larger batch sizes for the mSV estimator clearly comes at a computational cost.

Due to the results of Table 4.2, we will only consider the mBM estimator in our examples in the next chapter. This is since often either $p$ is large or the Monte Carlo sample is large so as to make it difficult to use the mSV estimator.

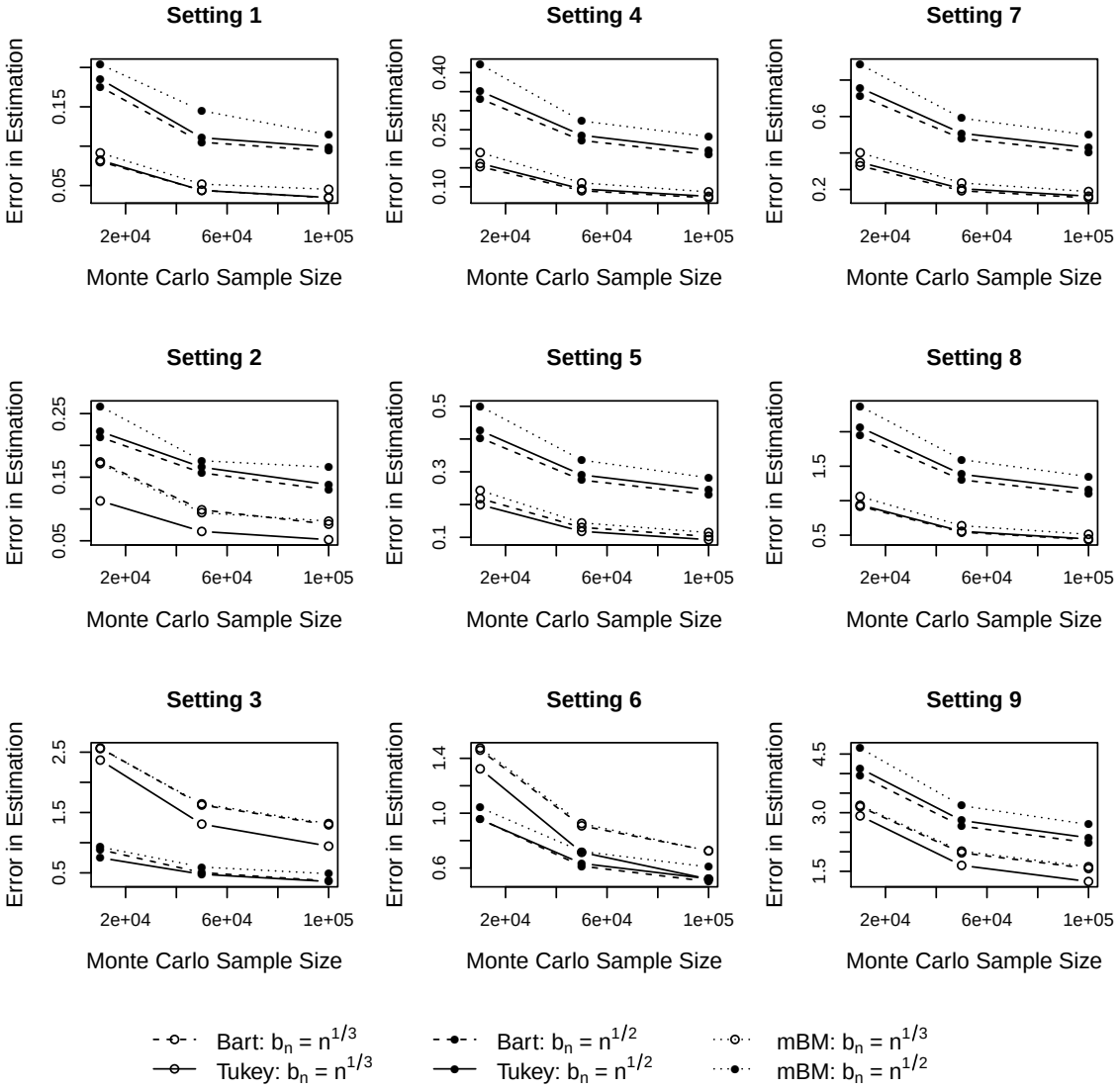
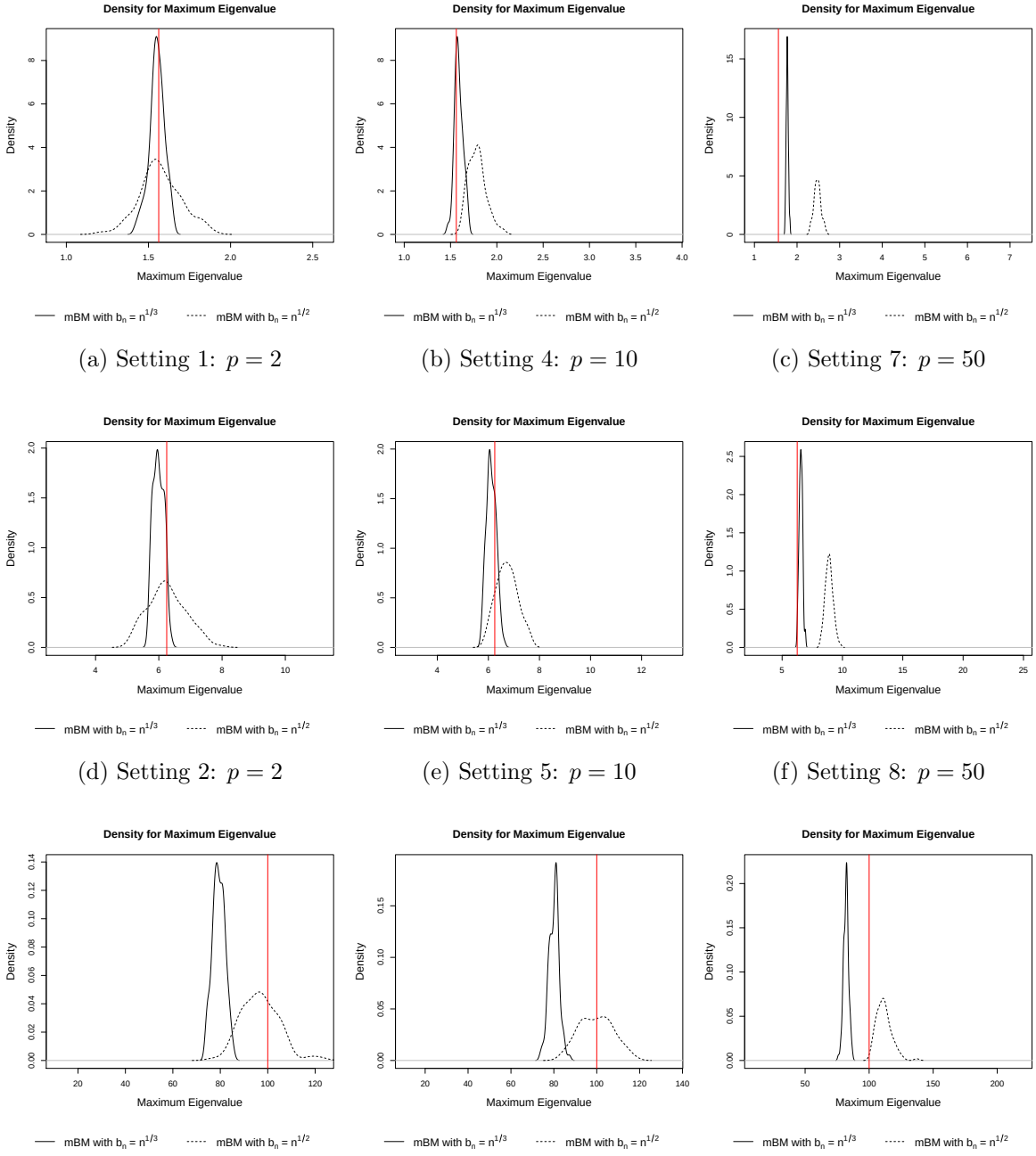



Figure 4.2: For $\Omega_1 = I_p$ we plot $\|\widehat{\Sigma} - \Sigma\|_F / \|\Sigma\|_F$ versus the Monte Carlo sample size for all nine settings. Standard errors were small.

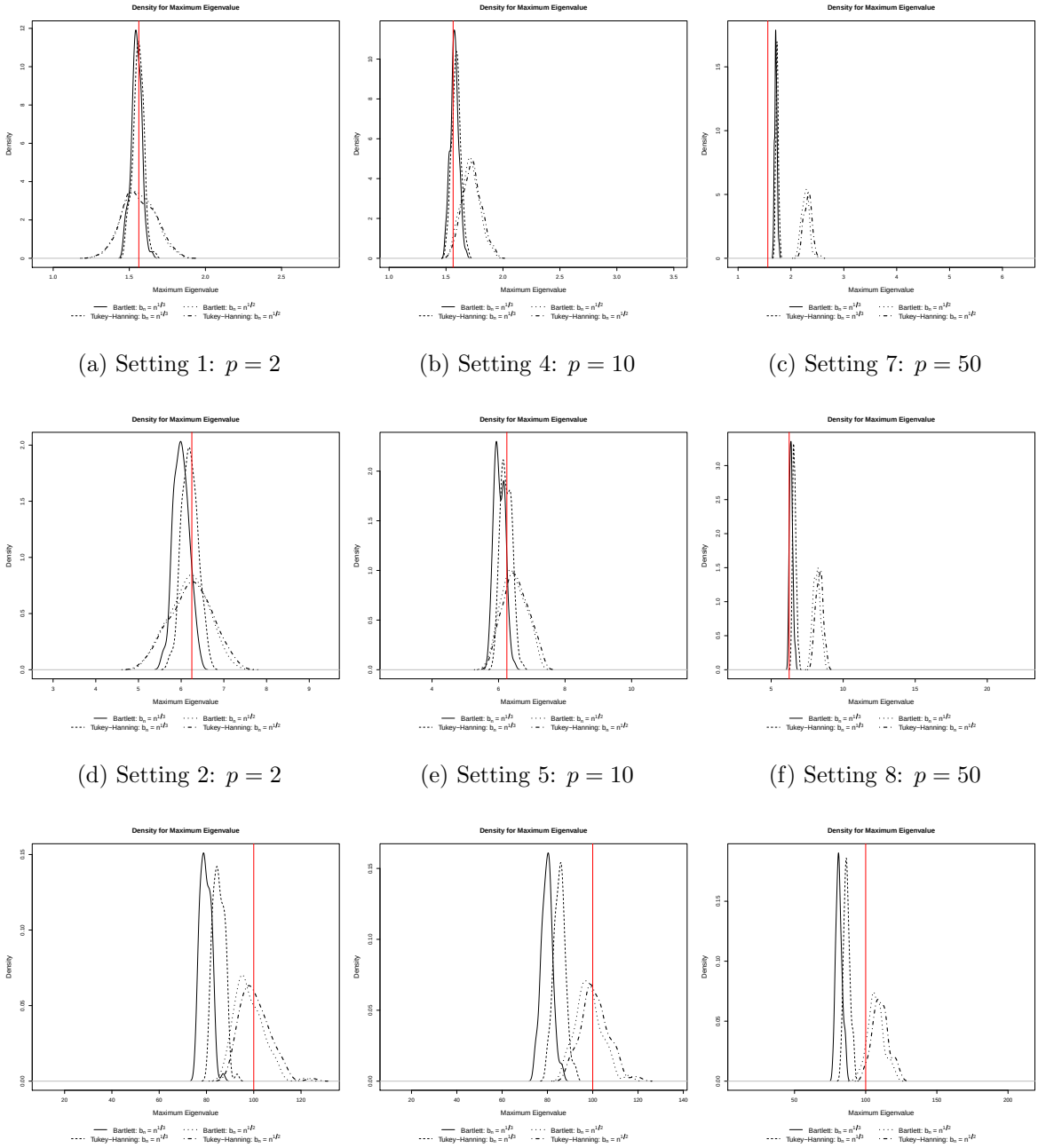



(a) Setting 1: $p = 2$ (b) Setting 4: $p = 10$ (c) Setting 7: $p = 50$

(d) Setting 2: $p = 2$ (e) Setting 5: $p = 10$ (f) Setting 8: $p = 50$

(g) Setting 3: $p = 2$ (h) Setting 6: $p = 10$ (i) Setting 9: $p = 50$

Figure 4.3: mBM for $\Omega_1$: Kernel density of the maximum eigenvalue for the mBM estimator for two batch lengths over 100 replications and Monte Carlo sample size $= 10^5$. The vertical line indicates the true eigenvalue. The first row is $\phi_{\max} = .20$ the second $\phi_{\max} = .60$, and the third $\phi_{\max} = .90$. It is clear that as mixing worsens, larger batch sizes are preferred.




(a) Setting 1: $p = 2$ (b) Setting 4: $p = 10$ (c) Setting 7: $p = 50$

(d) Setting 2: $p = 2$ (e) Setting 5: $p = 10$ (f) Setting 8: $p = 50$

(g) Setting 3: $p = 2$ (h) Setting 6: $p = 10$ (i) Setting 9: $p = 50$

Figure 4.4: mSV for $\Omega_1$: Kernel density of the maximum eigenvalue for the mSV estimators for all four lag window settings over 100 replications and Monte Carlo sample size = $10^5$. The vertical line indicates the true eigenvalue. The first row is $\phi_{\max} = .20$ the second $\phi_{\max} = .60$, and the third $\phi_{\max} = .90$. It is clear that as mixing worsens, larger batch sizes are preferred. The Tukey-Hanning window often performs slightly better.

| Estimator | $n = 10^3$ | $n = 10^4$ | $n = 10^5$ |
|---|---|---|---|
| | $b_n = \lfloor n^{1/3} \rfloor$ | | |
| mBM | 0.002 (0.0001) | 0.008 (0.0003) | 0.089 (0.0020) |
| Bartlett | 0.029 (0.0005) | 1.238 (0.0204) | 30.380 (0.4230) |
| Tukey-Hanning | 0.029 (0.0004) | 1.236 (0.0201) | 30.491 (0.4150) |
| | $b_n = \lfloor n^{1/2} \rfloor$ | | |
| mBM | 0.001 (0.0001) | 0.007 (0.0002) | 0.061 (0.0013) |
| Bartlett | 0.090 (0.0013) | 5.948 (0.0765) | 215.697 (2.3757) |
| Tukey-Hanning | 0.090 (0.0013) | 5.840 (0.0767) | 211.442 (2.7977) |

Table 4.2: Comparing computational time (in seconds) for setting 7 ($p = 50$) and $\Omega_1$ for the six estimators. Replications = 100 and standard errors are in parentheses.

# Chapter 5

# Examples

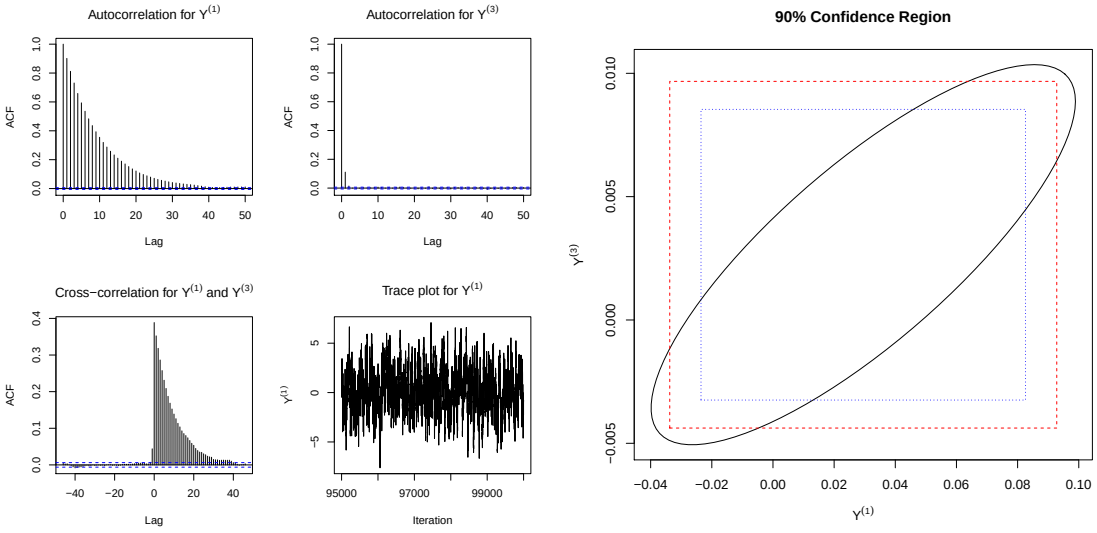
This chapter presents examples on which we test our multivariate termination rules. In each example we present a target distribution $F$, a Harris ergodic Markov chain with $F$ as its invariant distribution, we specify $g$, and are interested in estimating $\mathrm{E}_F g$. We consider the finite sample performance (based on 1000 independent replications) of the relative standard deviation fixed-volume sequential stopping rules and compare them to the relative standard deviation fixed-width sequential stopping rules (see Section 1.1). In each case we make 90% confidence regions for various choices of $\epsilon$ and specify our choice of $n^*$ and $b_n$. Since the stopping rules are sequential, theory dictates the termination criterion be checked at every new sample. This is quite impractical in real applications and so the sequential stopping rules are checked at 10% increments of the current Monte Carlo sample size.

## 5.1 Vector Autoregressive Process

We continue with the VAR(1) model and test our terminal rules. Recall that the VAR(1) process is defined for $t = 1, 2, \ldots,$ as

$$Y_t = \Phi Y_{t-1} + \epsilon_t,$$

where $Y_t \in \mathbb{R}^p$, $\Phi$ is a $p \times p$ matrix, $\epsilon_t \overset{iid}{\sim} N_p(0, \Omega)$, and $\Omega$ is a $p \times p$ positive definite




(a) (b)

Figure 5.1: VAR: (a) ACF plot for $Y^{(1)}$ and $Y^{(3)}$, CCF plot between $Y^{(1)}$ and $Y^{(3)}$, and trace plot for $Y^{(1)}$. Monte Carlo sample size is $10^5$. (b) Joint 90% confidence region for first the two components of $Y$. The solid ellipse is made using mBM, the dotted box using uBM uncorrected and the dashed line using uBM corrected by Bonferroni. Monte Carlo sample size is $10^5$.

matrix.

Consider the goal of estimating the mean of $F$, i.e. $\mathrm{E}_F Y = 0$, with $\bar{Y}_n$. Let $p = 5$, $\Phi = \mathrm{diag}(.9, .5, .1, .1, .1)$, and $\Omega$ be the AR(1) covariance matrix with autocorrelation 0.9. Since the first eigenvalue of $\Phi$ is large, the first component mixes slowest. We sample the process for $10^5$ iterations and in Figure 5.1a present the ACF plot for $Y^{(1)}$ and $Y^{(3)}$ and the CCF plot between $Y^{(1)}$ and $Y^{(3)}$ in addition to the trace plot for $Y^{(1)}$. Notice that $Y^{(1)}$ has larger significant lags than $Y^{(3)}$ and there is significant cross correlation between $Y^{(1)}$ and $Y^{(3)}$.

Figure 5.1b displays joint confidence regions for $Y^{(1)}$ and $Y^{(3)}$. Recall that the true mean is $(0, 0)$, and is present in all three regions, but the ellipse produced by mBM has significantly smaller volume than the uBM boxes. The orientation of the

ellipse is determined by the cross correlations shown in Figure 5.1a.

We assess the multivariate ESS and the relative standard deviation fixed-volume sequential stopping rule by comparing these methods to their corresponding univariate methods - the univariate ESS of Gong and Flegal (2016) and the relative standard deviation fixed-width sequential stopping rule of Flegal and Gong (2015). We set $n^* = 1000$, $b_n = \lfloor n^{1/3} \rfloor$, $\epsilon$ in $\{0.05, 0.02, 0.01\}$ and at termination of each method, calculate the coverage probabilities and effective sample size. Results are presented in Table 5.1. Note that as $\epsilon$ decreases, termination time increases and coverage probabilities tend to the 90% nominal for each method. Also note that the uncorrected methods produce confidence regions with undesirable coverage probabilities and thus are not of interest. Consider $\epsilon = .02$ in Table 5.1. Termination for mBM is at $8.8 \times 10^4$ iterations compared to $9.6 \times 10^5$ for uBM-Bonferroni. However, the estimates for multivariate ESS at $8.8 \times 10^4$ iterations is $4.7 \times 10^4$ samples compared to univariate ESS of $5.6 \times 10^4$ samples for $9.6 \times 10^5$ iterations. This is because the leading component $Y^{(1)}$ mixes much slower than the other components and defines the behavior of the univariate ESS.

A small study presented in Table 5.2 elaborates on this behavior. Over 100 replications of Monte Carlo sample sizes $10^5$ and $10^6$, we present the mean estimate of ESS using multivariate and univariate methods. The estimate of ESS for the first component is significantly smaller than all other components leading to a conservative univariate estimate of ESS.

Table 5.3 shows the coverage probabilities and volume to the $p$th root of 90% confidence regions averaged over 1000 replications. Clearly the univariate uncorrected method has less than desired coverage probability, and the mBM produces 90% confidence regions with volume much smaller than the uBM method corrected for Bonferroni.

The effect of such a large difference in volume can be seen in the ESS calculations

| | mBM | uBM-Bonferroni | uBM |
|---|---|---|---|
| Termination Iteration | | | |
| $\epsilon = 0.05$ | 14423 (10) | 141427 (93) | 66083 (52) |
| $\epsilon = 0.02$ | 88259 (23) | 956454 (286) | 478869 (617) |
| $\epsilon = 0.01$ | 360284 (454) | 3991753 (0) | 2043931 (902) |
| Effective Sample Size | | | |
| $\epsilon = 0.05$ | 7650 (6) | 9093 (6) | 4542 (3) |
| $\epsilon = 0.02$ | 46722 (21) | 55730 (25) | 28749 (29) |
| $\epsilon = 0.01$ | 192611 (217) | 223461 (56) | 116297 (51) |
| Volume to the $p$th root | | | |
| $\epsilon = 0.05$ | 0.0361 (1.52e-05) | 0.0234 (9.7e-06) | 0.0241 (1.24e-05) |
| $\epsilon = 0.02$ | 0.0146 (3.50e-06) | 0.0091 (2.1e-06) | 0.0091 (5.30e-06) |
| $\epsilon = 0.01$ | 0.0072 (4.20e-06) | 0.0045 (5.0e-07) | 0.0044 (1.00e-06) |
| Coverage Probabilities | | | |
| $\epsilon = 0.05$ | 0.886 (0.0101) | 0.945 (0.0072) | 0.757 (0.0136) |
| $\epsilon = 0.02$ | 0.883 (0.0102) | 0.942 (0.0074) | 0.765 (0.0134) |
| $\epsilon = 0.01$ | 0.900 (0.0095) | 0.941 (0.0075) | 0.778 (0.0131) |

Table 5.1: VAR: Over 1000 replications, we present termination iterations, effective sample size at termination and coverage probabilities at termination for each corresponding method. Standard errors are in parentheses.

| $n$ | ESS | $\text{ESS}_1$ | $\text{ESS}_2$ | $\text{ESS}_3$ | $\text{ESS}_4$ | $\text{ESS}_5$ |
|---|---|---|---|---|---|---|
| $10^5$ | $52902_{(71)}$ | $5447_{(41)}$ | $33863_{(284)}$ | $82986_{(664)}$ | $82727_{(639)}$ | $81923_{(647)}$ |
| $10^6$ | $538313_{(332)}$ | $53472_{(256)}$ | $334098_{(1665)}$ | $819760_{(3801)}$ | $820761_{(3639)}$ | $822912_{(3845)}$ |

Table 5.2: VAR: Effective sample sample (ESS) estimated using proposed multivariate method and the univariate method of Gong and Flegal (2016) for Monte Carlo sample sizes of $n = 10^5$ and $n = 10^6$ and 100 replications. Standard errors are in parentheses.

| Volume to the $p$th root | | | |
|---|---|---|---|
| $n$ | mBM | uBM-Bonferroni | uBM |
| $10^3$ | 0.149 (1.6e-04) | 0.254 (4.7e-04) | 0.179 (3.3e-04) |
| $10^4$ | 0.048 (2.3e-05) | 0.085 (7.4e-05) | 0.060 (5.2e-05) |
| $10^5$ | 0.015 (3.0e-06) | 0.028 (1.1e-05) | 0.020 (8.0e-06) |
| Coverage Probabilities | | | |
| $10^3$ | 0.815 (0.0123) | 0.836 (0.0117) | 0.627 (0.0153) |
| $10^4$ | 0.893 (0.0098) | 0.908 (0.0091) | 0.703 (0.0144) |
| $10^5$ | 0.892 (0.0098) | 0.928 (0.0082) | 0.753 (0.0136) |

Table 5.3: VAR: Volume to the $p$th ($p = 5$) root and coverage probabilities for 90% confidence regions constructed using mBM, uBM uncorrected and uBM corrected for Bonferroni. Replications = 1000 and $b_n = \lfloor n^{1/3} \rfloor$. Standard errors are in parentheses.

in Table 5.2. Over 100 replications, a Monte Carlo sample size of $10^5$ was obtained and its ESS calculated. The conservative univariate methods that ignore the cross correlations would estimate the effective sample size to be 5432, due to one slow mixing component in the process. Our estimate of the effective sample size is significantly larger at 55190.

## 5.2 Bayesian Logistic Regression

We revisit the Bayesian logistic regression example introduced in Chapter 1. Recall that for $i = 1, \ldots, K$, let $Y_i$ is a binary response variable and $X_i = (x_{i1}, x_{i2}, \ldots, x_{i5})$ is the observed predictors for the $i$th observation. Assume $\tau^2 > 0$ is known,

$$Y_i \mid X_i, \beta \overset{ind}{\sim} \text{Bernoulli}\left(\frac{1}{1+e^{-X_i\beta}}\right), \quad \text{and} \quad \beta \sim N_5(0, \tau^2 I_5)\,. \tag{5.1}$$

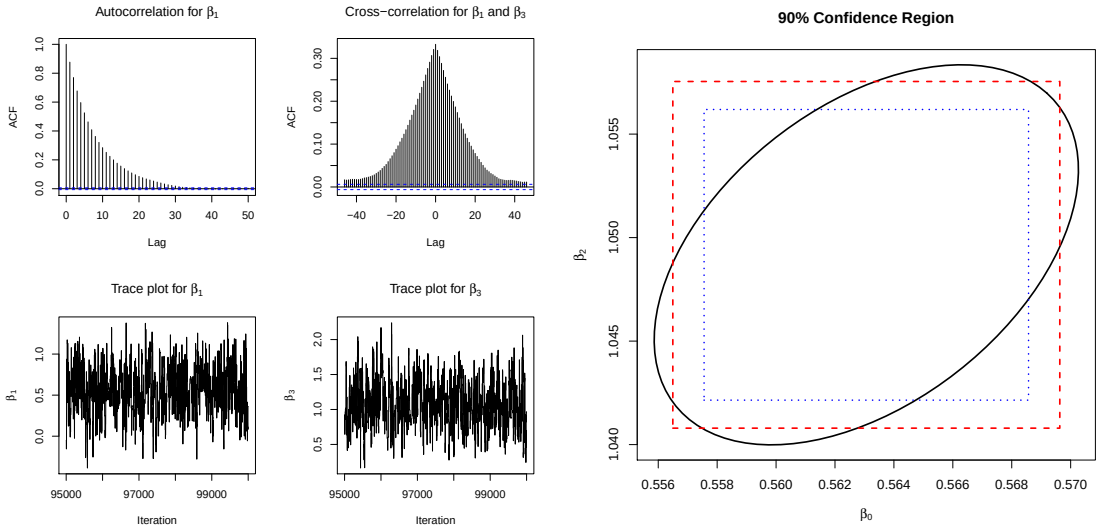
This simple hierarchical model results in an intractable posterior, $F$ on $\mathbb{R}^5$. The dataset used is the `logit` dataset in the `mcmc` R package. The goal is to estimate the

posterior mean of $\beta$, $\mathrm{E}_F\beta$. Thus $g$ here is the identity function mapping to $\mathbb{R}^5$. We implement a random walk Metropolis-Hastings algorithm with a multivariate normal proposal distribution $N_5(\,\cdot\,, 0.35^2 I_5)$ where $I_5$ is the $5 \times 5$ identity matrix and the 0.35 scaling ensures an optimal acceptance probability as suggested by Roberts et al. (1997).

**Theorem 5.1** The random walk based Metropolis-Hastings algorithm with invariant distribution given by the posterior from (5.1) is geometrically ergodic. □

We calculate the Monte Carlo estimate for $\mathrm{E}_F\beta$ from an MCMC sample of size $10^5$. The starting value for $\beta$ is a random draw from the prior distribution. The covariance matrix $\Sigma$ is estimated by the mBM estimator described in Section 4.2. We also implement the univariate batch means (uBM) methods described in Jones et al. (2006) to estimate $\sigma_i^2$, which captures the autocorrelation in each component while ignoring the cross-correlation. This cross-correlation is often significant as seen in Figure 5.2a, and can only be captured by multivariate methods like mBM. Figure 5.2b shows 90% confidence regions created using mBM and uBM estimators for $\beta_1$ and $\beta_3$ (for the purpose of this figure, we set $p = 2$).

To assess the confidence regions, we verify their coverage probabilities over 1000 independent replications with Monte Carlo sample sizes in $\{10^4, 10^5, 10^6\}$. The true posterior mean, $(0.5706, 0.7516, 1.0559, 0.4517, 0.6545)$, was obtained by averaging over $10^9$ iterations. For each of the 1000 replications, it was noted whether the confidence region contained the true posterior mean. The volume of the confidence region to the $p$th root was also observed. Table 5.4 summarizes the results. Note that though the uncorrected univariate methods produce the smallest confidence regions, their coverage probabilities are far from desirable. For a large enough Monte Carlo sample size, mBM produces 90% coverage probabilities with systematically lower volume than uBM corrected with Bonferroni (uBM-Bonferroni).




(a) (b)

Figure 5.2: (a) ACF plot for $\beta_1$, cross-correlation plot between $\beta_1$ and $\beta_3$, and trace plots for $\beta_1$ and $\beta_3$. (b) Joint 90% confidence region for $\beta_1$ and $\beta_3$. The ellipse is made using mBM, the dotted line using uncorrected uBM, and the dashed line using the uBM corrected by Bonferroni. Monte Carlo sample size is $10^5$ for both plots.

| $n$ | mBM | uBM-Bonferroni | uBM |
|---|---|---|---|
| Volume to the $p$th root | | | |
| $10^4$ | 0.062 (7.94e-05) | 0.066 (9.23e-05) | 0.046 (6.48e-05) |
| $10^5$ | 0.020 (1.20e-05) | 0.021 (1.42e-05) | 0.015 (1.00e-05) |
| $10^6$ | 0.006 (1.70e-06) | 0.007 (2.30e-06) | 0.005 (1.60e-06) |
| Coverage Probabilities | | | |
| $10^4$ | 0.876 (0.0104) | 0.889 (0.0099) | 0.596 (0.0155) |
| $10^5$ | 0.880 (0.0103) | 0.910 (0.0090) | 0.578 (0.0156) |
| $10^6$ | 0.894 (0.0097) | 0.913 (0.0094) | 0.627 (0.0153) |

Table 5.4: Logistic: Volume to the $p$th ($p = 5$) root and coverage probabilities for 90% confidence regions constructed using mBM, uBM uncorrected, and uBM corrected by Bonferroni. Replications = 1000 and standard errors are indicated in parenthesis.

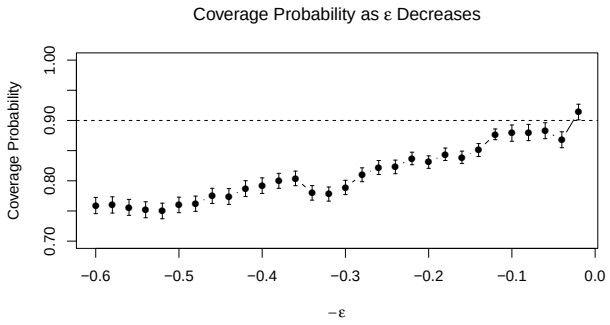



Figure 5.3: Logistic: Demonstration of asymptotic validity for the Bayesian logistic regression model using the relative standard deviation fixed-volume sequential stopping rule. Replications = 100 and standard error bars are indicated.

Figure 5.3 demonstrates the asymptotic validity result of Chapter 3. As $\epsilon$ decreases (or $-\epsilon$ increases), the coverage probability reaches the nominal level of .90. Notice that this behavior is not monotonic due to the random nature of the process. Thus, for smaller values of $\epsilon$ we expect better coverage probabilities.

## 5.3 Bayesian Lasso

Let $y$ be a $K \times 1$ response vector and $X$ be a $K \times r$ matrix of predictors. We consider the following Bayesian lasso formulation of Park and Casella (2008).

$$\begin{aligned} y|\beta, \sigma^2, \tau^2 &\sim N_K(X\beta, \sigma^2 I_n) \\ \beta|\sigma^2, \tau^2 &\sim N_r(0, \sigma^2 D_\tau) \quad \text{where} \quad D_\tau = \text{diag}(\tau_1^2, \tau_2^2, \ldots, \tau_r^2) \\ \sigma^2 &\sim \text{Inverse-Gamma}(\alpha, \xi) \\ \tau_j^2 &\overset{iid}{\sim} \text{Exponential}\left(\frac{\lambda^2}{2}\right) \quad \text{for } j = 1, \ldots, r, \end{aligned}$$

where $\lambda, \alpha$, and $\xi$ are fixed and the Inverse-Gamma$(a, b)$ distribution has density proportional to $x^{-a-1}e^{-b/x}$. We use a deterministic scan Gibbs sampler to draw

approximate samples from the posterior; see Khare and Hobert (2013) for a full description of the algorithm. Khare and Hobert (2013) showed that for $K \geq 3$, this Gibbs sampler is geometrically ergodic for arbitrary $r, X$ and $\lambda$.

We fit this model to the cookie dough dataset of Osborne et al. (1984). The data was collected to test the feasibility of near infra-red (NIR) spectroscopy for measuring the composition of biscuit dough pieces. There are 72 observations; the response variable is the amount of dry flour content measured and the predictor variables are 25 measurements of spectral data spaced equally between 1100 to 2498 nanometers. We are interested in estimating the posterior mean for $(\beta, \tau^2, \sigma^2)$, $p = 51$. The data is available in the `R` package `ppls`, and the Gibbs sampler is implemented in the function `blasso` in `R` package `monomvn`. The "truth" was declared by averaging posterior means from 1000 parallel chains of length $10^6$. We set $n^* = 2 \times 10^4$ and $b_n = \lfloor n^{1/3} \rfloor$.

Table 5.5 shows termination results from 1000 replications. With $p = 51$, the uncorrected univariate regions produce confidence regions with very low coverage probabilities. The uBM-Bonferroni and mBM provide competitive coverage probabilities at termination. However, termination for mBM is significantly earlier than univariate methods over all values of $\epsilon$. For $\epsilon = .05$ and .02 we observe zero standard error for termination using mBM since termination is achieved at the same 10% increment over all 1000 replications. Thus the variability in those estimates is less than 10% of the size of the estimate.

| | mBM | uBM-Bonferroni | uBM |
|---|---|---|---|
| Termination Iteration | | | |
| $\epsilon = 0.05$ | 20000 (0) | 69264 (76) | 20026 (7) |
| $\epsilon = 0.02$ | 69045 (0) | 445754 (664) | 122932 (103) |
| $\epsilon = 0.01$ | 271088 (393) | 1765008 (431) | 508445 (332) |
| Effective Sample Size | | | |
| $\epsilon = 0.05$ | 15631 (4) | 16143 (15) | 4778 (6) |
| $\epsilon = 0.02$ | 52739 (8) | 101205 (122) | 28358 (24) |
| $\epsilon = 0.01$ | 204801 (283) | 395480 (163) | 115108 (74) |
| Volume to the $p$th root | | | |
| $\epsilon = 0.05$ | 0.0444 (6.0e-06) | 0.0376 (2.0e-05) | 0.0370 (8.0e-06) |
| $\epsilon = 0.02$ | 0.0236 (2.0e-06) | 0.0149 (1.1e-05) | 0.0150 (5.9e-06) |
| $\epsilon = 0.01$ | 0.0119 (8.10e-06) | 0.0075 (9.0e-07) | 0.0074 (2.5e-06) |
| Coverage Probabilities | | | |
| $\epsilon = 0.05$ | 0.898 (0.0096) | 0.896 (0.0097) | 0.010 (0.0031) |
| $\epsilon = 0.02$ | 0.892 (0.0098) | 0.905 (0.0093) | 0.009 (0.0030) |
| $\epsilon = 0.01$ | 0.898 (0.0096) | 0.929 (0.0081) | 0.009 (0.0030) |

Table 5.5: Bayesian Lasso: Over 1000 replications, we present termination iterations, effective sample size at termination and coverage probabilities at termination for each corresponding method. Standard errors are in parentheses.

## 5.4 Bayesian Dynamic Spatio-Temporal Model

Gelfand et al. (2005) propose a Bayesian hierarchical model for modeling univariate and multivariate dynamic spatial data viewing time as discrete and space as continuous. The methods in their paper have been implemented in the R package `spBayes`. We present a simpler version of the dynamic model as described by Finley et al. (2015).

Let $s = 1, 2, \ldots, N_s$ be location sites and $t = 1, 2, \ldots, N_t$ be time-points. Let the observed measurement at location $s$ and time $t$ be denoted by $y_t(s)$. In addition, let $x_t(s)$ be the $r \times 1$ vector of predictors, observed at location $s$ and time $t$, and $\beta_t$ be the $r \times 1$ vector of coefficients. For $t = 1, 2, \ldots, N_t$,

$$y_t(s) = x_t(s)^T \beta_t + u_t(s) + \epsilon_t(s), \quad \epsilon_t(s) \overset{ind}{\sim} N(0, \tau_t^2); \tag{5.2}$$

$$\beta_t = \beta_{t-1} + \eta_t, \quad \eta_t \overset{iid}{\sim} N(0, \Sigma_\eta);$$

$$u_t(s) = u_{t-1}(s) + w_t(s), \quad w_t(s) \overset{ind}{\sim} GP(0, \sigma_t^2 \rho(\cdot; \phi_t)), \tag{5.3}$$

where $GP(0, \sigma_t^2 \rho(\cdot; \phi_t))$ denotes a spatial Gaussian process with covariance function $\sigma_t^2 \rho(\cdot; \phi_t)$. Here, $\sigma_t^2$ denotes the spatial variance component and $\rho(\cdot, \phi_t)$ is the correlation function with exponential decay. Equation (5.2) is referred to as the measurement equation and $\epsilon_t(s)$ denotes the measurement error, assumed to be independent of location and time. Equation (5.3) contains the transition equations which emulate the Markovian nature of dependence in time. To complete the Bayesian hierarchy, the following priors are assumed

$$\begin{aligned} \beta_0 &\sim N(m_0, C_0) \quad \text{and} \quad u_0(s) \equiv 0; \\ \tau_t^2 &\sim \text{IG}(a_\tau, b_\tau) \quad \text{and} \quad \sigma_t^2 \sim \text{IG}(a_s, b_s); \\ \Sigma_\eta &\sim \text{IW}(a_\eta, B_\eta) \quad \text{and} \quad \phi_t \sim \text{Unif}(a_\phi, b_\phi)\,, \end{aligned}$$

where IW denotes the Inverse-Wishart distribution with density proportional to $|\Sigma_\eta|^{-\frac{a_\eta+q+1}{2}} e^{-\frac{1}{2}tr(B_\eta \Sigma_\eta^{-1})}$ and IG$(a, b)$ is the inverse-Gamma distribution with density proportional to $x^{-a-1}e^{-b/x}$. We fit the model to the `NETemp` dataset in the `spBayes` package. This dataset contains monthly temperature measurements from 356 weather stations on the east coast of USA collected from January 2000 to December 2010. The elevation of the weather stations is also available as a covariate. We choose a subset of the data with 10 weather stations for the year 2000, and fit the model with an intercept. The resulting posterior has $p = 185$ components.

A conditional Metropolis-Hastings sampler is described in Gelfand et al. (2005) and implemented in the `spDynLM` function. Default hyper parameter settings were used. The posterior and the rate of convergence for this sampler have not been studied; thus we do not know if the conditions of our theoretical results are satisfied. Our goal is to estimate the posterior expectation of $\theta = (\beta_t, u_t(s), \sigma_t^2, \Sigma_\eta, \tau_t^2, \phi_t)$. For the calculation of coverage probabilities, 1000 parallel runs of a $2 \times 10^6$ MCMC sample were averaged and declared as the "truth". We set $b_n = \lfloor n^{1/2} \rfloor$ and $n^* = 5 \times 10^4$ so that $a_n > p$ to ensure positive definitiveness of $\Sigma_n$.

Due to the Markovian transition equations in (5.3), the $\beta_t$ and $u_t$ exhibit a significant covariance structure in the posterior distribution. This is evidenced in Figure 5.4 where for Monte Carlo sample size $n = 10^5$, we present confidence regions for $\beta_1^{(0)}$ and $\beta_2^{(0)}$, the intercept coefficient for the first and second months, and for $u_1(1)$ and $u_2(1)$, the additive spatial coefficient for the first and second weather stations. The thin ellipses indicate that the principal direction of variation is due to the correlation between the components. This significant reduction in volume, along with the conservative Bonferroni correction ($p = 185$) results in increased delay in termination when using univariate methods. For smaller values of $\epsilon$ it was not possible to store the MCMC output in memory on a 8 gigabyte machine using uBM-Bonferroni methods.

As a result (see Table 5.6), the univariate methods could not be implemented for

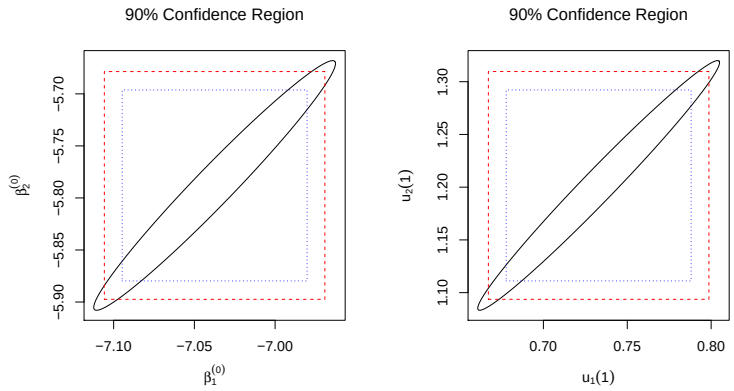



Figure 5.4: Bayesian Spatial: 90% confidence regions for $\beta_1^{(0)}$ and $\beta_2^{(0)}$ and $u_1(1)$ and $u_2(1)$. Monte Carlo sample size $= 10^5$.

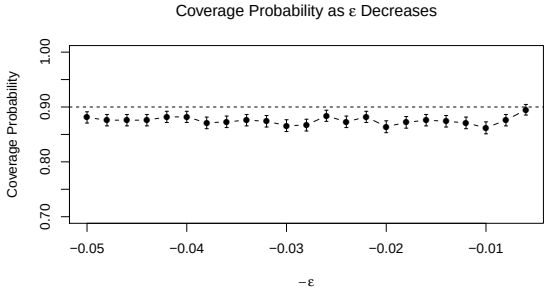



Figure 5.5: Bayesian Spatial: Plot of $-\epsilon$ versus observed coverage probability for mBM estimator over 1000 replications with $b_n = \lfloor n^{1/2} \rfloor$.

smaller $\epsilon$ values. For $\epsilon = .10$, termination for mBM was at $n^* = 5 \times 10^4$ for every replication. At these minimum iterations, the coverage probability for mBM is at 88%, whereas both the univariate methods have far lower coverage probabilities at 0.62 for uBM-Bonferroni and 0.003 for uBM. The coverage probabilities for the uncorrected methods are quite small since we are making 185 confidence regions simultaneously.

In Figure 5.5, we illustrate asymptotic validity of the confidence regions constructed using the relative standard deviation fixed-volume sequential stopping rule. We present the observed coverage probabilities over 1000 replications for several values of $\epsilon$.

| | mBM | uBM-Bonferroni | uBM |
|---|---|---|---|
| Termination Iteration | | | |
| $\epsilon = 0.10$ | 50000 (0) | 1200849 (28315) | 311856 (491) |
| $\epsilon = 0.05$ | 50030 (12) | - | 1716689 (2178) |
| $\epsilon = 0.02$ | 132748 (174) | - | - |
| $\epsilon = 0.01$ | 407380 (122) | - | - |
| Effective Sample Size | | | |
| $\epsilon = 0.10$ | 55170 (20) | 3184 (75) | 1130 (1) |
| $\epsilon = 0.05$ | 55190 (20) | - | 4525 (4) |
| $\epsilon = 0.02$ | 105166 (97) | - | - |
| $\epsilon = 0.01$ | 275073 (78) | - | - |
| Volume to the $p$th root | | | |
| $\epsilon = 0.010$ | 0.0308 (6.40e-06) | 0.0409 (9.63e-04) | 0.0657 (4.2e-05) |
| $\epsilon = 0.05$ | 0.0308 (1.11e-05) | - | 0.0315 (1.7e-05) |
| $\epsilon = 0.02$ | 0.0123 (9.80e-06) | - | - |
| $\epsilon = 0.01$ | 0.0062 (1.00e-06) | - | - |
| Coverage Probabilities | | | |
| $\epsilon = 0.10$ | 0.882 (0.0102) | 0.625 (0.0153) | 0.007 (0.0026) |
| $\epsilon = 0.05$ | 0.881 (0.0102) | - | 0.016 (0.0040) |
| $\epsilon = 0.02$ | 0.864 (0.0108) | - | - |
| $\epsilon = 0.01$ | 0.862 (0.0109) | - | - |

Table 5.6: Bayesian Spatial: Over 1000 replications, we present termination iteration, effective sample size at termination and coverage probabilities at termination for each corresponding method at 90% nominal levels. Standard errors are in parentheses.

# Chapter 6

# Convergence Rates of Markov Chains

The rate of convergence of a Markov chain impacts the Monte Carlo error in estimation. It is thus important to ensure that MCMC samplers used for statistical inference are at least polynomially ergodic in order for our theoretical results to hold. In this chapter we study the rates of convergence of some MCMC samplers for a variety of Bayesian models commonly used in statistics.

Recall that we showed that the random walk Metropolis-Hastings algorithm for the Bayesian logistic regression problem is geometrically ergodic. All of the samplers we study here are Gibbs samplers. Establishing rates of convergences for Gibbs samplers is comparatively easy due to the structure of the Markov chain transition density. For this reason, there has been a considerable amount of work in establishing geometric ergodicity of Gibbs samplers, many of which are two variable Gibbs samplers. Two variable Gibbs samplers are special because the marginal process for each variable is a Markov chain with the same rate of convergence as the joint chain (Roberts and Rosenthal (2001)). Thus it is often sufficient to study the marginal chains in order to study properties of the joint chain. Higher variable Gibbs samplers do not have this property and thus studying their convergence rates is often more challenging. Geometric ergodicity of the three variable Gibbs samplers in the

Bayesian lasso and the Bayesian elastic net was shown by Khare and Hobert (2013) and Roy and Chakraborty (2016); Khare and Hobert (2012) proved geometric ergodicity of the three variable Gibbs sampler in Bayesian quantile regression; and Doss and Hobert (2010) and Jones and Hobert (2004) proved geometric ergodicity of the three variable Gibbs sampler in hierarchical random effects models. Recently, Johnson and Jones (2015) established geometric ergodicity of a four variable random scan Gibbs sampler for a hierarchical random effects model.

## 6.1 Linchpin Variable Samplers

Let $f(x, y)$ be a probability density function on $\mathcal{X} \times \mathcal{Y} \subseteq \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ and $F(x, y)$ be the associated distribution. One prominent roadblock in implementing MCMC for modern problems is dimensionality of the state space, $d_1 + d_2$. A larger dimensional state space usually implies slower convergence of the Markov chain to the invariant distribution.

Let $f_{X|Y}$ be the probability density function of the conditional distribution of $X$ given $Y$; also let $f_Y$ be the probability density function of the marginal distribution of $Y$ and let $F_{X|Y}$ and $F_Y$ be the respective associated distributions. If exact sampling from $F_{X|Y}$ is straightforward, then $Y$ is called a *linchpin variable* since

$$f(x, y) = f_{X|Y}(x|y)\, f_Y(y).$$

Exact samples drawn from $F_Y$ can then be used to obtain a realization from the joint distribution by using $F_{X|Y}$. If exact sampling from $F$ is impossible or inefficient, we explore replacing exact samples with MCMC samples from $F_Y$ to obtain an MCMC realization from $F$. Acosta et al. (2015) show that the joint chain $\{(X_t, Y_t)\}$ has the same rate of convergence as $\{Y_t\}$. This result is unsurprising but important as it explains the benefits of using linchpin variable samplers. Additionally, the marginal

chain explores a $d_2$ dimensional space as opposed to a generic MCMC sampler exploring a $d_1 + d_2$ dimensional space.

To implement a linchpin variable sampler, the only requirement is exact sampling from $F_{X|Y}$. Thus whenever a Gibbs sampler can be implemented, a linchpin variable sampler can also be used.

### 6.1.1 Bayesian Variable Selection

The purpose of this example is to demonstrate the usability and immediate impact of a linchpin variable sampler when one component of the Markov chain takes values in a finite state space.

Let $y$ be a response vector in $\mathbb{R}^n$, $X$ be an $n \times p_n$ matrix of predictors and $\beta \in \mathbb{R}^{p_n}$ be the vector of coefficients. Narisetty and He (2014) introduced the following BAyesian Shrinkage And Diffusing priors (BASAD) model:

$$\begin{aligned} y|X, \beta, \sigma^2 &\sim N(X\beta, \sigma^2 I_n) \\ \beta_i|\sigma^2, Z_i = 0 &\sim N(0, \sigma^2 \tau_{0,n}^2) \\ \beta_i|\sigma^2, Z_i = 1 &\sim N(0, \sigma^2 \tau_{1,n}^2) \\ \Pr(Z_i = 1) = 1 - \Pr(Z_i = 0) &= q_n \\ \sigma^2 &\sim IG(\alpha_1, \alpha_2)\,, \end{aligned} \tag{6.1}$$

where $\tau_{0,n}^2$ and $\tau_{1,n}^2$ are positive functions of $n$ and $\alpha_1$ and $\alpha_2$ are hyper-parameters for the Inverse Gamma distribution. When $\tau_{0,n}^2$ and $\tau_{1,n}^2$ do not depend on $n$, the BASAD model corresponds to the seminal variable selection model of George and McCulloch (1993). The latent variables $Z_i$ are indicators of whether the $i$th variable is active or not. The introduction of $Z_i$ allows for variable selection, and the structure of $\tau_{\cdot,n}^2$ allows for shrinkage and diffusion of the priors. If the $i$th variable is active ($Z_i = 1$),

then $\tau^2_{1,n}$ will be large and if it is inactive ($Z_i = 0$), then $\tau^2_{0,n}$ will be small.

Narisetty and He (2014) proposed a Gibbs sampler to sample from the resulting posterior using the following full conditional distributions. Let $D_z$ be the $p_n \times p_n$ diagonal matrix with $\tau^2_{z_i,n}$ on the diagonal and $V_z = (X^T X + D_z^{-1})$. Let $\eta(x, 0, \tau^2)$ be the probability density function of a mean zero, variance $\tau^2$ normal random variable.

$$\begin{aligned}
&\beta \mid Z, \sigma^2, y \sim N(V_z^{-1} X^T y, \sigma^2 V_z^{-1}) \\
&\Pr(Z_i = 1 \mid \beta, \sigma^2, y) = \frac{q_n \eta(\beta_i, 0, \sigma^2 \tau^2_{1,n})}{q_n \eta(\beta_i, 0, \sigma^2 \tau^2_{1,n}) + (1 - q_n)\eta(\beta_i, 0, \sigma^2 \tau^2_{0,n})} \\
&\sigma^2 \mid \beta, Z, y \sim IG\left(\alpha_1 + \frac{n}{2} + \frac{p_n}{2}, \alpha_2 + \frac{\beta^T D_z^{-1} \beta + (y - X\beta)^T (y - X\beta)}{2}\right). \quad (6.2)
\end{aligned}$$

Note that each of the $Z_i$ is updated independent of the other $Z_i$ and thus all $Z_i$ can be updated in a block. The resulting MCMC sampler is a three variable deterministic scan Gibbs sampler that is updated according to $(\beta, Z, \sigma^2) \to (\beta', Z', \sigma'^2)$. The rate of convergence of the Gibbs sampler is unknown.

We will develop a linchpin variable sampler to sample from the joint posterior distribution. Notice that the joint posterior distribution takes the following decomposition.

$$f(\beta, \sigma^2, Z \mid y) = f(\beta, \sigma^2 \mid Z, y)\, f(Z \mid y).$$

If we can sample from the conditional distribution of $(\beta, \sigma^2)$ given $Z$, then $Z$ is a linchpin variable. We find that

$$\sigma^2 \mid Z, y \sim IG\left(\alpha_1 + \frac{n}{2}, \frac{2\alpha_2 + y^T (I_n - X(X^T X + D_z^{-1}) X^T) y}{2}\right).$$

Using $f(\beta, \sigma^2 \mid Z, y) = f(\beta \mid \sigma^2, Z, y)\, f(\sigma^2 \mid Z, y)$, and the fact that $\beta \mid Z, \sigma^2, y \sim N(V_z^{-1} X^T y, \sigma^2 V_z^{-1})$, exact samples can be drawn from $(\beta, \sigma^2) \mid Z, y$. Thus, $Z$ is a

linchpin variable. A Metropolis-Hastings sampler will be used to sample from $F_Z$. This Markov chain will explore a $p$-dimensional space whereas the Gibbs sampler explores a $2p + 1$ dimensional space.

Consider the independence proposal distribution of independent Bernoullis. That is, the proposal distribution density function is $g(Z) = \prod_{i=1}^{p_n} q_n^{Z_i}(1 - q_n)^{1-Z_i}$. In Appendix D.1, we show that

$$\frac{f(Z \mid y)}{g(Z)} \leq k\alpha_2^{-(n/2+\alpha_1)},$$

where $k$ is the unknown normalizing constant for $f(Z \mid y)$. Thus, a rejection sampler can be implemented, however since $p$ is often large, the probability of acceptance is low enough so as to make it impossible to implement this in practice.

Nonetheless, by Acosta et al. (2015), the linchpin variable sampler that uses $g(z)$ as the proposal distribution in the independence Metropolis-Hastings step is uniformly ergodic. In fact, since $Z \in \{0, 1\}^p$, most reasonable proposal distributions for the Metropolis-Hastings step in the linchpin variable sampler will lead to uniformly ergodic Markov chains (see Section 3.4 in Roberts and Rosenthal (2004)). In this way, the use of a linchpin variable makes it easier to assess the rate of convergence of the Markov chain. This is not to say that the linchpin variable sampler used in this case will converge faster than the Gibbs sampler, but at least a rate of convergence will be known.

### 6.1.2 Latent Dirichlet Allocation

Analyzing a collection of documents and identifying underlying "topics" associated with each document is called topic modeling. A popular tool in topic modeling is the Latent Dirichlet allocation (LDA) of Blei et al. (2003). LDA is a Bayesian hierarchical model that allows the interpretation of a document as a mixture over latent topics,

where the topics themselves are mixture over words. LDA treats a document as a set of words, or the more common expression "bag of words"; that is, the ordering of words is not modeled in this analytic framework. We present the model formally.

Suppose there are $D$ documents, each of which has $N_i$ words, $i = 1, \ldots, D$, and let $N = \sum_{i=1}^{D} N_i$. Let there be $W$ unique words across all documents. The $j$th word from the $i$th document is denoted by $w_{ij}$, $j = 1, \ldots, N_i$. We assume the existence of $K$ latent topics. Each word $w_{ij}$ is assigned a topic $z_{ij} = k$, where $k = 1, \ldots, K$.

For $k = 1, \ldots, K$, let $\phi_k$ be a $W$-dimensional vector of probabilities such that $\sum_{t=1}^{W} \phi_{kt} = 1$. The data (words) are assumed to arise as following.

$$\textbf{Likelihood}: \qquad w_{ij} \mid z_{ij} = k, \phi_k \overset{ind}{\sim} \text{Multinomial}(\phi_k), \tag{6.3}$$

The vector $\phi_k$ represents the meaning of topic $k$ as defined by its mixture over the $W$ words. Thus, given a topic assignment of $k$, for word $j$ in document $i$, the word is assumed to be drawn from the dictionary with probability $\phi_k$. Below are the prior specifications.

$$\begin{aligned} \textbf{Priors}: \qquad & z_{ij} \mid \theta_i \overset{ind}{\sim} \text{Multinomial}(\theta_i) \\ & \theta_i \sim \text{Dirichlet}(a), \quad \text{independently for } i = 1, \ldots, D \\ & \phi_k \sim \text{Dirichlet}(b), \quad \text{independently for } k = 1, \ldots, K. \end{aligned} \tag{6.4}$$

In (6.4), $\theta_i$ is a $K$-dimensional vector of probabilities such that for each document $i$, $\sum_{t=1}^{K} \theta_{it} = 1$. Thus, $\theta_i$ stores the mixture of topic assignments for the $i$th document. Both $\theta_i$ and $\phi_k$ are given Dirichlet priors on their respective $K$-dimensional and $W$-dimensional spaces. The hyperparameters $a$ and $b$ are positive scalars of symmetric Dirichlet priors.

The purpose of LDA is to model each document as a mixture over $K$ topics. For

this reason, the vectors $\theta_i$ are of prominent interest. In addition, $\phi_k$ is used to infer the interpretation of each of the topics. As in any Bayesian model, inference in made using the posterior distribution of these parameters, $F$. Let $\theta$ denote all vectors $\theta_i$, $\phi$ denote all vectors $\phi_k$ and $z$ denote all $z_{ij}$.

$$\textbf{Posterior}: \qquad f(\theta, \phi, z \mid w, a, b) = \frac{f(w \mid z, \phi) f(z \mid \theta) f(\theta \mid a) f(\phi \mid b)}{f(w)}. \tag{6.5}$$

This distribution is not available in closed form, since $f(w)$ is intractable and thus MCMC methods are used for inference. It is also important to mention that the quantities needed to be estimated are the posterior means of $\theta$ and $\phi$ (Blei and Lafferty, 2009). That is, we are interested in estimating $\mathrm{E}_F[\theta]$ and $\mathrm{E}_F[\phi]$.

Define

$$\begin{aligned}
n_{i\cdot k} &= \sum_{j=1}^{W} n_{ijk} = \text{the number of words assigned topic } k \text{ in document } i, \\
n_{\cdot jk} &= \sum_{i=1}^{D} n_{ijk} \\
&= \text{the number of times word } j \text{ is assigned topic } k \text{ over all documents, and} \\
n_{\cdot\cdot k} &= \sum_{i=1}^{D} \sum_{j=1}^{N_i} n_{ijk} \\
&= \text{the number of times topic } k \text{ assigned to any word over all documents.}
\end{aligned}$$

Notice that

$$f(\theta, \phi, z \mid w, a, b) = f(\theta, \phi \mid z, w, a, b)\, f(z \mid w, a, b)$$

Griffiths and Steyvers (2004) noted that $f(\theta, \phi \mid z, w, a, b)$ is available in closed form and both $\theta$ and $\phi$ can be integrated out of the posterior to obtain the following

marginal posterior of $z$.

$$f(z \mid w, a, b) \propto \prod_{k=1}^{K} \frac{\prod_{t=1}^{W} \Gamma(n_{\cdot tk} + b)}{\Gamma(n_{\cdot\cdot k} + Wb)} \prod_{i=1}^{D} \frac{\prod_{k=1}^{K} \Gamma(n_{i\cdot k} + a)}{\Gamma(n_{i\cdot\cdot} + Ka)} .$$

Samples are drawn from the marginal posterior distribution of $z$ using a Gibbs sampler, leading to a collapsed Gibbs sampler for the full posterior distribution. Thus, in this case, $z$ is a linchpin variable and the Markov chain that samples from the marginal posterior of $z$ is described by a Gibbs sampler.

In addition, the full conditional distribution of each $z_{ij}$ is available in closed form, allowing for a Gibbs sampler with invariant distribution being the marginal posterior of $z$. Specifically the full conditional is,

$$\Pr(z_{ij} = k \mid z_{-ij}, w, a, b) \propto \frac{\left(n_{\cdot jk}^{-ij} + b\right)\left(n_{i\cdot k}^{-ij} + a\right)}{n_{\cdot\cdot k} + Wb} .$$

In this way the collapsed Gibbs sampler is a linchpin variable sampler. Since the state space for the marginal chain of $z$ is finite and the full conditional distribution is well defined, the marginal chain is uniformly ergodic. By Acosta et al. (2015), the joint chain is also uniformly ergodic. This was also noted by Pazhayidam George (2015).

In addition to faster convergence, there are clear computational gains from using the linchpin variable sampler. The full posterior lies in a $K(D+W)+N$ dimensional space, whereas the marginal posterior for $z$ lies in an $N$-dimensional space. As $D$ and $W$ are often large enough to inhibit obtaining samples from the full posterior, the posterior mean of $\theta$ and $\phi$ can be estimated by the Rao-Blackwellized estimators. Note that,

$$\mathrm{E}[\theta_{ik} \mid z, w, a, b] = \frac{n_{i\cdot k} + a}{n_{i\cdot\cdot} + Ka} \quad \text{and} \quad \mathrm{E}[\phi_{kt} \mid z, w, a, b] = \frac{n_{\cdot tk} + b}{n_{\cdot\cdot k} + Wb} , \tag{6.6}$$

If for samples $l = 1, \ldots, n$, $n_{(\cdot)}^{l}$ denotes the respective sample counts, the Rao-

Blackwellized estimates for the posterior mean of $\theta$ and $\phi$ are

$$\hat{\theta}_{ik,RB} = \frac{1}{n}\sum_{l=1}^{n} \frac{n_{i\cdot k}^{l} + a}{n_{i\cdot\cdot}^{l} + Ka} \quad \text{and} \quad \hat{\phi}_{kt,RB} = \frac{1}{n}\sum_{l=1}^{n} \frac{n_{\cdot tk}^{l} + b}{n_{\cdot\cdot k}^{l} + Wb}\,. \tag{6.7}$$

Geyer (1995) demonstrated that Rao-Blackwellized estimators can have larger variance than standard Monte Carlo estimators. However, in the context of LDA, the computational gains can be significant since obtaining samples from the full posterior can be expensive. Thus, the linchpin variable sampler allows for lower computational costs than the full Gibbs sampler.

## 6.2 Bayesian Penalized Regression

We study the rates of convergence for the Gibbs samplers used in three different Bayesian penalized regression models. For all three Gibbs samplers we conclude that the Markov chains are geometrically ergodic regardless of the number of covariates. The content of the paper is primarily contained in Vats (2016).

Let $X$ be an $F$-invariant Harris ergodic Markov chain defined on the state space $\mathcal{X}$. Recall that a Markov chain is geometrically ergodic if there exists a function $M : \mathcal{X} \to \mathbb{R}^{+}$ and some $0 \leq t < 1$ such that for all $n \in \mathbb{N}$ and for all $x \in \mathcal{X}$

$$\|P^{n}(x, \cdot) - F(\cdot)\|_{TV} \leq M(x)t^{n}\,. \tag{6.8}$$

Geometric ergodicity is often demonstrated by establishing a drift condition and an associated *minorization* condition. A drift condition is said to hold if there exists a function $V : \mathcal{X} \to [0, \infty)$, and constants $0 < \phi < 1$ (this $\phi$ being different from the $\phi$ in LDA) and $L < \infty$ such that for all $x_0 \in \mathcal{X}$

$$\mathrm{E}\left[V(x) \mid x_0\right] \leq \phi V(x_0) + L\,, \tag{6.9}$$

where the expectation is with respect to the Markov chain transition kernel. For $d > 0$, consider the set $C_d = \{x : V(x) \leq d\}$. A minorization condition holds if there exists an $\epsilon > 0$ and a distribution $Q$ such that for all $x_0 \in C_d$

$$P(x_0, \cdot) \geq \epsilon Q(\cdot). \tag{6.10}$$

It is well known that both (6.9) and (6.10) together imply geometric ergodicity (see Meyn and Tweedie (2009) and Jones and Hobert (2001)). The *drift rate* $\phi$ determines how fast the Markov chain drifts back to the *small set* $C_d$. A drift rate close to one signifies slower convergence and a smaller value indicates faster convergence. See Jones and Hobert (2001) for a heuristic explanation.

When a drift condition holds, Meyn and Tweedie (2009) explain that the function $M$ in (6.8) is proportional to the *drift function* $V$. Thus, minimizing $V$ over the state space leads to the tightest bound for that choice of $V$. This leads to default starting values for the Markov chain.

### 6.2.1 Bayesian Fused Lasso

The Bayesian fused lasso (BFL) model we consider here is different from the BFL model formulated in Kyung et al. (2010). Let $y \in \mathbb{R}^n$ be the observed realization of the response $Y$, $X$ be the $n \times p$ model matrix, and $\beta \in \mathbb{R}^p$ be a regression coefficient vector. Kyung et al. (2010) present the following BFL hierarchical structure:

$$\begin{aligned} Y \mid \beta, \sigma^2, \tau^2 &\sim N_n(X\beta, \sigma^2 I_n) \\ \beta \mid \tau^2, w^2, \sigma^2 &\sim N_p(0, \sigma^2 \Sigma_\beta) \\ \tau^2 &\sim \prod_{i=1}^{p} \frac{\lambda_1^2}{2} e^{-\lambda_1 \tau_i^2/2} d\tau_i^2 \qquad \text{where } \tau_i^2 > 0 \end{aligned} \tag{6.11}$$

$$w^2 \sim \prod_{i=1}^{p-1} \frac{\lambda_2^2}{2} e^{-\lambda_2 w_i^2/2} dw_i^2 \qquad \text{where } w_i^2 > 0$$
$$\sigma^2 \sim \text{Inverse-Gamma}(\alpha, \xi),$$

where $\alpha, \xi \geq 0$ and $\lambda_1, \lambda_2 > 0$ are known and $\Sigma_\beta$ is such that $\Sigma_\beta^{-1}$ is the following tridiagonal matrix,

$$\begin{aligned} \text{Main diagonal:} \quad & \left\{ \frac{1}{\tau_i^2} + \frac{1}{w_{i-1}^2} + \frac{1}{w_i^2} \ : i = 1, \ldots, p \right\} \\ \text{Off diagonals:} \quad & \left\{ -\frac{1}{w_i^2} \ : i = 1, \ldots, p-1 \right\}. \end{aligned}$$

Here we assume that $(1/w_0^2) = (1/w_p^2) = 0$. Specifically, $\Sigma_\beta^{-1}$ takes the following form,

$$\Sigma_\beta^{-1} = \begin{bmatrix} \frac{1}{\tau_1^2} + \frac{1}{w_1^2} & -\frac{1}{w_1^2} & 0 & \ldots & 0 \\ -\frac{1}{w_1^2} & \frac{1}{\tau_2^2} + \frac{1}{w_1^2} + \frac{1}{w_2^2} & -\frac{1}{w_2^2} & \ldots & 0 \\ 0 & -\frac{1}{w_2^2} & \frac{1}{\tau_3^2} + \frac{1}{w_2^2} + \frac{1}{w_3^2} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ddots & \ldots \\ 0 & 0 & \ldots & \frac{1}{\tau_{p-1}^2} + \frac{1}{w_{p-2}^2} + \frac{1}{w_{p-1}^2} & -\frac{1}{w_{p-1}^2} \\ 0 & 0 & \ldots & -\frac{1}{w_{p-1}^2} & \frac{1}{\tau_p^2} + \frac{1}{w_{p-1}^2} \end{bmatrix}. \tag{6.12}$$

Kyung et al. (2010) incorrectly state that the priors in (6.11) lead to the following marginal prior on $\beta$ given $\sigma^2$.

$$\pi(\beta \mid \sigma^2) \propto \exp\left( -\frac{\lambda_1}{\sigma} \sum_{j=1}^{p} |\beta_j| - \frac{\lambda_2}{\sigma} \sum_{j=1}^{p-1} |\beta_{j+1} - \beta_j| \right). \tag{6.13}$$

The independent exponential priors on $\tau^2$ and $w^2$ do not lead to the correct marginal prior in (6.13). We show that the correct prior leading to (6.13) is

$$\pi(\tau^2, w^2) \propto \det\left(\Sigma_\beta\right)^{1/2} \left(\prod_{i=1}^{p} \left(\tau_i^2\right)^{-1/2} e^{-\lambda_1 \tau_i^2/2}\right) \left(\prod_{i=1}^{p-1} \left(w_i^2\right)^{-1/2} e^{-\lambda_2 w_i^2/2}\right). \quad (6.14)$$

In Appendix D.4, we show that the prior on $(\tau^2, w^2)$ in (6.14) is proper and leads to the prior in (6.13). Thus, our model formulation is a correct BFL model.

The resulting full conditionals are,

$$\begin{aligned}
\beta \mid \sigma^2, \tau^2, \gamma^2, y &\sim N_p\left((X^TX + \Sigma_\beta^{-1})^{-1}X^Ty, \sigma^2(X^TX + \Sigma_\beta^{-1})^{-1}\right) \\
\frac{1}{\tau_i^2} \mid \beta, \sigma^2, y &\overset{ind}{\sim} \text{Inv-Gaussian}\left(\sqrt{\frac{\lambda_1^2\sigma^2}{\beta_i^2}}, \lambda_1^2\right),\ i = 1, \ldots, p \\
\frac{1}{w_i^2} \mid \beta, \sigma^2, y &\overset{ind}{\sim} \text{Inv-Gaussian}\left(\sqrt{\frac{\lambda_2^2\sigma^2}{(\beta_{i+1} - \beta_i)^2}}, \lambda_2^2\right),\ i = 1, \ldots, p-1 \\
\sigma^2 \mid \beta, \tau^2, \lambda^2, y & \\
&\sim \text{Inv-Gamma}\left(\frac{n + p + 2\alpha}{2}, \frac{(y - X\beta)^T(y - X\beta) + \beta^T\Sigma_\beta^{-1}\beta + 2\xi}{2}\right).
\end{aligned} \quad (6.15)$$

Notice that the full conditionals for $\tau^2$ and $w^2$ are independent and thus can be updated in one block. This reduces the four variable Gibbs sampler to a three variable Gibbs sampler. If $(\beta_{(n)}, \tau^2_{(n)}, w^2_{(n)}, \sigma^2_{(n)})$ is the current state of the Gibbs sampler the $(n+1)$th state is obtained as follows.

---

1. Draw $\sigma^2_{(n+1)}$ from $f(\sigma^2 \mid \beta_{(n)}, \tau^2_{(n)}, w^2_{(n)}, y)$.
2. Draw $\left(1/\tau^2_{(n+1)}, 1/w^2_{(n+1)}\right)$ from $f(1/\tau^2 \mid \beta_{(n)}, \sigma^2_{(n+1)}, y)\ f(1/w^2 \mid \beta_{(n)}, \sigma^2_{(n+1)}, y)$.
3. Draw $\beta_{(n+1)}$ from $f(\beta \mid \tau^2_{(n+1)}, w^2_{(n+1)}, \sigma^2_{(n+1)}, y)$.

---

The full conditionals in (6.15) lead to a three variable deterministic scan Gibbs sampler. Note that the full conditional distribution of $1/\tau_i^2$ is an Inverse-Gaussian with mean parameter $\sqrt{\lambda_1^2\sigma^2/\beta_i^2}$. If the starting value for any $\beta_i$ is zero, this Inverse-Gaussian is still well defined as it is an Inverse-Gamma distribution with shape parameter 1/2 and rate parameter $\lambda_1^2/2$. The same is true for the full conditional of $1/w_i^2$.

We define the drift function $V_{BFL} : \mathbb{R}^p \times \mathbb{R}_+^p \times \mathbb{R}_+^{p-1} \times \mathbb{R}_+ \to [0, \infty)$ as

$$V_{BFL}(\beta, \tau^2, w^2, \sigma^2) = (y - X\beta)^T(y - X\beta) + \beta^T\Sigma_\beta^{-1}\beta + \frac{\lambda_1^2}{4}\sum_{i=1}^{p}\tau_i^2 + \frac{\lambda_2^2}{4}\sum_{i=1}^{p-1}w_i^2. \quad (6.16)$$

**Theorem 6.1** If $n \geq 3$, a drift condition and an associated minorization condition hold for $V_{BFL}$. Thus, the three variable Gibbs sampler for the BFL is geometrically ergodic. □

**Remark 6.1** In Appendix D.5, we arrive at the drift rate

$$\phi_{BFL} = \max\left\{\frac{p}{n + p + 2\alpha - 2}, \frac{1}{2}\right\}.$$

Thus, $\phi_{BFL}$ is no better than 1/2 and as $p$ increases, the drift rate approaches one. This leads us to conclude that convergence may be slower for large $p$ problems. □

**Remark 6.2** The tightest bound for a given $V_{BFL}$ is obtained by minimizing $V_{BFL}$ over the state space. Thus, a default starting value is $\beta_0$ being the frequentist fused lasso estimate, $\tau_{0,i}^2 = 2|\beta_{0,i}|/\lambda_1$ and $w_{0,i}^2 = 2|\beta_{0,i+1} - \beta_{0,i}|/\lambda_2$. See Appendix D.5.1 for details. □

**Remark 6.3** Our result requires no conditions on the design matrix $X$, the dimension of the regression coefficient vector $\beta$, or the tuning parameters $\lambda_1, \lambda_2$. □

### 6.2.2 Bayesian Group Lasso

Recall that $y$ is the observed realization of the response $Y$, $X$ is the $n \times p$ design matrix and $\beta$ is a $p$-dimensional regression coefficient vector. For a fixed $K$, partition $\beta$ into $K$ groups of sizes $m_1, m_2, \ldots, m_K$, the groups being denoted by $\beta_{G_1}, \beta_{G_2}, \ldots, \beta_{G_K}$. Kyung et al. (2010) present the following Bayesian group lasso (BGL) model:

$$\begin{aligned} Y \mid \beta, \sigma^2 &\sim N_n(X\beta, \sigma^2 I_n) \\ \beta_{G_k} \mid \sigma^2, \tau^2 &\overset{ind}{\sim} N_{m_k}(0, \sigma^2 \tau_k^2 I_{m_k}) \quad k = 1, \ldots, K \\ \tau_k^2 &\overset{ind}{\sim} \text{Gamma}\left(\frac{m_k + 1}{2}, \frac{\lambda^2}{2}\right) \quad k = 1, \ldots, K \\ \sigma^2 &\sim \text{Inverse-Gamma}(\alpha, \xi), \end{aligned} \tag{6.17}$$

where $\lambda > 0$, $\alpha, \xi \geq 0$ are fixed and the probability density function of a Gamma$(a, b)$ is proportional to $x^{a-1}e^{-bx}$. Define

$$D_\tau = \text{diag}( \underbrace{\tau_1^2, \ldots, \tau_1^2}_{m_1}, \underbrace{\tau_2^2, \ldots, \tau_2^2}_{m_2}, \ldots, \underbrace{\tau_K^2, \ldots, \tau_K^2}_{m_K}) .$$

The BGL model in (6.17) leads to the following full conditionals for $\beta, \tau^2$ and $\sigma^2$:

$$\begin{aligned} &\beta \mid \sigma^2, \tau^2, y \sim N_p\left((X^T X + D_\tau^{-1})^{-1} X^T y, \sigma^2 (X^T X + D_\tau^{-1})^{-1}\right) \\ &\frac{1}{\tau_k^2} \mid \beta, \sigma^2, y \overset{ind}{\sim} \text{Inv-Gaussian}\left(\sqrt{\frac{\lambda^2 \sigma^2}{\beta_{G_k}^T \beta_{G_k}}}, \lambda^2\right), \text{ for } k = 1, \ldots, K \\ &\sigma^2 \mid \beta, \tau^2, y \sim \text{Inv-Gamma}\left(\frac{n + p + 2\alpha}{2}, \frac{(y - X\beta)^T (y - X\beta) + \beta^T D_\tau^{-1} \beta + 2\xi}{2}\right). \end{aligned} \tag{6.18}$$

The full conditionals lead to a three variable Gibbs sampler.

**Remark 6.4** Kyung et al. (2010) propose a $K + 2$ variable Gibbs sampler where the variables are $\beta_{G_1}, \beta_{G_2}, \ldots, \beta_{G_K}, \tau^2$ and $\sigma^2$. For this sampler the full conditionals for

$\sigma^2$ and $\tau^2$ are the same as above, but the full conditional for each $\beta_{G_k}$ is

$$\begin{aligned}&\beta_{G_k} \mid \beta_{-G_k}, \sigma^2, \tau^2, y \\ &\sim N_{m_k}\left(\left(X_k^T X_k + \tau_k^{-2} I_{m_k}\right)^{-1} X_k^T \left(y - \sum_{k' \neq k} X_{k'} \beta_{G_{k'}}\right), \sigma^2 \left(X_k^T X_k + \tau_k^{-2} I_{m_k}\right)^{-1}\right).\end{aligned}$$

Here $X_k$ is the submatrix of $X$ with columns corresponding to the group $\beta_{G_k}$. Kyung et al. (2010) had an error in their full conditional; they had

$$\left(y - \frac{1}{2}\sum_{k' \neq k} X_{k'} \beta_{G_{k'}}\right) \text{ instead of } \left(y - \sum_{k' \neq k} X_{k'} \beta_{G_{k'}}\right).$$

The motivation for using the $(K+2)$-variable sampler is to avoid the $p \times p$ matrix inversion of $(X^T X + D_\tau^{-1})$, and instead to do $K$ matrix inversions each of size $m_k \times m_k$. This reduces the computational cost from $O(p^3)$ to $O(\sum_{k=1}^K m_k^3)$. Such a technique was also discussed in Ishwaran and Rao (2005). In addition, Bhattacharya et al. (2015) recently proposed a linear (in $p$) time sampling algorithm to sample from high-dimensional normal distributions of the form in (6.18). Using their method, the computational cost of drawing from the full conditional of $\beta$ is $O(p)$, and thus the $K+2$ variable Gibbs sampler is not required. □

We will study the convergence of the three variable Gibbs sampler which we describe as follows. If $(\beta_{(n)}, \tau^2_{(n)}, \sigma^2_{(n)})$ is the current state of the Gibbs sampler, the $(n+1)$st state is obtained as follows.

---

1. Draw $\sigma^2_{(n+1)}$ from $f(\sigma^2 \mid \beta_{(n)}, \tau^2_{(n)}, y)$.
2. Draw $1/\tau^2_{(n+1)}$ from $f(1/\tau^2 \mid \beta_{(n)}, \sigma^2_{(n+1)}, y)$.
3. Draw $\beta_{(n+1)}$ from $f(\beta \mid \tau^2_{(n+1)}, \sigma^2_{(n+1)}, y)$.

---

The full conditionals in (6.18) lead to a three variable deterministic scan Gibbs sampler that samples from the posterior distribution of $(\beta, \tau^2, \sigma^2)$. We again note that the full conditional distribution of $1/\tau_k^2$ is an Inverse-Gaussian with mean parameter $\sqrt{\lambda^2\sigma^2/\beta_{G_k}^T\beta_{G_k}}$. If the starting value for any $\beta_{G_k}$ is the zero vector, this Inverse-Gaussian is still well defined as it is an Inverse-Gamma distribution with shape parameter $1/2$ and rate parameter $\lambda^2/2$.

As in the proof of geometric ergodicity of the Gibbs sampler in BFL, we establish a drift and a minorization condition. Define the drift function $V_{BGL} : \mathbb{R}^p \times \mathbb{R}_+^K \times \mathbb{R}_+ \to [0, \infty)$ as

$$V_{BGL}(\beta, \tau^2, \sigma^2) = (y - X\beta)^T(y - X\beta) + \beta^T D_\tau^{-1}\beta + \frac{\lambda^2}{4}\sum_{k=1}^{K}\tau_k^2\,. \tag{6.19}$$

**Theorem 6.2** If $n \geq 3$, a drift condition and an associated minorization condition hold for $V_{BGL}$. Thus, the three variable Gibbs sampler for the BGL is geometrically ergodic. □

**Remark 6.5** As in BFL, the drift rate

$$\phi_{BGL} = \max\left\{\frac{p}{n + p + 2\alpha - 2}, \frac{1}{2}\right\},$$

is no better than $1/2$ as $n$ increases and approaches 1 as $p$ increases. Thus, convergence may be slower for large $p$ problems. □

**Remark 6.6** The tightest bound is constructed by minimizing $V_{BGL}$ over the space of starting values of the chain. Thus a reasonable starting value for the Markov chain is $\beta_0$, the frequentist group lasso estimate and $\tau_{0,k}^2 = 2\sqrt{\beta_{0,G_k}^T\beta_{0,G_k}}/\lambda$. See Appendix D.6.1 for details. □

**Remark 6.7** Our result requires no conditions on the model matrix $X$, the dimension of the regression coefficient vector $\beta$, or the tuning parameter $\lambda$. □

**Remark 6.8** Since for $K = p$, the Bayesian group lasso is the Bayesian lasso, our geometric ergodicity result holds for the Bayesian lasso as well. Geometric ergodicity of the Bayesian lasso was demonstrated by Khare and Hobert (2013) under exactly the same conditions. Our result on the starting values in Remark 6.6 also holds for the Bayesian lasso. □

### 6.2.3 Bayesian Sparse Group Lasso

As before, let $y$ be the observed realization of the response $Y$, $X$ be the $n \times p$ model matrix and $\beta$ be the $p$-dimensional regression coefficient vector. For a fixed $K$, partition $\beta$ into $K$ groups of sizes $m_1, m_2, \ldots, m_K$, the groups being denoted by $\beta_{G_1}, \beta_{G_2}, \ldots, \beta_{G_K}$. The Bayesian sparse group lasso (BSGL) model introduced by Xu and Ghosh (2015) induces sparsity on individual coefficients in addition to the groups. Before introducing the model, we present some definitions.

Let $\gamma^2_{1,1}, \gamma^2_{1,2}, \ldots, \gamma^2_{1,m_1}, \ldots, \gamma^2_{K,m_K}$ and $\tau^2_1, \ldots, \tau^2_K$ be variables defined on the positive reals. For each group $k$ define,

$$V_k = \text{Diag}\left\{ \left( \frac{1}{\tau_k^2} + \frac{1}{\gamma_{k,j}^2} \right)^{-1} \; : \; j = 1, \ldots, m_k \right\}.$$

The notation $\gamma^2_{k,j}$ is used purely for convenience and can easily be replaced with $\gamma^2_i$ for $i = 1, \ldots, p$, since each $\gamma$ in and across $V_k$ can be different. The BSGL model formulated by Xu and Ghosh (2015) is

$$\begin{aligned} Y \mid \beta, \sigma^2, \tau^2 &\sim N_n(X\beta, \sigma^2 I_n) \\ \beta_{G_k} \mid \sigma^2, \tau^2 &\overset{iid}{\sim} N_{m_k}(0, \sigma^2 V_k) \quad \text{for } k = 1, \ldots, K \end{aligned} \tag{6.20}$$

$$\begin{aligned}\pi(\gamma_{k,1},\ldots,\gamma_{k,m_k},\tau_k^2) &= \pi_k \quad \text{independently for } k=1,\ldots,K\\ \sigma^2 &\sim \text{Inverse-Gamma}(\alpha,\xi),\end{aligned}$$

where $\alpha,\xi \geq 0$ are fixed and the independent prior on each $(\gamma_{k,1},\ldots,\gamma_{k,m_k},\tau_k^2)$ is

$$\pi_k \propto \prod_{j=1}^{m_k}\left[\left(\gamma_{k,j}^2\right)^{-\frac{1}{2}}\left(\frac{1}{\gamma_{k,j}^2}+\frac{1}{\tau_k^2}\right)^{-\frac{1}{2}}\right]\left(\tau_k^2\right)^{-\frac{1}{2}}\exp\left\{-\frac{\lambda_2^2}{2}\sum_{j=1}^{m_k}\gamma_{k,j}^2-\frac{\lambda_1^2}{2}\tau_k^2\right\}. \quad (6.21)$$

Here $\lambda_1,\lambda_2>0$ are fixed. Xu and Ghosh (2015) show that the prior in (6.21) is proper with the normalizing constant being a function of $\lambda_1$ and $\lambda_2$. They also derive the full conditionals for $\beta,\gamma^2,\tau^2$ and $\sigma^2$ leading to a four variable deterministic scan Gibbs sampler. We will note that $\gamma^2$ and $\tau^2$ can be updated together in a block leading to a three variable Gibbs sampler.

Define $V_{\tau,\gamma}$ to be the diagonal matrix with its diagonals being the diagonals of $V_1,\ldots,V_K$ in that sequence. In addition, when we use two subscripts on $\beta$ as in $\beta_{k,j}$, then we are referring to the $j$th coefficient in the $k$th group. The BSGL model in (6.20) leads to the following full conditionals for $\beta,\tau^2,\gamma^2$ and $\sigma^2$.

$$\begin{aligned}&\beta \mid \sigma^2,\tau^2,\gamma^2,y \sim N_p\left((X^TX+V_{\tau,\gamma}^{-1})^{-1}X^Ty,\sigma^2(X^TX+V_{\tau,\gamma}^{-1})^{-1}\right)\\ &\frac{1}{\tau_k^2}\mid \beta,\sigma^2,y \sim \text{Inv-Gaussian}\left(\sqrt{\frac{\lambda_1^2\sigma^2}{\beta_{G_k}^T\beta_{G_k}}},\lambda_1^2\right),\ \text{independently for all } k\\ &\frac{1}{\gamma_{k,j}^2}\mid \beta,\sigma^2,y \sim \text{Inv-Gaussian}\left(\sqrt{\frac{\lambda_2^2\sigma^2}{\beta_{k,j}^2}},\lambda_2^2\right),\ \text{independently for all } k,j\\ &\sigma^2 \mid \beta,\tau^2,\lambda^2,y\\ &\sim \text{Inv-Gamma}\left(\frac{n+p+2\alpha}{2},\frac{(y-X\beta)^T(y-X\beta)+\beta^TV_{\tau,\gamma}^{-1}\beta+2\xi}{2}\right).\end{aligned} \quad (6.22)$$

Notice that the full conditionals for $\tau^2$ and $\gamma^2$ are independent and thus can be updated in one block. This reduces the four variable Gibbs sampler to a three variable Gibbs sampler. If $(\beta_{(n)}, \tau^2_{(n)}, \gamma^2_{(n)}, \sigma^2_{(n)})$ is the current state of the Gibbs sampler, the $(n+1)$th state is obtained as follows.

---

1. Draw $\sigma^2_{(n+1)}$ from $f(\sigma^2 \mid \beta_{(n)}, \tau^2_{(n)}, \gamma^2_{(n)}, y)$.
2. Draw $\left(1/\tau^2_{(n+1)}, 1/\gamma^2_{(n+1)}\right)$ from $f(1/\tau^2 \mid \beta_{(n)}, \sigma^2_{(n+1)}, y)\ f(1/\gamma^2 \mid \beta_{(n)}, \sigma^2_{(n+1)}, y)$.
3. Draw $\beta_{(n+1)}$ from $f(\beta \mid \tau^2_{(n+1)}, \gamma^2_{(n+1)}, \sigma^2_{(n+1)}, y)$.

---

The full conditionals in (6.22) lead to a three variable Gibbs sampler that samples from the posterior distribution of $(\beta, \tau^2, \gamma^2, \sigma^2)$. Define the drift function $V_{BSGL} : \mathbb{R}^p \times \mathbb{R}^K_+ \times \mathbb{R}^p_+ \times \mathbb{R}_+ \to [0, \infty)$ as

$$V_{BSGL}(\beta, \tau^2, \gamma^2, \sigma^2) = (y - X\beta)^T(y - X\beta) + \beta^T V^{-1}_{\tau,\gamma}\beta + \frac{\lambda_1^2}{4}\sum_{k=1}^{K}\tau_k^2 + \frac{\lambda_2^2}{4}\sum_{k=1}^{K}\sum_{j=1}^{m_k}\gamma^2_{k,j}. \tag{6.23}$$

**Theorem 6.3** If $n \geq 3$, a drift condition and an associated minorization condition hold for $V_{BSGL}$. Thus, the three variable Gibbs sampler for the BSGL is geometrically ergodic.

**Remark 6.9** Let $M = \max_k m_k$. Then in Appendix D.7 the drift rate is determined to be

$$\phi_{BSGL} = \max\left\{\frac{p}{n+p+2\alpha-2}, \frac{\left(1+\frac{\lambda_2^2}{\lambda_1^2}\right)}{2\left(1+\frac{\lambda_1^2}{\lambda_2^2}+\frac{\lambda_2^2}{\lambda_1^2}\right)}, \frac{\left(1+\frac{\lambda_1^2}{\lambda_2^2}\right)}{2M\left(1+\frac{\lambda_1^2}{\lambda_2^2}+\frac{\lambda_2^2}{\lambda_1^2}\right)}\right\}.$$

Unlike the drift rate in BFL and BGL, the drift rate here can be less than 1/2. However, it is likely that $p$ is large enough so that $\phi_{BSGL}$ is determined by the first

term $p/(n+p+2\alpha-2)$. In this case again, the drift rate will tend to 1 as $p$ increases and the thus convergence may be slower for large $p$ problems. □

**Remark 6.10** The tightest bound for this choice of $V_{BSGL}$ and resulting $\phi_{BSGL}$ and $L_{BSGL}$ is obtained by minimizing $V_{BSGL}$ over the space of starting values of the chain. Thus a reasonable starting value for the Markov chain is $\beta_0$, the sparse group lasso estimate, $\tau^2_{0,k} = 2\sqrt{\beta^T_{0,G_k}\beta_{0,G_k}}/\lambda_1$ and $\gamma^2_{0,k} = 2|\beta_{0,k,j}|/\lambda_2$. See Appendix D.7.1. □

**Remark 6.11** We require no conditions on $p$, the model matrix $X$ or the tuning parameters $\lambda_1, \lambda_2$. □

# Chapter 7

# A Guide for Users

The goal of this dissertation was to develop termination procedures for MCMC using multivariate output analysis tools. The literature in multivariate output analysis for MCMC has been lacking, and the work done through this dissertation serves as a building block for further research. The relative volume sequential stopping rule, multivariate effective sample size, and strong consistency of the mBM estimator have been presented in Vats et al. (2016a). Strong consistency of the mSV estimator is shown in Vats et al. (2016b).

The `R` package `mcmcse` (Flegal et al. (2015)) computes the mBM estimator, the mSV estimator, the effective sample size, and the minimum number of effective samples needed. The package is open source and is available on CRAN. Our methods can be implemented by users in the following way.

1. The user must first identify the target distribution, $F$. Further, the user should decide the vector of quantities of interest. That is, $g$ and $\theta$ should be chosen.

2. Since $\theta$ is now fixed, $p$ is known. In addition, the relative tolerance level for termination, $\epsilon$, and the confidence level $1-\alpha$ must be decided. We recommend using $\epsilon \leq .05$. This allows the calculation of the minimum number of effective sample size needed to be within an $\epsilon$ level of relative tolerance using (3.5).

3. The user chooses an MCMC sampler to draw samples from $F$. The sampler should be such that it is polynomially ergodic. Since this can often be difficult to show, we recommend using the "best" possible sampler.

4. At this time the user must also decide their choice of estimator for $\Sigma$. We recommend using the mBM estimator as it is significantly faster to compute than the mSV estimator. We also recommend using batch size $b_n = \lfloor n^{1/2} \rfloor$ when the rate of convergence of the Markov chain is unknown.

5. The MCMC sampler should be first run for some $n^*$ iterations. The choice of $n^*$ is such that reasonable estimates of $\Sigma$ and $\Lambda$ can be obtained. We recommend $n^*$ to be larger than the lower bound in (3.5) and be such that $\Sigma_n$ is positive definite.

6. After $n^*$ iterations, simulation is terminated when the estimated ESS is larger than the lower bound in (3.5). At termination, the final estimate of $\theta$ is calculated.

The performance of the methods above depend on the rate of convergence of the chosen MCMC sampler. The choice of $\epsilon, b_n$, and $n^*$ may be influenced by the convergence rate of the Markov chain. For this reason, we do not recommend using our method as a black box and encourage users to adapt our methods according to the chosen MCMC sampler.

# Appendix A

# Proof of Theorem 3.1

By Vats et al. (2016b), since $\mathrm{E}_F\|g\|^{2+\delta} < \infty$ and $\{X_t\}$ is polynomially ergodic of order $m > (1+\epsilon_1)(1+2/\delta)$, (2.4) holds with $\psi(n) = n^{1/2-\lambda}$ for $\lambda > 0$. This implies the existence of an functional CLT. Now, recall that

$$T^*(\epsilon) = \inf\left\{n \geq 0 : \mathrm{Vol}(C_\alpha(n))^{1/p} + s(n) \leq \epsilon K_n(g(X), p)\right\}.$$

For cleaner notation we will use $K_n$ for $K_n(g(X), p)$ and $K$ for $K(g(X), p)$. First, we show that as $\epsilon \to 0$, $T^*(\epsilon) \to \infty$. Recall $s(n) = \epsilon K_n I(n < n^*) + n^{-1}$. Consider the rule,

$$t(\epsilon) = \inf\{n \geq 0 : s(n) < \epsilon K_n\} = \inf\{n \geq 0 : \epsilon I(n < n^*) + (K_n n)^{-1} < \epsilon\}.$$

As $\epsilon \to 0$, $t(\epsilon) \to \infty$. Since $T^*(\epsilon) > t(\epsilon)$, $T^*(\epsilon) \to \infty$ as $\epsilon \to 0$.

Define $V(n) = \mathrm{Vol}(C_\alpha(n))^{1/p} + s(n)$. Then $T^*(\epsilon) = \inf\{n \geq 0 : V(n) \leq \epsilon K_n\}$. Let

$$d_{\alpha,p} = \frac{2\pi^{p/2}}{p\Gamma(p/2)}\left(\chi^2_{1-\alpha,p}\right)^{p/2}.$$

Since $s(n) = o(n^{-1/2})$ and $\Sigma$ is positive definite,

$$\begin{aligned} &n^{1/2}V(n) \\ &= n^{1/2}\left[n^{-1/2}\left(\frac{2\pi^{p/2}}{p\Gamma(p/2)}\left(\frac{p(a_n-1)}{(a_n-p)}F_{1-\alpha,p,a_n-p}\right)^{p/2}|\Sigma_n|^{1/2}\right)^{1/p} + s(n)\right] \end{aligned}$$

$$= \left( \frac{2\pi^{p/2}}{p\Gamma(p/2)} \left( \frac{p(a_n - 1)}{(a_n - p)} F_{1-\alpha,p,a_n-p} \right)^{p/2} |\Sigma_n|^{1/2} \right)^{1/p} + n^{1/2} s(n)$$

$$\to (d_{\alpha,p}|\Sigma|^{1/2})^{1/p} > 0 \text{ w.p. } 1 \text{ as } n \to \infty. \tag{A.1}$$

By definition of $T^*(\epsilon)$

$$V(T^*(\epsilon) - 1) > \epsilon K_{T^*(\epsilon)-1}, \tag{A.2}$$

and there exists a random variable $Z(\epsilon)$ on $[0, 1]$ such that,

$$V(T^*(\epsilon) + Z(\epsilon)) \leq \epsilon K_{T^*(\epsilon)+Z(\epsilon)}. \tag{A.3}$$

Since $K_n$ is strongly consistent for $K$ and $T^*(\epsilon) \to \infty$ w.p. 1 as $\epsilon \to 0$,

$$K_{T^*(\epsilon)} \to K \quad \text{w.p. } 1, \tag{A.4}$$

Using (A.1), (A.2), (A.3), and (A.4)

$$\limsup_{\epsilon \to 0} \epsilon T^*(\epsilon)^{1/2} \leq \limsup_{\epsilon \to 0} \frac{T^*(\epsilon)^{1/2} V(T^*(\epsilon) - 1)}{K_{T^*(\epsilon)-1}} = d_{\alpha,p}^{1/p} \frac{|\Sigma|^{1/2p}}{K} \quad \text{w.p. } 1$$

$$\liminf_{\epsilon \to 0} \epsilon T^*(\epsilon)^{1/2} \geq \liminf_{\epsilon \to 0} \frac{T^*(\epsilon)^{1/2} V(T^*(\epsilon) + Z(\epsilon))}{K_{T^*(\epsilon)+Z(\epsilon)}} = d_{\alpha,p}^{1/p} \frac{|\Sigma|^{1/2p}}{K} \quad \text{w.p. } 1.$$

Thus,

$$\lim_{\epsilon \to 0} \epsilon T^*(\epsilon)^{1/2} = d_{\alpha,p}^{1/p} \frac{|\Sigma|^{1/2p}}{K}. \tag{A.5}$$

Using (A.5) and the existence of a functional CLT, by a standard random-time-

change argument (Billingsley, 1968, p. 151)

$$\sqrt{T^*(\epsilon)}\Sigma_{T^*(\epsilon)}^{-1/2}(\theta_{T^*(\epsilon)} - \theta) \xrightarrow{d} N_p(0, I_p) \quad \text{as } \epsilon \to 0.$$

Finally,

$$\begin{aligned}
&Pr[\theta \in C_\alpha(T^*(\epsilon))] \\
&= Pr\left[T^*(\epsilon)(\theta_{T^*(\epsilon)} - \theta)^T \Sigma_{T(\epsilon)}^{-1}(\theta_{T^*(\epsilon)} - \theta)\right. \\
&\qquad \left. \leq \frac{p(a_{T^*(\epsilon)} - 1)}{(a_{T^*(\epsilon)} - p)} F_{1-\alpha,p,a_{T^*(\epsilon)}-p}; |\Sigma_{T^*(\epsilon)}| \neq 0\right] \\
&\quad + Pr\left[\theta \in C_\alpha(T^*(\epsilon)); |\Sigma_{T^*(\epsilon)}| = 0\right] \\
&\to 1 - \alpha \text{ w.p. } 1 \text{ as } n \to \infty \text{ since } Pr(|\Sigma_{T^*(\epsilon)}| = 0) \to 0 \text{ as } \epsilon \to 0.
\end{aligned}$$

# Appendix B

# Proofs from Chapter 4

Throughout this chapter we use the notation $|\cdot|$ to denote absolute value and not determinant.

## B.1 Preliminaries

Let $\{B(t)\}_{t\geq 0}$ denote a $p$-dimensional standard Brownian motion and let $B^{(i)}$ denote the $i$th component of $B(t)$.

**Lemma B.1 (Csörgő and Révész (1981))** Suppose Condition 4.2 holds, then for all $\epsilon > 0$ and for almost all sample paths, there exists $n_0(\epsilon)$ such that for all $n \geq n_0$ and all $i = 1, \ldots, p$

$$\sup_{0\leq t\leq n-b_n} \sup_{0\leq s\leq b_n} \left|B^{(i)}(t+s) - B^{(i)}(t)\right| < (1+\epsilon)\left(2b_n\left(\log\frac{n}{b_n} + \log\log n\right)\right)^{1/2},$$

$$\sup_{0\leq s\leq b_n} \left|B^{(i)}(n) - B^{(i)}(n-s)\right| < (1+\epsilon)\left(2b_n\left(\log\frac{n}{b_n} + \log\log n\right)\right)^{1/2}, \text{ and}$$

$$\left|B^{(i)}(n)\right| < (1+\epsilon)\sqrt{2n\log\log n}.$$

□

**Corollary B.1 (Damerdji (1994))** Suppose Condition 4.5a holds, then for all $\epsilon > 0$ and for almost all sample paths, there exists $n_0(\epsilon)$ such that for all $n \geq n_0$ and all $i = 1, \ldots, p$

$$|\bar{B}_k^{(i)}(b_n)| \leq \frac{\sqrt{2}}{\sqrt{b_n}}(1+\epsilon)\left(\log\frac{n}{b_n} + \log\log n\right)^{1/2}. \qquad \square$$

Let $L$ be a lower triangular matrix and set $\Sigma = LL^T$. Define $C(t) := LB(t)$ and if $C^{(i)}(t)$ is the $i$th component of $C(t)$, define

$$\bar{C}_l^{(i)}(k) := \frac{1}{k}\left(C^{(i)}(l+k) - C^{(i)}(l)\right) \text{ and } \bar{C}_n^{(i)} := \frac{1}{n}C^{(i)}(n)\,.$$

Since $C^{(i)}(t) \sim N(0, t\Sigma_{ii})$, where $\Sigma_{ii}$ is the $i$th diagonal of $\Sigma$, $C^{(i)}/\sqrt{\Sigma_{ii}}$ is a 1-dimensional standard Brownian motion. As a consequence, we have the following corollaries of Lemma B.1.

**Corollary B.2** Suppose Condition 4.2 holds, then for all $\epsilon > 0$ and for almost all sample paths there exists $n_0(\epsilon)$ such that for all $n \geq n_0$ and all $i = 1, \ldots, p$

$$\left|C^{(i)}(n)\right| < (1+\epsilon)(2n\Sigma_{ii}\log\log n)^{1/2}, \tag{B.1}$$

where $\Sigma_{ii}$ is the $i$th diagonal entry of $\Sigma$. $\square$

**Corollary B.3** Suppose Condition 4.2 holds, then for all $\epsilon > 0$ and for almost all sample paths, there exists $n_0(\epsilon)$ such that for all $n \geq n_0$ and all $i = 1, \ldots, p$

$$\left|\bar{C}_l^{(i)}(k)\right| \leq \frac{1}{k}\sup_{0\leq l\leq n-b_n}\sup_{0\leq s\leq b_n}\left|C^{(i)}(l+s) - C^{(i)}(l)\right| < \frac{1}{k}2(1+\epsilon)(b_n\Sigma_{ii}\log n)^{1/2}, \tag{B.2}$$

where $\Sigma_{ii}$ is the $i$th diagonal entry of $\Sigma$. $\square$

**Lemma B.2 (Kendall and Stuart (1963))** If $Z \sim \chi^2_v$, then for all positive integers $r$ there exists a constant $K := K(r)$ such that $\mathrm{E}[(Z-v)^{2r}] \leq Kv^r$. □

**Lemma B.3 (Billingsley (2008))** For a family of random variables $\{X_n : n \geq 1\}$, if $\mathrm{E}(|X_n|) \leq s_n$ where $s_n$ is a sequence such that $\sum_{n=1}^{\infty} s_n < \infty$, then $X_n \to 0$ w.p. 1 as $n \to \infty$. □

## B.2 Proof of Theorems 4.1 and 4.2

Recall that the lag window $w_n(\cdot)$ is such that it satisfies Condition 4.1. We will require the following results about the lag window $w_n(\cdot)$.

**Lemma B.4 (Damerdji (1991))** Under Condition 4.1,

(i) $\Delta_1 w_n(s) = \sum_{k=s}^{b_n} \Delta_2 w_n(k)$,

(ii) $\sum_{k=s+1}^{b_n} \Delta_1 w_n(k) = w_n(s)$, and

(iii) $\sum_{k=1}^{b_n} \Delta_1 w_n(k) = 1$. □

We will first prove Theorem 4.1 and then use it to prove Theorem 4.2. The proof for Theorem 4.1 is split into several lemmas. We first present some definitions.

Define for $l = 0, \ldots, (n-b_n)$, $\bar{Y}_l(k) = k^{-1}\sum_{t=1}^{k} Y_{l+t}$ and

$$\widehat{\Sigma}_{w,n} = \frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{k=1}^{b_n} k^2 \Delta_2 w_n(k)[\bar{Y}_l(k) - \bar{Y}_n][\bar{Y}_l(k) - \bar{Y}_n]^T . \tag{B.3}$$

For $t = 1, 2, \ldots, n$, define $Z_t = Y_t - \bar{Y}_n$ and

$$d_n = \frac{1}{n}\left\{\sum_{t=1}^{b_n}\Delta_1 w_n(t)\left(\sum_{l=1}^{t-1} Z_l Z_l^T + \sum_{l=n-b_n+t+1}^{n} Z_l Z_l^T\right)\right.$$
$$+\sum_{s=1}^{b_n-1}\left[\sum_{t=1}^{b_n-s}\Delta_1 w_n(s+t)\left(\sum_{l=1}^{t-1}\left(Z_l Z_{l+s}^T + Z_{l+s}Z_l^T\right)\right.\right.$$
$$\left.\left.\left.+\sum_{l=n-b_n+t+1}^{n-s}\left(Z_l Z_{l+s}^T + Z_{l+s}Z_l^T\right)\right)\right]\right\}. \tag{B.4}$$

Notice that in (B.4) we use the convention that empty sums are zero.

**Lemma B.5** Under Condition 4.1, $\widehat{\Sigma}_{w,n} = \Sigma_{SV} - d_n$. □

**Proof**

For $i, j = 1, \ldots, p$, let $\widehat{\Sigma}_{w,ij}$ denote the $(i,j)$th entry of $\widehat{\Sigma}_{w,n}$. Then,

$$\begin{aligned}
&\widehat{\Sigma}_{w,ij}\\
&= \frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{k=1}^{b_n} k^2\Delta_2 w_n(k)\left[\bar{Y}_l^{(i)}(k) - \bar{Y}_n^{(i)}\right]\left[\bar{Y}_l^{(j)}(k) - \bar{Y}_n^{(j)}\right]\\
&= \frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{k=1}^{b_n}\Delta_2 w_n(k)\left[\sum_{t=1}^{k} Y_{l+t}^{(i)} - k\bar{Y}_n^{(i)}\right]\left[\sum_{t=1}^{k} Y_{l+t}^{(j)} - k\bar{Y}_n^{(j)}\right]\\
&= \frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{k=1}^{b_n}\Delta_2 w_n(k)\left[\sum_{t=1}^{k} Z_{l+t}^{(i)}\right]\left[\sum_{t=1}^{k} Z_{l+t}^{(j)}\right]\\
&= \frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{k=1}^{b_n}\Delta_2 w_n(k)\left[\sum_{t=1}^{k} Z_{l+t}^{(i)}Z_{l+t}^{(j)} + \sum_{s=1}^{k-1}\sum_{t=1}^{k-s} Z_{l+t}^{(i)}Z_{l+t+s}^{(j)} + \sum_{s=1}^{k-1}\sum_{t=1}^{k-s} Z_{l+t}^{(j)}Z_{l+t+s}^{(i)}\right].
\end{aligned} \tag{B.5}$$

Notice that in (B.5), we use the convention that empty sums are zero. We will consider each term in (B.5) separately. For the first term, changing the order of summation

and then using Lemma B.4,

$$
\begin{aligned}
&\frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{k=1}^{b_n}\sum_{t=1}^{k}\Delta_2 w_n(k)Z_{l+t}^{(i)}Z_{l+t}^{(j)}\\
&=\frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{t=1}^{b_n}\sum_{k=t}^{b_n}\Delta_2 w_n(k)Z_{l+t}^{(i)}Z_{l+t}^{(j)}\\
&=\frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{t=1}^{b_n}\Delta_1 w_n(t)Z_{l+t}^{(i)}Z_{l+t}^{(j)}\\
&=\frac{1}{n}\sum_{t=1}^{b_n}\Delta_1 w_n(t)\sum_{l=0}^{n-b_n}Z_{l+t}^{(i)}Z_{l+t}^{(j)}\\
&=\sum_{t=1}^{b_n}\Delta_1 w_n(t)\left[\gamma_{n,ij}(0)-\frac{1}{n}\left(Z_1^{(i)}Z_1^{(j)}+\dots\right.\right.\\
&\quad\left.\left.+Z_{t-1}^{(i)}Z_{t-1}^{(j)}+Z_{n-b_n+t+1}^{(i)}Z_{n-b_n+t+1}^{(j)}+\cdots+Z_n^{(i)}Z_n^{(j)}\right)\right]\\
&=\gamma_{n,ij}(0)-\frac{1}{n}\sum_{t=1}^{b_n}\Delta_1 w_n(t)\left(\sum_{l=1}^{t-1}Z_l^{(i)}Z_l^{(j)}+\sum_{l=n-b_n+t+1}^{n}Z_l^{(i)}Z_l^{(j)}\right)\text{ by Lemma B.4.}
\end{aligned}
\tag{B.6}
$$

For the second term in (B.5) we change the order of summation from $l,k,s,t$ to $l,s,k,t$ to $l,s,t,k$ to get

$$
\begin{aligned}
&\frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{k=1}^{b_n}\sum_{s=1}^{k-1}\sum_{t=1}^{k-s}\Delta_2 w_n(k)Z_{l+t}^{(i)}Z_{l+t+s}^{(j)}\\
&=\frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{s=1}^{b_n-1}\sum_{k=s+1}^{b_n}\sum_{t=1}^{k-s}\Delta_2 w_n(k)Z_{l+t}^{(i)}Z_{l+t+s}^{(j)}\\
&=\frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{s=1}^{b_n-1}\sum_{t=1}^{b_n-s}\sum_{k=t+s}^{b_n}\Delta_2 w_n(k)Z_{l+t}^{(i)}Z_{l+t+s}^{(j)}\\
&=\frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{s=1}^{b_n-1}\sum_{t=1}^{b_n-s}\Delta_1 w_n(s+t)Z_{l+t}^{(i)}Z_{l+t+s}^{(j)}\qquad\text{by Lemma B.4}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{s=1}^{b_n-1}\sum_{l=0}^{n-b_n}\sum_{t=1}^{b_n-s}\Delta_1 w_n(s+t)Z_{l+t}^{(i)}Z_{l+t+s}^{(j)} \\
&= \frac{1}{n}\sum_{s=1}^{b_n-1}\sum_{t=1}^{b_n-s}\sum_{l=0}^{n-b_n}\Delta_1 w_n(s+t)Z_{l+t}^{(i)}Z_{l+t+s}^{(j)} \\
&= \frac{1}{n}\sum_{s=1}^{b_n-1}\sum_{t=1}^{b_n-s}\Delta_1 w_n(s+t)\sum_{l=0}^{n-b_n}Z_{l+t}^{(i)}Z_{l+t+s}^{(j)} \\
&= \sum_{s=1}^{b_n-1}\sum_{t=1}^{b_n-s}\Delta_1 w_n(s+t)\left[\gamma_{n,ij}(s) - \frac{1}{n}\sum_{l=1}^{t-1}Z_l^{(i)}Z_{l+s}^{(j)} - \frac{1}{n}\sum_{l=n-b_n+t+1}^{n-s}Z_l^{(i)}Z_{l+s}^{(j)}\right] \\
&= \sum_{s=1}^{b_n-1} w_n(s)\gamma_{n,ij}(s) - \frac{1}{n}\sum_{s=1}^{b_n-1}\sum_{t=1}^{b_n-s}\Delta_1 w_n(s+t)\left[\sum_{l=1}^{t-1}Z_l^{(i)}Z_{l+s}^{(j)}\right. \\
&\quad \left. + \sum_{l=n-b_n+t+1}^{n-s}Z_l^{(i)}Z_{l+s}^{(j)}\right] \quad \text{by Lemma B.4.}
\end{aligned} \tag{B.7}
$$

Repeating the same steps as in the second term we reduce the third term in (B.5) to

$$
\begin{aligned}
&\frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{k=1}^{b_n}\sum_{s=1}^{k-1}\sum_{t=1}^{k-s}\Delta_2 w_n(k)Z_{l+t}^{(j)}Z_{l+t+s}^{(i)} \\
&= \sum_{s=1}^{b_n-1} w_n(s)\gamma_{n,ji}(s) - \frac{1}{n}\sum_{s=1}^{b_n-1}\sum_{t=1}^{b_n-s}\Delta_1 w_n(s+t) \\
&\quad \times \left[\sum_{l=1}^{t-1}Z_l^{(j)}Z_{l+s}^{(i)} + \sum_{l=n-b_n+t+1}^{n-s}Z_l^{(j)}Z_{l+s}^{(i)}\right] \\
&= \sum_{s=1}^{b_n-1} w_n(-s)\gamma_{n,ij}(-s) - \frac{1}{n}\sum_{s=1}^{b_n-1}\sum_{t=1}^{b_n-s}\Delta_1 w_n(s+t) \\
&\quad \times \left[\sum_{l=1}^{t-1}Z_l^{(j)}Z_{l+s}^{(i)} + \sum_{l=n-b_n+t+1}^{n-s}Z_l^{(j)}Z_{l+s}^{(i)}\right] .
\end{aligned} \tag{B.8}
$$

Using (B.6), (B.7), and (B.8) in (B.5)

$$\widehat{\Sigma}_{w,ij}$$

$$
\begin{aligned}
&= \gamma_{n,ij}(0) + \sum_{s=1}^{b_n-1} w_n(s)\gamma_{n,ij}(s) + \sum_{s=-(b_n-1)}^{-1} w_n(s)\gamma_{n,ij}(s) \\
&- \frac{1}{n}\sum_{t=1}^{b_n} \Delta_1 w_n(t) \left( \sum_{l=1}^{t-1} Z_l^{(i)} Z_l^{(j)} + \sum_{l=n-b_n+t+1}^{n} Z_l^{(i)} Z_l^{(j)} \right) \\
&- \frac{1}{n}\sum_{s=1}^{b_n-1}\sum_{t=1}^{b_n-s} \Delta_1 w_n(s+t) \left[ \sum_{l=1}^{t-1} (Z_l^{(i)} Z_{l+s}^{(j)} + Z_{l+s}^{(i)} Z_l^{(j)}) + \sum_{l=n-b_n+t+1}^{n-s} (Z_l^{(i)} Z_{l+s}^{(j)} \right. \\
&\quad \left. + Z_{l+s}^{(i)} Z_l^{(j)}) \right] \\
&= \sum_{s=-(b_n-1)}^{b_n-1} \gamma_{n,ij}(s) w_n(s) - d_{n,ij} \\
&= \Sigma_{SV,ij} - d_{n,ij}. \qquad \square
\end{aligned}
$$

Let $\widetilde{\gamma}_n(s)$, $\widetilde{\Sigma}_{SV}, \widetilde{\Sigma}_{w,n}$ and $\tilde{d}_n$ be the Brownian motion analogs of (4.1), (4.2), (B.3), and (B.4). Specifically, for $t = 1, \ldots, n$, define Brownian motion increments $U_t = B(t) - B(t-1)$, so that $U_1, \ldots, U_n$ are $\overset{iid}{\sim} \mathrm{N}_p(0, I_p)$. For $l = 0, \ldots, n - b_n$ and $k = 1, \ldots, b_n$ define $\bar{B}_l(k) = k^{-1}(B(l+k) - B(l))$, $\bar{B}_n = n^{-1}B(n)$, and $T_t = U_t - \bar{B}_n$. Then

$$
\widetilde{\gamma}_n(s) = \frac{1}{n}\sum_{t \in I_s} (U_t - \bar{B}_n)(U_{t+s} - \bar{B}_n)^T = \frac{1}{n}\sum_{t \in I_s} T_t T_{t+s}^T, \tag{B.9}
$$

$$
\widetilde{\Sigma}_{SV} = \sum_{s=-(b_n-1)}^{b_n-1} w_n(s)\widetilde{\gamma}_n(s), \tag{B.10}
$$

$$
\widetilde{\Sigma}_{w,n} = \frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{k=1}^{b_n} k^2 \Delta_2 w_n(k)[\bar{B}_l(k) - \bar{B}_n][\bar{B}_l(k) - \bar{B}_n]^T, \tag{B.11}
$$

$$
\begin{aligned}
\tilde{d}_n = \frac{1}{n}\Bigg\{ &\sum_{t=1}^{b_n} \Delta_1 w_n(t) \left( \sum_{l=1}^{t-1} T_l T_l^T + \sum_{l=n-b_n+t+1}^{n} T_l T_l^T \right) \\
&+ \sum_{s=1}^{b_n-1} \left[ \sum_{t=1}^{b_n-s} \Delta_1 w_n(s+t) \left( \sum_{l=1}^{t-1} (T_l T_{l+s}^T + T_{l+s} T_l^T) \right. \right.
\end{aligned} \tag{B.12}
$$

$$+ \sum_{l=n-b_n+t+1}^{n-s} (T_l T_{l+s}^T + T_{l+s} T_l^T) \Bigg) \Bigg] \Bigg\}. \tag{B.13}$$

Notice that in (B.13) we use the convention that empty sums are zero. Our goal is to show that $\widetilde{\Sigma}_{w,n} \to I_p$ as $n \to \infty$ with probability 1 in the following way. In Lemma B.6 we show that $\widetilde{\Sigma}_{w,n} = \widetilde{\Sigma}_{SV} - \tilde{d}_n$ and in Lemma B.8 we show that the end term $\tilde{d}_n \to 0$ as $n \to \infty$ with probability 1. Lemma B.12 shows that $\widetilde{\Sigma}_{SV} \to I_p$ as $n \to \infty$ with probability 1, and hence $\widetilde{\Sigma}_{w,n} \to I_p$ as $n \to \infty$ with probability 1.

**Lemma B.6** Under Condition 4.1, $\widetilde{\Sigma}_{w,n} = \widetilde{\Sigma}_{SV} - \tilde{d}_n$. □

**Proof**

For $i, j = 1, \ldots, p$, let $\widetilde{\Sigma}_{w,ij}$ denote the $(i,j)$th entry of $\widetilde{\Sigma}_{w,n}$. Then,

$$\begin{aligned}
&\widetilde{\Sigma}_{w,ij} \\
&= \frac{1}{n} \sum_{l=0}^{n-b_n} \sum_{k=1}^{b_n} k^2 \Delta_2 w_n(k) \left[ \bar{B}_l^{(i)}(k) - \bar{B}_n^{(i)} \right] \left[ \bar{B}_l^{(j)}(k) - \bar{B}_n^{(j)} \right] \\
&= \frac{1}{n} \sum_{l=0}^{n-b_n} \sum_{k=1}^{b_n} \Delta_2 w_n(k) \left[ B^{(i)}(k+l) - B^{(i)}(l) - k\bar{B}_n^{(i)} \right] \\
&\quad \times \left[ B^{(j)}(k+l) - B^{(j)}(l) - k\bar{B}_n^{(j)} \right] \\
&= \frac{1}{n} \sum_{l=0}^{n-b_n} \sum_{k=1}^{b_n} \Delta_2 w_n(k) \left[ \sum_{t=1}^{k} U_{t+l}^{(i)} - k\bar{B}_n^{(i)} \right] \left[ \sum_{t=1}^{k} U_{t+l}^{(j)} - k\bar{B}_n^{(j)} \right] \\
&= \frac{1}{n} \sum_{l=0}^{n-b_n} \sum_{k=1}^{b_n} \Delta_2 w_n(k) \left[ \sum_{t=1}^{k} T_{t+l}^{(i)} \right] \left[ \sum_{t=1}^{k} T_{t+l}^{(j)} \right] \\
&= \frac{1}{n} \sum_{l=0}^{n-b_n} \sum_{k=1}^{b_n} \Delta_2 w_n(k) \left[ \sum_{t=1}^{k} T_{l+t}^{(i)} T_{l+t}^{(j)} + \sum_{s=1}^{k-1} \sum_{t=1}^{k-s} T_{l+t}^{(i)} T_{l+t+s}^{(j)} + \sum_{s=1}^{k-1} \sum_{t=1}^{k-s} T_{l+t}^{(j)} T_{l+t+s}^{(i)} \right].
\end{aligned} \tag{B.14}$$

In (B.14), we continue to use convention that empty sums are zero. We will look

at each of the terms in (B.14) separately. For the first term, changing the order of summation and then using Lemma B.4,

$$\begin{aligned}
&\frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{k=1}^{b_n}\sum_{t=1}^{k}\Delta_2 w_n(k)T_{l+t}^{(i)}T_{l+t}^{(j)}\\
&=\frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{t=1}^{b_n}\sum_{k=t}^{b_n}\Delta_2 w_n(k)T_{l+t}^{(i)}T_{l+t}^{(j)}\\
&=\frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{t=1}^{b_n}T_{l+t}^{(i)}T_{l+t}^{(j)}\Delta_1 w_n(t)\\
&=\frac{1}{n}\sum_{t=1}^{b_n}\Delta_1 w_n(t)\sum_{l=0}^{n-b_n}T_{l+t}^{(i)}T_{l+t}^{(j)}\\
&=\widetilde{\gamma}_{n,ij}(0)-\frac{1}{n}\sum_{t=1}^{b_n}\Delta_1 w_n(t)\left(\sum_{l=1}^{t-1}T_l^{(i)}T_l^{(j)}+\sum_{l=n-b_n+t+1}^{n}T_l^{(i)}T_l^{(j)}\right). \qquad \text{(B.15)}
\end{aligned}$$

For the second term in (B.14) we change the order of summation from $l,k,s,t$ to $l,s,k,t$ then to $l,s,t,k$ and use Lemma B.4 to get

$$\begin{aligned}
&\frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{k=1}^{b_n}\sum_{s=1}^{k-1}\sum_{t=1}^{k-s}\Delta_2 w_n(k)T_{l+t}^{(i)}T_{l+t+s}^{(j)}\\
&=\frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{s=1}^{b_n-1}\sum_{k=s+1}^{b_n}\sum_{t=1}^{k-s}\Delta_2 w_n(k)T_{l+t}^{(i)}T_{l+t+s}^{(j)}\\
&=\frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{s=1}^{b_n-1}\sum_{t=1}^{b_n-s}\sum_{k=t+s}^{b_n}\Delta_2 w_n(k)T_{l+t}^{(i)}T_{l+t+s}^{(j)}\\
&=\frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{s=1}^{b_n-1}\sum_{t=1}^{b_n-s}\Delta_1 w_n(s+t)T_{l+t}^{(i)}T_{l+t+s}^{(j)}\\
&=\frac{1}{n}\sum_{s=1}^{b_n-1}\sum_{l=0}^{n-b_n}\sum_{t=1}^{b_n-s}\Delta_1 w_n(s+t)T_{l+t}^{(i)}T_{l+t+s}^{(j)}\\
&=\frac{1}{n}\sum_{s=1}^{b_n-1}\sum_{t=1}^{b_n-s}\sum_{l=0}^{n-b_n}\Delta_1 w_n(s+t)T_{l+t}^{(i)}T_{l+t+s}^{(j)}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{s=1}^{b_n-1}\sum_{t=1}^{b_n-s}\Delta_1 w_n(s+t)\sum_{l=0}^{n-b_n} T_{l+t}^{(i)}T_{l+t+s}^{(j)} \\
&= \frac{1}{n}\sum_{s=1}^{b_n-1}\sum_{t=1}^{b_n-s}\Delta_1 w_n(s+t)\sum_{l=t}^{n-b_n+t} T_{l}^{(i)}T_{l+s}^{(j)} \\
&= \sum_{s=1}^{b_n-1}\sum_{t=1}^{b_n-s}\Delta_1 w_n(s+t)\left[\tilde{\gamma}_{n,ij}(s) - \frac{1}{n}\sum_{l=1}^{t-1} T_{l}^{(i)}T_{l+s}^{(j)} - \frac{1}{n}\sum_{l=n-b_n+t+1}^{n-s} T_{l}^{(i)}T_{l+s}^{(j)}\right] \\
&= \sum_{s=1}^{b_n-1} w_n(s)\tilde{\gamma}_{n,ij}(s) - \frac{1}{n}\sum_{s=1}^{b_n-1}\sum_{t=1}^{b_n-s}\Delta_1 w_n(s+t) \\
&\quad\times\left[\sum_{l=1}^{t-1} T_{l}^{(i)}T_{l+s}^{(j)} + \sum_{l=n-b_n+t+1}^{n-s} T_{l}^{(i)}T_{l+s}^{(j)}\right]. \qquad \text{(B.16)}
\end{aligned}
$$

Repeating the same steps as in the second term, the third term in (B.14) reduces to

$$
\begin{aligned}
&\frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{k=1}^{b_n}\sum_{s=1}^{k-1}\sum_{t=1}^{k-s}\Delta_2 w_n(k) T_{l+t}^{(j)}T_{l+t+s}^{(i)} \\
&= \sum_{s=1}^{b_n-1} w_n(s)\tilde{\gamma}_{n,ji}(s) - \frac{1}{n}\sum_{s=1}^{b_n-1}\sum_{t=1}^{b_n-s}\Delta_1 w_n(s+t) \\
&\quad\times\left[\sum_{l=1}^{t-1} T_{l}^{(j)}T_{l+s}^{(i)} + \sum_{l=n-b_n+t+1}^{n-s} T_{l}^{(j)}T_{l+s}^{(i)}\right] \\
&= \sum_{s=1}^{b_n-1} w_n(-s)\tilde{\gamma}_{n,ij}(-s) - \frac{1}{n}\sum_{s=1}^{b_n-1}\sum_{t=1}^{b_n-s}\Delta_1 w_n(s+t) \\
&\quad\times\left[\sum_{l=1}^{t-1} T_{l}^{(j)}T_{l+s}^{(i)} + \sum_{l=n-b_n+t+1}^{n-s} T_{l}^{(j)}T_{l+s}^{(i)}\right]. \qquad \text{(B.17)}
\end{aligned}
$$

Using (B.15), (B.16), and (B.17) in (B.14), we get

$$
\begin{aligned}
&\widetilde{\Sigma}_{w,ij} \\
&= \tilde{\gamma}_{n,ij}(0) + \sum_{s=1}^{b_n-1} w_n(s)\tilde{\gamma}_{n,ij}(s) + \sum_{s=-(b_n-1)}^{-1} w_n(s)\tilde{\gamma}_{n,ij}(s)
\end{aligned}
$$

$$
\begin{aligned}
&- \frac{1}{n}\sum_{t=1}^{b_n} \Delta_1 w_n(t) \left( \sum_{l=1}^{t-1} T_l^{(i)} T_l^{(j)} + \sum_{l=n-b_n+t+1}^{n} T_l^{(i)} T_l^{(j)} \right) \\
&- \frac{1}{n}\sum_{s=1}^{b_n-1}\sum_{t=1}^{b_n-s} \Delta_1 w_n(s+t) \left[ \sum_{l=1}^{t-1} (T_l^{(i)} T_{l+s}^{(j)} + T_{l+s}^{(i)} T_l^{(j)}) \right. \\
&\qquad \left. + \sum_{l=n-b_n+t+1}^{n-s} (T_l^{(i)} T_{l+s}^{(j)} + T_{l+s}^{(i)} T_l^{(j)}) \right] \\
&= \sum_{s=-(b_n-1)}^{b_n-1} \tilde{\gamma}_{n,ij}(s) w_n(s) - \tilde{d}_{n,ij} \\
&= \widetilde{\Sigma}_{SV,ij} - \tilde{d}_{n,ij}. \qquad \square
\end{aligned}
$$

Next, we show that as $n \to \infty$, $\tilde{d}_n \to 0$ with probability 1 implying $\widetilde{\Sigma}_{w,n} - \widetilde{\Sigma}_{SV} \to 0$ with probability 1 as $n \to \infty$. To do so we require a strong invariance principle for independent and identically distributed random variables.

**Theorem B.1 (Komlós et al. (1975))** Let $B(n)$ be a 1-dimensional standard Brownian motion. If $X_1, X_2, X_3 \ldots$ are independent and identically distributed univariate random variables with mean $\mu$ and standard deviation $\sigma$, such that $\mathrm{E}\left[e^{|tX_1|}\right] < \infty$ in a neighborhood of $t = 0$, then as $n \to \infty$

$$
\sum_{i=1}^{n} X_i - n\mu - \sigma B(n) = O(\log n)\ . \qquad \square
$$

We begin with a technical lemma that will be used in a couple of places in the rest of the proof.

**Lemma B.7** Let Conditions 4.1 and 4.2 hold. If $b_n n^{-1} \sum_{k=1}^{b_n} k|\Delta_1 w_n(k)| \to 0$ as

$n \to \infty$, then

$$\frac{b_n}{n}\left(\sum_{t=1}^{b_n}|\Delta_1 w_n(t)| + 2\sum_{s=1}^{b_n-1}\sum_{t=1}^{b_n-s}|\Delta_1 w_n(s+t)|\right) \to 0\ .$$

□

**Proof**

$$\begin{aligned}
&\frac{b_n}{n}\left(\sum_{t=1}^{b_n}|\Delta_1 w_n(t)| + 2\sum_{s=1}^{b_n-1}\sum_{t=1}^{b_n-s}|\Delta_1 w_n(s+t)|\right)\\
&= \frac{b_n}{n}\left(\sum_{t=1}^{b_n}|\Delta_1 w_n(t)| + 2\sum_{s=1}^{b_n-1}\sum_{k=s+1}^{b_n}|\Delta_1 w_n(k)|\right)\\
&= \frac{b_n}{n}\left(\sum_{t=1}^{b_n}|\Delta_1 w_n(t)| + 2\sum_{k=2}^{b_n}\sum_{s=1}^{k-1}|\Delta_1 w_n(k)|\right)\\
&= \frac{b_n}{n}\left(\sum_{t=1}^{b_n}|\Delta_1 w_n(t)| + 2\sum_{k=2}^{b_n}(k-1)|\Delta_1 w_n(k)|\right)\\
&= \frac{b_n}{n}\left(\sum_{t=1}^{b_n}|\Delta_1 w_n(t)| + 2\sum_{k=1}^{b_n}(k-1)|\Delta_1 w_n(k)|\right)\\
&\le \frac{b_n}{n}\left(2\sum_{k=1}^{b_n}k|\Delta_1 w_n(k)|\right)\\
&\to 0 \text{ by assumption.}
\end{aligned}$$

□

**Lemma B.8** Let Conditions 4.1 and 4.2 hold and let $n > 2b_n$. If as $n \to \infty$ $b_n n^{-1}\sum_{k=1}^{b_n} k|\Delta_1 w_n(k)| \to 0$ and $b_n^{-1}\log n = O(1)$, then $\tilde{d}_n \to 0$ with probability 1 as $n \to \infty$. □

**Proof**

For $i, j = 1, \ldots, p$, we will show that as $n \to \infty$ with probability 1, $\tilde{d}_{n,ij} \to 0$. Recall

$$
\begin{aligned}
&\tilde{d}_{n,ij} \\
&= \frac{1}{n} \sum_{t=1}^{b_n} \Delta_1 w_n(t) \left( \sum_{l=1}^{t-1} T_l^{(i)} T_l^{(j)} + \sum_{l=n-b_n+t+1}^{n} T_l^{(i)} T_l^{(j)} \right) \\
&\quad + \frac{1}{n} \sum_{s=1}^{b_n-1} \sum_{t=1}^{b_n-s} \Delta_1 w_n(s+t) \left[ \sum_{l=1}^{t-1} (T_l^{(i)} T_{l+s}^{(j)} + T_{l+s}^{(i)} T_l^{(j)}) \right. \\
&\quad \left. + \sum_{l=n-b_n+t+1}^{n-s} (T_l^{(i)} T_{l+s}^{(j)} + T_{l+s}^{(i)} T_l^{(j)}) \right] \\
&|\tilde{d}_{n,ij}| \\
&\leq \frac{1}{n} \sum_{t=1}^{b_n} |\Delta_1 w_n(t)| \left( \sum_{l=1}^{t-1} \left| T_l^{(i)} T_l^{(j)} \right| + \sum_{l=n-b_n+t+1}^{n} \left| T_l^{(i)} T_l^{(j)} \right| \right) \\
&\quad + \frac{1}{n} \sum_{s=1}^{b_n-1} \sum_{t=1}^{b_n-s} |\Delta_1 w_n(s+t)| \times \\
&\quad \times \left[ \sum_{l=1}^{t-1} \left( \left| T_l^{(i)} T_{l+s}^{(j)} \right| + \left| T_{l+s}^{(i)} T_l^{(j)} \right| \right) + \sum_{l=n-b_n+t+1}^{n-s} \left( \left| T_l^{(i)} T_{l+s}^{(j)} \right| + \left| T_{l+s}^{(i)} T_l^{(j)} \right| \right) \right],
\end{aligned}
\tag{B.18}
$$

where we use the convention that empty sums are zero. Using the inequality $|ab| \leq (a^2 + b^2)/2$ in the first and second terms in (B.18), we have for $t = 1, \ldots, b_n$

$$
\sum_{l=1}^{t-1} \left| T_l^{(i)} T_l^{(j)} \right| \leq \frac{1}{2} \sum_{l=1}^{t-1} \left( T_l^{(i)2} + T_l^{(j)2} \right) \leq \frac{1}{2} \sum_{l=1}^{2b_n} T_l^{(i)2} + \frac{1}{2} \sum_{l=1}^{2b_n} T_l^{(j)2}
$$

$$
\begin{aligned}
\sum_{l=n-b_n+t+1}^{n} \left| T_l^{(i)} T_l^{(j)} \right| &\leq \frac{1}{2} \sum_{l=n-b_n+t+1}^{n} \left( T_l^{(i)2} + T_l^{(j)2} \right) \\
&\leq \frac{1}{2} \sum_{l=n-2b_n+1}^{n} T_l^{(i)2} + \frac{1}{2} \sum_{l=n-2b_n+1}^{n} T_l^{(j)2}.
\end{aligned}
$$

Similarly, for the third and fourth terms in (B.18), for $t = 1, \dots b_n - 1$ and $s = 1, \dots, b_n - 1$

$$\begin{aligned}
&\sum_{l=1}^{t-1} \left| T_l^{(i)} T_{l+s}^{(j)} \right| + \sum_{l=1}^{t-1} \left| T_{l+s}^{(i)} T_l^{(j)} \right| \\
&\leq \frac{1}{2} \sum_{l=1}^{2b_n} T_l^{(i)2} + \frac{1}{2} \sum_{l=1}^{2b_n} T_l^{(j)2} + \frac{1}{2} \sum_{l=1}^{b_n} T_{l+s}^{(j)2} + \frac{1}{2} \sum_{l=1}^{b_n} T_{l+s}^{(i)2} \\
&\leq \frac{1}{2} \sum_{l=1}^{2b_n} T_l^{(i)2} + \frac{1}{2} \sum_{l=1}^{2b_n} T_l^{(j)2} + \frac{1}{2} \sum_{l=1}^{2b_n} T_l^{(j)2} + \frac{1}{2} \sum_{l=1}^{2b_n} T_l^{(i)2} \\
&= \sum_{l=1}^{2b_n} T_l^{(i)2} + \sum_{l=1}^{2b_n} T_l^{(j)2}.
\end{aligned}$$

$$\begin{aligned}
&\sum_{l=n-b_n+t+1}^{n} \left| T_l^{(i)} T_{l+s}^{(j)} \right| + \sum_{l=n-b_n+t+1}^{n} \left| T_{l+s}^{(i)} T_l^{(j)} \right| \\
&\leq \frac{1}{2} \sum_{l=n-2b_n+1}^{n} (T_l^{(i)2} + T_l^{(j)2}) + \frac{1}{2} \sum_{l=n-b_n+1}^{n} (T_{l+s}^{(j)2} + T_{l+s}^{(i)2}) \\
&\leq \frac{1}{2} \sum_{l=n-2b_n+1}^{n} (T_l^{(i)2} + T_l^{(j)2}) + \frac{1}{2} \sum_{l=n-2b_n+1}^{n} (T_l^{(j)2} + T_l^{(i)2}) \\
&= \sum_{l=n-2b_n+1}^{n} T_l^{(i)2} + \sum_{l=n-2b_n+1}^{n} T_l^{(j)2}.
\end{aligned}$$

Combining the above results in (B.18) we get,

$$\begin{aligned}
&|\tilde{d}_{n,ij}| \\
&\leq \frac{1}{n} \left( \frac{1}{2} \sum_{l=1}^{2b_n} T_l^{(i)2} + \frac{1}{2} \sum_{l=1}^{2b_n} T_l^{(j)2} + \frac{1}{2} \sum_{l=n-2b_n+1}^{n} T_l^{(i)2} + \frac{1}{2} \sum_{l=n-2b_n+1}^{n} T_l^{(j)2} \right) \\
&\quad \times \sum_{t=1}^{b_n} |\Delta_1 w_n(t)| + \frac{1}{n} \sum_{s=1}^{b_n-1} \left[ \left( \sum_{l=1}^{2b_n} T_l^{(i)2} + \sum_{l=1}^{2b_n} T_l^{(j)2} \right. \right.
\end{aligned}$$

$$+\sum_{l=n-2b_n+1}^{n} T_l^{(i)2}+\sum_{l=n-2b_n+1}^{n} T_l^{(j)2}\Bigg)\sum_{t=1}^{b_n-s}|\Delta_1 w_n(s+t)|\Bigg]$$
$$=\frac{1}{b_n}\left(\frac{1}{2}\sum_{l=1}^{2b_n} T_l^{(i)2}+\frac{1}{2}\sum_{l=1}^{2b_n} T_l^{(j)2}+\frac{1}{2}\sum_{l=n-2b_n+1}^{n} T_l^{(i)2}+\frac{1}{2}\sum_{l=n-2b_n+1}^{n} T_l^{(j)2}\right)$$
$$\times\frac{b_n}{n}\left(\sum_{t=1}^{b_n}|\Delta_1 w_n(t)|+2\sum_{s=1}^{b_n-1}\sum_{t=1}^{b_n-s}|\Delta_1 w_n(s+t)|\right). \tag{B.19}$$

We will show that the first term in the product in (B.19) remains bounded with probability 1 as $n\to\infty$. Consider,

$$\begin{aligned}
\frac{1}{2b_n}\sum_{l=1}^{2b_n} T_l^{(i)2} &= \frac{1}{2b_n}\sum_{l=1}^{2b_n}\left(U_l^{(i)}-\bar{B}_n^{(i)}\right)^2\\
&= \frac{1}{2b_n}\sum_{l=1}^{2b_n} U_l^{(i)2}-2\bar{B}_n^{(i)}\frac{1}{2b_n}\sum_{l=1}^{2b_n} U_l^{(i)}+\left(\bar{B}_n^{(i)}\right)^2\\
&\le \left|\frac{1}{2b_n}\sum_{l=1}^{2b_n} U_l^{(i)2}\right|+2\left|\bar{B}_n^{(i)}\right|\left|\frac{1}{2b_n}\sum_{l=1}^{2b_n} U_l^{(i)}\right|+\left|\bar{B}_n^{(i)}\right|^2\\
&< \left|\frac{1}{2b_n}\sum_{l=1}^{2b_n} U_l^{(i)2}\right|+\left|\frac{1}{2b_n}\sum_{l=1}^{2b_n} U_l^{(i)}\right|\left(\frac{2}{n}(1+\epsilon)(2n\log\log n)^{1/2}\right)\\
&\quad+\left(\frac{1}{n}(1+\epsilon)(2n\log\log n)^{1/2}\right)^2 \qquad \text{by Lemma B.1}\\
&< \left|\frac{1}{2b_n}\sum_{l=1}^{2b_n} U_l^{(i)2}\right|+\left|\frac{1}{2b_n}\sum_{l=1}^{2b_n} U_l^{(i)}\right| O((n^{-1}\log n)^{1/2})+O(n^{-1}\log n)\,.
\end{aligned}$$

Since $U_l^{(i)}$ are Brownian motion increments, $U_l^{(i)}\overset{iid}{\sim}N(0,1)$ and by the classical strong law of large numbers, the above remains bounded with probability 1. Similarly $(2b_n)^{-1}\sum_{l=1}^{2b_n} T_l^{(j)2}$ remains bounded with probability 1 as $n\to\infty$. Next, consider $R_n=\sum_{l=1}^{n} U_l^{(i)2}$. Since $U_l^{(i)}\sim N(0,1)$, $R_n\sim\chi_n^2$. Thus $R_n$ has a moment generating function and an application of Theorem B.1 implies there exists a finite random

variable $C_R$ such that, for sufficiently large $n$,

$$\left|R_n - n - 2B^{(i)}(n)\right| < C_R \log n \ . \tag{B.20}$$

Consider

$$\begin{aligned}
&|R_n - R_{n-2b_n}| \\
&= \Big| \left(R_n - n - 2B^{(i)}(n)\right) - \left(R_{n-2b_n} - (n-2b_n) - 2B^{(i)}(n-2b_n)\right) \\
&\quad - (n-2b_n) + n + 2B^{(i)}(n) - 2B^{(i)}(n-2b_n)\Big| \\
&\le \left|\left(R_n - n - 2B^{(i)}(n)\right)\right| + \left|\left(R_{n-2b_n} - (n-2b_n) - 2B^{(i)}(n-2b_n)\right)\right| \\
&\quad + \left|2b_n + 2B^{(i)}(n) - 2B^{(i)}(n-2b_n)\right| \\
&< C_R \log n + C_R \log (n-b_n) + 2b_n \\
&\quad + 2(1+\epsilon)\left(2(2b_n)\left(\log \frac{n}{2b_n} + \log\log n\right)\right)^{1/2} \quad \text{by (B.20) and Lemma B.1} \\
&< 2C_R \log n + 2b_n + 4(1+\epsilon)(2b_n \log n)^{1/2}.
\end{aligned} \tag{B.21}$$

Finally,

$$\begin{aligned}
&\frac{1}{2b_n} \sum_{l=n-2b_n+1}^{n} T_l^{(i)2} \\
&= \frac{1}{2b_n} \sum_{l=n-2b_n+1}^{n} \left(U_l^{(i)} - \bar{B}_n^{(i)}\right)^2 \\
&= \frac{1}{2b_n} \sum_{l=n-2b_n+1}^{n} U_l^{(i)2} - \frac{2\bar{B}_n^{(i)}}{2b_n} \sum_{l=n-2b_n+1}^{n} U_l^{(i)} + \left(\bar{B}_n^{(i)}\right)^2 \\
&= \frac{1}{2b_n}(R_n - R_{n-2b_n}) - \frac{2}{n}B^{(i)}(n)\frac{1}{2b_n}\left(B^{(i)}(n) - B^{(i)}(n-2b_n)\right) + \left(\frac{1}{n}B^{(i)}(n)\right)^2 \\
&< \frac{1}{2b_n}|R_n - R_{n-2b_n}| + \frac{2}{n}\left|B^{(i)}(n)\right| \frac{1}{2b_n}\left|B^{(i)}(n) - B^{(i)}(n-2b_n)\right|
\end{aligned}$$

$$
\begin{aligned}
&+ \left(\frac{1}{n}\left|B^{(i)}(n)\right|\right)^2 \\
&< \frac{1}{2b_n}\left(2C_R \log n + 2b_n + 4(1+\epsilon)(2b_n \log n)^{1/2}\right) \\
&+ \frac{2}{n}(1+\epsilon)(2n\log\log n)^{1/2}\frac{1}{2b_n}(1+\epsilon)\left(2(2b_n)\left(\log\frac{n}{2b_n} + \log\log n\right)\right)^{1/2} \\
&+ \left(\frac{1}{n}(1+\epsilon)(2n\log\log n)^{1/2}\right)^2 \qquad \text{by (B.21) and Lemma B.1} \\
&< C_R b_n^{-1}\log n + 1 + \frac{4(1+\epsilon)(2b_n\log n)^{1/2}}{2b_n} \\
&+ \frac{1}{nb_n}(1+\epsilon)^2(2n\log n)^{1/2}(8b_n\log n)^{1/2} + \frac{(1+\epsilon)^2}{n}(2\log n) \\
&< C_R b_n^{-1}\log n + 1 + 2(1+\epsilon)(2b_n^{-1}\log n)^{1/2} \\
&+ 4(1+\epsilon)^2\left(\frac{\log n}{n}\right)^{1/2}\left(b_n^{-1}\log n\right)^{1/2} + 2(1+\epsilon)^2\frac{\log n}{n}.
\end{aligned}
$$

Since $b_n^{-1}\log n = O(1)$ as $n \to \infty$, the above term remains bounded with probability 1 as $n \to \infty$. Similarly, $(2b_n)^{-1}\sum_{l=n-2b_n+1}^{n} T_l^{(j)2}$ remains bounded with probability 1 as $n \to \infty$. The second term in the product in (B.19) converges to 0 by Lemma B.7 and hence $\tilde{d}_{n,ij} \to 0$ with probability 1 as $n \to \infty$. □

Recall that $h(X_t) = Y_t^2$ for $t = 1, 2, 3, \ldots$, where the square is element-wise.

**Lemma B.9** Let a strong invariance principle for $h$ hold as in (4.3). If Condition 4.2 holds, $b_n^{-1}\psi_h(n) \to 0$ and $b_n^{-1}\log n = O(1)$ as $n \to \infty$, then

$$\frac{1}{b_n}\sum_{k=1}^{b_n} h(X_k) \text{ and } \frac{1}{b_n}\sum_{k=n-b_n+1}^{n} h(X_k),$$

stay bounded with probability 1 as $n \to \infty$. □

**Proof**

Equation (4.3) implies that $b_n^{-1}\sum_{k=1}^{b_n} h(X_k) \to \mathrm{E}_F h$ if $b_n^{-1}\psi_h(b_n) \to 0$ as $n \to \infty$. Since by assumption $b_n^{-1}\psi_h(n) \to 0$ as $n \to \infty$ and $\psi_h$ is increasing, $b_n^{-1}\sum_{k=1}^{b_n} h(X_k)$

remains bounded w.p. 1 as $n \to \infty$. Next, for all $\epsilon > 0$ and sufficiently large $n(\epsilon)$,

$$
\begin{aligned}
&\frac{1}{b_n}\left\|\sum_{k=n-b_n+1}^{n} h(X_k)\right\| \\
=&\frac{1}{b_n}\left\|\sum_{k=1}^{n} h(X_k) - \sum_{k=1}^{n-b_n} h(X_k)\right\| \\
=&\frac{1}{b_n}\Bigg\|\sum_{k=1}^{n} h(X_k) - n\mathrm{E}_F h + (n-b_n)\mathrm{E}_F h + b_n \mathrm{E}_F h - L_h B(n) + L_h B(n-b_n) \\
&+ L_h(B(n) - B(n-b_n)) - \sum_{k=1}^{n-b_n} h(X_k)\Bigg\| \\
\le&\frac{1}{b_n}\left\|\sum_{k=1}^{n} h(X_k) - n\mathrm{E}_F h - L_h B(n)\right\| \\
&+ \frac{1}{b_n}\left\|\sum_{k=1}^{n-b_n} h(X_k) - (n-b_n)\mathrm{E}_F h - L_h B(n-b_n)\right\| \\
&+ \frac{1}{b_n}\left\|L_h(B(n) - B(n-b_n)) + b_n \mathrm{E}_F h\right\| \\
<&\frac{1}{b_n} D_h \psi_h(n) + \frac{1}{b_n} D_h \psi_h(n-b_n) \\
&+ \frac{1}{b_n}\left\|L_h(B(n) - B(n-b_n))\right\| + \left\|\mathrm{E}_F h\right\| \quad \text{by (4.3)} \\
\le&\frac{1}{b_n} D_h \psi_h(n) + \frac{1}{b_n} D_h \psi_h(n-b_n) \\
&+ \frac{1}{b_n}\|L_h\| \left(\sum_{i=1}^{p} \left|B^{(i)}(n) - B^{(i)}(n-b_n)\right|^2\right)^{1/2} + \left\|\mathrm{E}_F h\right\| \\
\le&\frac{1}{b_n} D_h \psi_h(n) + \frac{1}{b_n} D_h \psi_h(n-b_n) \\
&+ \frac{1}{b_n}\|L_h\| \left(\sum_{i=1}^{p} \sup_{0 \le s \le b_n} \left|B^{(i)}(n) - B^{(i)}(n-s)\right|^2\right)^{1/2} + \left\|\mathrm{E}_F h\right\|
\end{aligned}
$$

by Lemma B.1

$$
\begin{aligned}
&< \frac{2}{b_n} D_h \psi_h(n) + \frac{p^{1/2}}{b_n} \|L_h\| (1+\epsilon) \left( 2b_n \left( \log \frac{n}{b_n} + \log\log n \right) \right)^{1/2} + \|\mathrm{E}_F h\| \\
&< \|\mathrm{E}_F h\| + \frac{2}{b_n} D_h \psi_h(n) + O((b_n^{-1} \log n)^{1/2}).
\end{aligned}
$$

Thus by the assumptions $b_n^{-1} \left\| \sum_{k=n-b_n+1}^{n} h(X_k) \right\|$ stays bounded w.p. 1 as $n \to \infty$.□

**Lemma B.10** Suppose the strong invariance principles (2.4) and (4.3) hold. In addition, suppose Conditions 4.1 and 4.2 hold and $n > 2b_n$, $b_n n^{-1} \sum_{k=1}^{b_n} k|\Delta_1 w_n(k)| \to 0$, $b_n^{-1}\psi(n) \to 0$, $b_n^{-1}\psi_h(n) \to 0$ as $n \to \infty$ and $b_n^{-1} \log n = O(1)$. Then, $d_n \to 0$ with probability 1 as $n \to \infty$. □

**Proof**

For $i, j = 1, \ldots, p$, let $d_{n,ij}$ denote the $(i,j)$th element of the matrix $d_n$. We can follow the same steps as in Lemma B.8 to obtain

$$
\begin{aligned}
&|d_{n,ij}| \\
&\leq \frac{1}{b_n} \left( \frac{1}{2} \sum_{l=1}^{2b_n} Z_l^{(i)2} + \frac{1}{2} \sum_{l=1}^{2b_n} Z_l^{(j)2} + \frac{1}{2} \sum_{l=n-2b_n+1}^{n} Z_l^{(i)2} + \frac{1}{2} \sum_{l=n-2b_n+1}^{n} Z_l^{(j)2} \right) \\
&\times \frac{b_n}{n} \left( \sum_{t=1}^{b_n} |\Delta_1 w_n(t)| + 2 \sum_{s=1}^{b_n-1} \sum_{t=1}^{b_n-s} |\Delta_1 w_n(s+t)| \right).
\end{aligned}
$$

The second term in the product converges to 0 by Lemma B.7. It remains to show that the following remains bounded with probability 1 as $n \to \infty$,

$$
\frac{1}{b_n} \left( \frac{1}{2} \sum_{l=1}^{2b_n} Z_l^{(i)2} + \frac{1}{2} \sum_{l=1}^{2b_n} Z_l^{(j)2} + \frac{1}{2} \sum_{l=n-2b_n+1}^{n} Z_l^{(i)2} + \frac{1}{2} \sum_{l=n-2b_n+1}^{n} Z_l^{(j)2} \right).
$$

We have,

$$\frac{1}{2b_n}\sum_{l=1}^{2b_n} Z_l^{(i)2} = \frac{1}{2b_n}\sum_{l=1}^{2b_n}\left(Y_l^{(i)} - \bar{Y}_n^{(i)}\right)^2 = \frac{1}{2b_n}\sum_{l=1}^{2b_n} Y_l^{(i)2} - 2\bar{Y}_{2b_n}^{(i)}\bar{Y}_n^{(i)} + \left(\bar{Y}_n^{(i)}\right)^2.$$

By the strong invariance principle for $g$, $\bar{Y}_n^{(i)} \to 0$, $\bar{Y}_{2b_n}^{(i)} \to 0$, and $(\bar{Y}_n^{(i)})^2 \to 0$ w.p. 1 as $n \to \infty$. By Lemma B.9, $(2b_n)^{-1}\sum_{l=1}^{2b_n} Y_l^{(i)2}$ remains bounded w.p. 1 as $n \to \infty$. Thus $(2b_n)^{-1}\sum_{l=1}^{2b_n} Z_l^{(i)2}$ remains bounded w.p. 1 as $n \to \infty$. Similarly $(2b_n)^{-1}\sum_{l=1}^{2b_n} Z_l^{(j)2}$ stay bounded w.p. 1 as $n \to \infty$. Now consider

$$\begin{aligned}\frac{1}{2b_n}\sum_{l=n-2b_n+1}^{n} Z_l^{(i)2} &= \frac{1}{2b_n}\sum_{l=n-2b_n+1}^{n}\left(Y_l^{(i)} - \bar{Y}_n^{(i)}\right)^2 \\ &= \frac{1}{2b_n}\sum_{l=n-2b_n+1}^{n} Y_l^{(i)2} - 2\bar{Y}_n^{(i)}\frac{1}{2b_n}\sum_{l=n-2b_n+1}^{n} Y_l^{(i)} + \left(\bar{Y}_n^{(i)}\right)^2.\end{aligned} \tag{B.22}$$

We will first show that $(2b_n)^{-1}\sum_{l=n-2b_n+1}^{n} Y_l^{(i)}$ remains bounded with probability 1. Let $\Sigma_{ii}$ denote the $i$th diagonal entry of $\Sigma$, then

$$\begin{aligned}&\frac{1}{2b_n}\sum_{l=n-2b_n+1}^{n} Y_l^{(i)} \\ &= \frac{1}{2b_n}\left(\sum_{l=1}^{n} Y_l^{(i)} - \sum_{l=1}^{n-2b_n} Y_l^{(i)}\right) \\ &= \frac{1}{2b_n}\left(\sum_{l=1}^{n} Y_l^{(i)} - \sqrt{\Sigma_{ii}}B^{(i)}(n)\right) - \frac{1}{2b_n}\left(\sum_{l=1}^{n-2b_n} Y_l^{(i)} - \sqrt{\Sigma_{ii}}B^{(i)}(n-2b_n)\right) \\ &\quad + \frac{1}{2b_n}\sqrt{\Sigma_{ii}}\left(B^{(i)}(n) - B^{(i)}(n-2b_n)\right) \\ &< \frac{1}{2b_n}\left|\sum_{l=1}^{n} Y_l^{(i)} - \sqrt{\Sigma_{ii}}B^{(i)}(n)\right| + \frac{1}{2b_n}\left|\sum_{l=1}^{n-2b_n} Y_l^{(i)} - \sqrt{\Sigma_{ii}}B^{(i)}(n-2b_n)\right| \\ &\quad + \frac{1}{2b_n}\left|\sqrt{\Sigma_{ii}}\left(B^{(i)}(n) - B^{(i)}(n-2b_n)\right)\right|\end{aligned}$$

$$
\begin{aligned}
&< \frac{1}{2b_n} D\psi(n) + \frac{1}{2b_n} D\psi(n-2b_n) \\
&\quad + \frac{1}{2b_n}\sqrt{\Sigma_{ii}} \sup_{0\le s\le 2b_n} |B^{(i)}(n) - B^{(i)}(n-s)| \quad \text{by (2.4)} \\
&< \frac{2D}{2b_n}\psi(n) + \sqrt{\Sigma_{ii}}\frac{1}{2b_n}(1+\epsilon)\left[2(2b_n)\left(\log\frac{n}{2b_n} + \log\log n\right)\right]^{1/2} \text{by Lemma B.1} \\
&< Db_n^{-1}\psi(n) + \sqrt{\Sigma_{ii}}(1+\epsilon)(2b_n^{-1}\log n)^{1/2} \\
&= O(b_n^{-1}\psi(n)) + O((b_n^{-1}\log n)^{1/2}) \,.
\end{aligned}
$$

By the strong invariance principle for $g$, $\bar{Y}_n^{(i)} \to 0$ and $(\bar{Y}_n^{(i)})^2 \to 0$ w.p. 1 as $n \to \infty$. By Lemma B.9, $(2b_n)^{-1}\sum_{l=n-2b_n+1}^{n} Y_l^{(i)2}$ remains bounded w.p. 1 as $n \to \infty$. Combining these results in (B.22), $(2b_n)^{-1}\sum_{l=n-2b_n+1}^{n} Z_l^{(i)2}$ remains bounded w.p. 1 as $n \to \infty$. Similarly $(2b_n)^{-1}\sum_{l=n-2b_n+1}^{n} Z_l^{(j)2}$ remains bounded w.p. 1 as $n \to \infty$. □

**Lemma B.11** (Whittle, 1960) Let $R_1, \ldots, R_n$ be i.i.d standard normal variables and $A = \sum_{l=1}^{n}\sum_{k=1}^{n} a_{lk}R_lR_k$ where $a_{lk}$ are real coefficients, then for $c \ge 1$ and for some constant $K_c$, we have

$$
\mathrm{E}[|A - \mathrm{E}A|^{2c}] \le K_c\left(\sum_l\sum_k a_{lk}^2\right)^c \,.
$$

□

**Lemma B.12** Let Conditions 4.1 and 4.2 hold and assume that

(a) there exists a constant $c \ge 1$ such that $\sum_n (b_n/n)^c < \infty$,

(b) $b_n n^{-1}\log n \to 0$ as $n \to \infty$,

then $\widetilde{\Sigma}_{SV} \to I_p$ w.p. 1 as $n \to \infty$. □

**Proof**

Under the same conditions, Theorem 4.1 in Damerdji (1991) shows $\widetilde{\Sigma}_{SV,ii} \to 1$ as $n \to \infty$ w.p. 1. It is left to show that for all $i, j = 1, \ldots, p$, and $i \ne j$, $\widetilde{\Sigma}_{SV,ij} \to$

0 w.p. 1 as $n \to \infty$. Recall that

$$
\begin{aligned}
&\widetilde{\Sigma}_{SV,ij} \\
&= \sum_{s=-(b_n-1)}^{b_n-1} w_n(s)\tilde{\gamma}_{n,ij}(s) \\
&= \tilde{\gamma}_{n,ij}(0) + \frac{1}{n}\left[\sum_{s=1}^{b_n-1} w_n(s)\sum_{t=1}^{n-s}(U_t^{(i)} - \bar{B}_n^{(i)})(U_{t+s}^{(j)} - \bar{B}_n^{(j)})\right. \\
&\quad \left. + \sum_{s=-(b_n-1)}^{-1} w_n(s)\sum_{t=1-s}^{n}(U_t^{(i)} - \bar{B}_n^{(i)})(U_{t+s}^{(j)} - \bar{B}_n^{(j)})\right] \\
&= \tilde{\gamma}_{n,ij}(0) + \frac{1}{n}\left[\sum_{s=1}^{b_n-1} w_n(s)\sum_{t=1}^{n-s}(U_t^{(i)} - \bar{B}_n^{(i)})(U_{t+s}^{(j)} - \bar{B}_n^{(j)})\right. \\
&\quad \left. + \sum_{s=1}^{b_n-1} w_n(s)\sum_{t=1+s}^{n}(U_t^{(i)} - \bar{B}_n^{(i)})(U_{t-s}^{(j)} - \bar{B}_n^{(j)})\right] \\
&= \tilde{\gamma}_{n,ij}(0) + \sum_{s=1}^{b_n-1} w_n(s)\frac{1}{n}\left[\sum_{t=1}^{n-s}\left(U_t^{(i)}U_{t+s}^{(j)} - \bar{B}_n^{(i)}U_{t+s}^{(j)} - \bar{B}_n^{(j)}U_t^{(i)} + \bar{B}_n^{(i)}\bar{B}_n^{(j)}\right)\right. \\
&\quad \left. + \sum_{t=1+s}^{n}\left(U_t^{(i)}U_{t-s}^{(j)} - \bar{B}_n^{(i)}U_{t-s}^{(j)} - \bar{B}_n^{(j)}U_t^{(i)} + \bar{B}_n^{(i)}\bar{B}_n^{(j)}\right)\right].
\end{aligned}
$$

Since

$$
\sum_{t=1}^{n-s} U_{t+s}^{(j)} = B^{(j)}(n) - B^{(j)}(s), \qquad \sum_{t=1}^{n-s} U_t^{(i)} = B^{(i)}(n-s),
$$
$$
\sum_{t=1+s}^{n} U_{t-s}^{(j)} = B^{(j)}(n-s) \qquad \text{and} \qquad \sum_{t=1+s}^{n} U_t^{(i)} = B^{(i)}(n) - B^{(i)}(s),
$$

we get

$$
\begin{aligned}
&\widetilde{\Sigma}_{SV,ij} \\
&= \tilde{\gamma}_{n,ij}(0) + \sum_{s=1}^{b_n-1} w_n(s)\left[\frac{1}{n}\sum_{t=1}^{n-s} U_t^{(i)}U_{t+s}^{(j)} - \frac{1}{n}\bar{B}_n^{(i)}(B^{(j)}(n) - B^{(j)}(s)\right.
\end{aligned}
$$

$$
\begin{aligned}
&- \frac{1}{n}\bar{B}_n^{(j)}B^{(i)}(n-s) + \left(\frac{n-s}{n}\right)\bar{B}_n^{(i)}\bar{B}_n^{(j)} + \frac{1}{n}\sum_{t=1+s}^{n} U_t^{(i)}U_{t-s}^{(j)} \\
&- \frac{1}{n}\bar{B}_n^{(i)}B^{(j)}(n-s) - \frac{1}{n}\bar{B}_n^{(j)}(B^{(i)}(n) - B^{(i)}(s)) + \left(\frac{n-s}{n}\right)\bar{B}_n^{(i)}\bar{B}_n^{(j)}\Bigg] \\
&= \tilde{\gamma}_{n,ij}(0) + \sum_{s=1}^{b_n-1} w_n(s)\Bigg[\frac{1}{n}\sum_{t=1}^{n-s} U_t^{(i)}U_{t+s}^{(j)} + \frac{1}{n}\sum_{t=1+s}^{n} U_t^{(i)}U_{t-s}^{(j)} + 2\left(1-\frac{s}{n}\right)\bar{B}_n^{(i)}\bar{B}_n^{(j)} \\
&- 2\bar{B}_n^{(i)}\bar{B}_n^{(j)} + \frac{1}{n}\bar{B}_n^{(i)}B^{(j)}(s) - \frac{1}{n}\bar{B}_n^{(j)}B^{(i)}(n-s) \\
&- \frac{1}{n}\bar{B}_n^{(i)}B^{(j)}(n-s) + \frac{1}{n}\bar{B}_n^{(j)}B^{(i)}(s)\Bigg] \\
&= \tilde{\gamma}_{n,ij}(0) + \sum_{s=1}^{b_n-1} w_n(s)\Bigg[\frac{1}{n}\sum_{t=1}^{n-s} U_t^{(i)}U_{t+s}^{(j)} + \frac{1}{n}\sum_{t=1+s}^{n} U_t^{(i)}U_{t-s}^{(j)} - 2\left(1+\frac{s}{n}\right)\bar{B}_n^{(i)}\bar{B}_n^{(j)} \\
&+ \bar{B}_n^{(i)}\bar{B}_n^{(j)} - \frac{1}{n}\bar{B}_n^{(j)}B^{(i)}(n-s) + \bar{B}_n^{(i)}\bar{B}_n^{(j)} - \frac{1}{n}\bar{B}_n^{(i)}B^{(j)}(n-s) \\
&+ \frac{1}{n}\bar{B}_n^{(i)}B^{(j)}(s) + \frac{1}{n}\bar{B}_n^{(j)}B^{(i)}(s)\Bigg] \\
&= \tilde{\gamma}_{n,ij}(0) + \sum_{s=1}^{b_n-1} w_n(s)\Bigg[\frac{1}{n}\sum_{t=1}^{n-s} U_t^{(i)}U_{t+s}^{(j)} + \frac{1}{n}\sum_{t=1+s}^{n} U_t^{(i)}U_{t-s}^{(j)} - 2\left(1+\frac{s}{n}\right)\bar{B}_n^{(i)}\bar{B}_n^{(j)} \\
&+ \frac{1}{n}\bar{B}_n^{(j)}(B^{(i)}(n) - B^{(i)}(n-s)) + \frac{1}{n}\bar{B}_n^{(i)}(B^{(j)}(n) - B^{(j)}(n-s)) \\
&+ \frac{1}{n}\bar{B}_n^{(i)}B^{(j)}(s) + \frac{1}{n}B^{(i)}(s)\bar{B}_n^{(j)}\Bigg]. \qquad \text{(B.23)}
\end{aligned}
$$

We will show that each of the terms goes to 0 with probability 1 as $n \to \infty$.

1.

$$
\begin{aligned}
\tilde{\gamma}_{n,ij}(0) &= \frac{1}{n}\sum_{t=1}^{n} T_t^{(i)}T_t^{(j)} \\
&= \frac{1}{n}\sum_{t=1}^{n}\left(U_t^{(i)} - \bar{B}_n^{(i)}\right)\left(U_t^{(j)} - \bar{B}_n^{(j)}\right) \\
&= \frac{1}{n}\sum_{t=1}^{n} U_t^{(i)}U_t^{(j)} - \bar{B}_n^{(j)}\frac{1}{n}\sum_{t=1}^{n} U_t^{(i)} - \bar{B}_n^{(i)}\frac{1}{n}\sum_{t=1}^{n} U_t^{(j)} + \bar{B}_n^{(i)}\bar{B}_n^{(j)}. \qquad \text{(B.24)}
\end{aligned}
$$

We will show that each of the terms in (B.24) goes to 0 with probability 1, as $n \to \infty$. First, we will use Lemma B.11 to show that $n^{-1}\sum_{t=1}^{n} U_t^{(i)}U_t^{(j)} \to 0$ with probability 1 as $n \to \infty$. Define

$$R_1 = U_1^{(i)},\ R_2 = U_2^{(i)}, \ldots,\ R_n = U_n^{(i)},\ R_{n+1} = U_1^{(j)}, \ldots,\ R_{2n} = U_n^{(j)}.$$

Thus, $\{R_i : 1 \le i \le 2n\}$ is an i.i.d sequence of normally distributed random variables. Define for $1 \le l, k, \le 2n$,

$$a_{lk} = \begin{cases} \dfrac{1}{n}, & \text{if } 1 \le l \le n \text{ and } k = l + n \\ 0 & \text{otherwise .} \end{cases}$$

Then,

$$A := \sum_{l=1}^{2n}\sum_{k=1}^{2n} a_{lk}R_lR_k = \sum_{t=1}^{n} \frac{1}{n} U_t^{(i)}U_t^{(j)}\,.$$

By Lemma B.11, for all $c \ge 1$ there exists $K_c$ such that

$$\mathrm{E}[|A - \mathrm{E}A|^{2c}] \le K_c \left(\sum_l \sum_k a_{lk}^2\right)^c\,.$$

Since $i \ne j, \mathrm{E}[A] = 0$,

$$\mathrm{E}\left(\left|\frac{1}{n}\sum_{t=1}^{n} U_t^{(i)}U_t^{(j)}\right|^{2c}\right) \le K_c \left(\sum_{l=1}^{2n}\sum_{k=1}^{2n} a_{lk}^2\right)^c = K_c \left(\sum_{t=1}^{n} \frac{1}{n^2}\right)^c = K_c\, n^{-c}\,.$$

Note that $\sum_{n=0}^{\infty} n^{-c} < \infty$ for all $c > 1$, hence by Lemma B.3, $n^{-1}\sum_{t=1}^{n} U_t^{(i)}U_t^{(j)} \to 0$ with probability 1 as $n \to \infty$. Next in (B.24),

$$\bar{B}_n^{(j)} \frac{1}{n}\sum_{t=1}^{n} U_t^{(i)} \le \frac{1}{n}\left|B^{(j)}(n)\right| \left|\frac{1}{n}\sum_{t=1}^{n} U_t^{(i)}\right|$$

$$< \frac{1}{n}(1+\epsilon)\sqrt{2n\log\log n}\left|\frac{1}{n}\sum_{t=1}^{n} U_t^{(i)}\right| \quad \text{by Lemma B.1}$$
$$< \sqrt{2}(1+\epsilon)\left(\frac{\log n}{n}\right)^{1/2}\left|\frac{1}{n}\sum_{t=1}^{n} U_t^{(i)}\right| .$$

By the classical SLLN

$$\left|\frac{1}{n}\sum_{t=1}^{n} U_t^{(i)}\right| \to 0 \quad \text{w.p. 1 as } n \to \infty.$$

Similarly,

$$\bar{B}_n^{(i)}\frac{1}{n}\sum_{t=1}^{n} U_t^{(j)} \to 0 \quad \text{w.p. 1 as } n \to \infty .$$

Finally,

$$\begin{aligned}\bar{B}_n^{(i)}\bar{B}_n^{(j)} &\le \frac{1}{n^2}\left|B^{(i)}(n)\right|\left|B^{(j)}(n)\right| \\ &< \frac{1}{n^2}(1+\epsilon)^2(2n\log\log n) \quad \text{by Lemma B.1} \\ &< 2(1+\epsilon)^2\left(\frac{\log n}{n}\right) \\ &\to 0 \text{ as } n \to \infty .\end{aligned}$$

Thus, $\tilde{\gamma}_{n,ij}(0) \to 0$ with probability 1 as $n \to \infty$.

2. Now consider the term $\sum_{s=1}^{b_n-1} w_n(s)n^{-1}\sum_{t=1}^{n-s} U_t^{(i)}U_{t+s}^{(j)}$. Define

$$R_1 = U_1^{(i)},\ R_2 = U_2^{(i)}, \ldots,\ R_n = U_n^{(i)},\ R_{(n+1)} = U_1^{(j)}, \ldots,\ R_{2n} = U_n^{(j)} .$$

Thus, $\{R_i : 1 \le i \le 2n\}$ is an i.i.d sequence of normally distributed random

variables. Next, define for $1 \le l, k \le 2n$

$$a_{lk} = \begin{cases} \frac{1}{n} w_n(k-(n+l)), & \text{if } 1 \le l \le n-1, n+2 \le k \le 2n, \\ & \text{and } 1 \le k-(n+l) \le b_n - 1 \\ 0 & \text{otherwise.} \end{cases}$$

Then,

$$\begin{aligned} A &:= \sum_{l=1}^{2n} \sum_{k=1}^{2n} a_{lk} R_l R_k \\ &= \sum_{l=1}^{n-1} \sum_{k=n+2}^{2n} I\{1 \le k-(n+l) \le b_n - 1\} \frac{1}{n} w_n(k-(n+l)) R_l R_k \\ &= \sum_{l=1}^{n-1} \sum_{s=2-l}^{n-l} I\{1 \le s \le b_n - 1\} \frac{1}{n} w_n(s) R_l R_{n+l+s} \quad \text{letting } k-(n+l) = s \\ &= \sum_{s=1}^{n-1} \sum_{l=1}^{n-s} I\{1 \le s \le b_n - 1\} \frac{1}{n} w_n(s) R_l R_{n+l+s} \\ &\quad + \sum_{s=(3-n)}^{0} \sum_{l=(2-s)}^{n-1} I\{1 \le s \le b_n - 1\} \frac{1}{n} w_n(s) R_l R_{n+l+s} \\ &= \sum_{s=1}^{b_n-1} \sum_{l=1}^{n-s} \frac{1}{n} w_n(s) U_l^{(i)} U_{l+s}^{(j)} \qquad \text{since } n > 2b_n \ge 2 \\ &= \sum_{s=1}^{b_n-1} \sum_{l=1}^{n-s} \frac{1}{n} w_n(s) U_l^{(i)} U_{l+s}^{(j)}. \end{aligned}$$

Using Lemma B.11, for $c \ge 1$ and some constant $K_c$,

$$\mathrm{E}\left[ \left( \sum_{s=1}^{b_n-1} w_n(s) \frac{1}{n} \sum_{t=1}^{n-s} U_t^{(i)} U_{t+s}^{(j)} \right)^{2c} \right] \le K_c \left( \sum_l \sum_k a_{lk}^2 \right)^c,$$

where

$$\sum_l \sum_k a_{lk}^2 = \sum_{s=1}^{b_n-1} \sum_{t=1}^{n-s} \frac{1}{n^2} w_n^2(s) = \frac{1}{n^2} \sum_{s=1}^{b_n-1} (n-s) w_n^2(s) \le \frac{n}{n^2} \sum_{s=1}^{b_n-1} 1 \le \frac{b_n}{n}.$$

Thus, by Assumption (a) and Lemma B.3,

$$\sum_{s=1}^{b_n-1} w_n(s)\frac{1}{n}\sum_{t=1}^{n-s} U_t^{(i)}U_{t+s}^{(j)} \to 0 \text{ w.p. } 1 \text{ as } n\to\infty.$$

3. By letting $t-s=l$,

$$\sum_{s=1}^{b_n-1} w_n(s)\frac{1}{n}\sum_{t=1+s}^{n} U_t^{(i)}U_{t-s}^{(j)} = \sum_{s=1}^{b_n-1} w_n(s)\frac{1}{n}\sum_{l=1}^{n-s} U_{l+s}^{(i)}U_l^{(j)}.$$

This is similar to the previous part with just the $i$ and $j$ components interchanged. A similar argument will lead to $\sum_{s=1}^{b_n-1} w_n(s)n^{-1}\sum_{t=1+s}^{n} U_t^{(i)}U_{t-s}^{(j)} \to 0$ with probability 1 as $n\to\infty$.

4.

$$\begin{aligned}
&\sum_{s=1}^{b_n-1} 2w_n(s)\left(1+\frac{s}{n}\right)\bar{B}_n^{(i)}\bar{B}_n^{(j)}\\
&\le \left|\sum_{s=1}^{b_n-1} 2w_n(s)\left(1+\frac{s}{n}\right)\bar{B}_n^{(i)}\bar{B}_n^{(j)}\right|\\
&\le \sum_{s=1}^{b_n-1} 2|w_n(s)|\left(1+\frac{s}{n}\right)\left|\bar{B}_n^{(i)}\right|\left|\bar{B}_n^{(j)}\right|\\
&\le \frac{2}{n^2}\sum_{s=1}^{b_n-1}\left(1+\frac{s}{n}\right)\left|B^{(i)}(n)\right|\left|B^{(j)}(n)\right| \qquad \text{since } |w_n(s)|\le 1\\
&< \frac{2}{n^2}(1+\epsilon)^2 2n\log\log n\sum_{s=1}^{b_n-1}\left(1+\frac{s}{n}\right) \qquad \text{by Lemma B.1}\\
&< 4(1+\epsilon)^2 n^{-1}\log n\sum_{s=1}^{b_n-1} 2\\
&\le 8(1+\epsilon)^2 b_n n^{-1}\log n\\
&\to 0\,.
\end{aligned}$$

5. Next

$$
\begin{aligned}
&\sum_{s=1}^{b_n-1} w_n(s)\frac{1}{n}\bar{B}_n^{(j)}\left(B^{(i)}(n)-B^{(i)}(n-s)\right)\\
&\le\left|\sum_{s=1}^{b_n-1} w_n(s)\frac{1}{n}\bar{B}_n^{(j)}\left(B^{(i)}(n)-B^{(i)}(n-s)\right)\right|\\
&\le\sum_{s=1}^{b_n-1}\frac{1}{n^2}\left|B^{(j)}(n)\right|\left|B^{(i)}(n)-B^{(i)}(n-s)\right| \qquad \text{since } |w_n(s)|\le 1\\
&\le\frac{1}{n^2}\left|B^{(j)}(n)\right|\sum_{s=1}^{b_n-1}\sup_{0\le m\le b_n}|B^{(i)}(n)-B^{(i)}(n-m)|\\
&<\frac{1}{n^2}\left((1+\epsilon)(2n\log\log n)^{1/2}\right)(1+\epsilon)\\
&\quad\times\left(2b_n\left(\log\frac{n}{b_n}+\log\log n\right)\right)^{1/2}\sum_{s=1}^{b_n-1}1 \quad \text{by Lemma B.1}\\
&<2^{1/2}(1+\epsilon)^2\frac{1}{n^2}(n\log n)^{1/2}(4b_n\log n)^{1/2}b_n\\
&<2^{3/2}(1+\epsilon)^2\left(\frac{b_n}{n}\right)^{1/2}n^{-1}b_n\log n\\
&\to 0\ .
\end{aligned}
$$

6. Similar to the previous term, but exchanging the $i$ and $j$ indices,

$$\sum_{s=1}^{b_n-1} w_n(s)\frac{1}{n}\bar{B}_n^{(i)}\left(B^{(j)}(n)-B^{(j)}(n-s)\right)\to 0 \text{ with probability 1 as } n\to\infty.$$

7.

$$
\begin{aligned}
&\sum_{s=1}^{b_n-1} w_n(s)\frac{1}{n}\bar{B}_n^{(i)}B^{(j)}(s)\\
&\le\left|\sum_{s=1}^{b_n-1} w_n(s)\frac{1}{n}\bar{B}_n^{(i)}B^{(j)}(s)\right|\\
&\le\sum_{s=1}^{b_n-1}|w_n(s)|\frac{1}{n}\left|\bar{B}_n^{(i)}\right|\left|B^{(j)}(s)\right|
\end{aligned}
$$

$$
\begin{aligned}
&\leq \frac{1}{n^2}\left|B^{(i)}(n)\right| \sum_{s=1}^{b_n-1} \left|B^{(j)}(s)\right| \qquad \text{since } |w_n(s)| \leq 1 \\
&< \frac{1}{n^2}(1+\epsilon)(2n\log\log n)^{1/2} \sum_{s=1}^{b_n-1} \sup_{1\leq m\leq b_n} |B^{(j)}(m)| \qquad \text{by Lemma B.1} \\
&< \frac{1}{n^2}(1+\epsilon)(2n\log\log n)^{1/2} \sup_{1\leq m\leq b_n} |B^{(j)}(m+0) - B^{(j)}(0)| \sum_{s=1}^{b_n-1} 1 \\
&< \frac{b_n}{n^2}(1+\epsilon)(2n\log\log n)^{1/2} \sup_{0\leq t\leq n-b_n} \sup_{0\leq m\leq b_n} |B^{(j)}(t+m) - B^{(j)}(t)| \\
&< \frac{b_n}{n^2}(1+\epsilon)(2n\log\log n)^{1/2}(1+\epsilon)\left(2b_n\left(\log\frac{n}{b_n} + \log\log n\right)\right)^{1/2} \\
&= 2^{3/2}(1+\epsilon)^2 \frac{b_n^{1/2}}{n^{1/2}} b_n n^{-1}\log n \\
&\to 0\,.
\end{aligned}
$$

8. Similar to the previous term, by exchanging the $i$ and $j$ index,

$$\sum_{s=1}^{b_n-1} w_n(s)\frac{1}{n}\bar{B}_n^{(j)}B^{(i)}(s) \to 0 \text{ w.p. 1 as } n\to\infty.$$

Since each term in (B.23) goes to 0, we get that

$$\widetilde{\Sigma}_{SV,ij} \to 0 \text{ w.p. 1 as } n\to\infty.$$

□

**Lemma B.13** Let Conditions 4.1 and 4.2 hold. In addition, suppose there exists a constant $c \geq 1$ such that $\sum_n (b_n/n)^c < \infty$, $n > 2b_n$, $b_n n^{-1}\log n \to 0$, and $b_n n^{-1}\sum_{k=1}^{b_n} k|\Delta_1 w_n(k)| \to 0$, then $\widetilde{\Sigma}_{w,n} \to I_p$ w.p. 1 as $n\to\infty$. □

**Proof**

The result follows from Lemmas B.6, B.8 and B.12. □

The following corollary is an immediate consequence of the previous lemma.

**Corollary B.4** Under the conditions of Lemma B.13, $L\widetilde{\Sigma}_{w,n}L^T \to LL^T = \Sigma$ w.p. 1 as $n \to \infty$. □

**Lemma B.14** Suppose (2.4) holds and Conditions 4.1 and 4.2 hold. If as $n \to \infty$,

$$b_n \psi(n)^2 \log n \left( \sum_{k=1}^{b_n} |\Delta_2 w_n(k)| \right)^2 \to 0 \quad \text{and} \tag{B.25}$$

$$\psi(n)^2 \sum_{k=1}^{b_n} |\Delta_2 w_n(k)| \to 0, \tag{B.26}$$

then $\widehat{\Sigma}_{w,n} \to \Sigma$ w.p. 1. □

**Proof**

For $i, j = 1, \ldots, p$, let $\Sigma_{ij}$ and $\widehat{\Sigma}_{w,ij}$ denote the $(i,j)$th element of $\Sigma$ and $\widehat{\Sigma}_{w,n}$ respectively. Recall

$$\widehat{\Sigma}_{w,ij} = \frac{1}{n} \sum_{l=0}^{n-b_n} \sum_{k=1}^{b_n} k^2 \Delta_2 w_n(k) \left[ \bar{Y}_l^{(i)}(k) - \bar{Y}_n^{(i)} \right] \left[ \bar{Y}_l^{(j)}(k) - \bar{Y}_n^{(j)} \right].$$

We have

$$\begin{aligned}
&|\widehat{\Sigma}_{w,ij} - \Sigma_{ij}| \\
&= \left| \frac{1}{n} \sum_{l=0}^{n-b_n} \sum_{k=1}^{b_n} k^2 \Delta_2 w_n(k) \left[ \bar{Y}_l^{(i)}(k) - \bar{Y}_n^{(i)} \right] \left[ \bar{Y}_l^{(j)}(k) - \bar{Y}_n^{(j)} \right] - \Sigma_{ij} \right| \\
&= \left| \frac{1}{n} \sum_{l=0}^{n-b_n} \sum_{k=1}^{b_n} k^2 \Delta_2 w_n(k) \left[ \bar{Y}_l^{(i)}(k) - \bar{Y}_n^{(i)} \pm \bar{C}_l^{(i)}(k) \pm \bar{C}_n^{(i)} \right] \right. \\
&\quad \left. \times \left[ \bar{Y}_l^{(j)}(k) - \bar{Y}_n^{(j)} \pm \bar{C}_l^{(j)}(k) \pm \bar{C}_n^{(j)} \right] - \Sigma_{ij} \right| \\
&= \left| \frac{1}{n} \sum_{l=0}^{n-b_n} \sum_{k=1}^{b_n} k^2 \Delta_2 w_n(k) \left[ \left( \bar{Y}_l^{(i)}(k) - \bar{C}_l^{(i)}(k) \right) \right. \right. \\
&\quad \left. \left. + \left( \bar{C}_l^{(i)}(k) - \bar{C}_n^{(i)} \right) - \left( \bar{Y}_n^{(i)} - \bar{C}_n^{(i)} \right) \right] \right.
\end{aligned}$$

$$
\begin{aligned}
&\times \left[\left(\bar{Y}_l^{(j)}(k) - \bar{C}_l^{(j)}(k)\right) + \left(\bar{C}_l^{(j)}(k) - \bar{C}_n^{(j)}\right) - \left(\bar{Y}_n^{(j)} - \bar{C}_n^{(j)}\right)\right] - \Sigma_{ij}\Bigg| \\
\leq &\left|\frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{k=1}^{b_n} k^2\Delta_2 w_n(k)\left(\bar{C}_l^{(i)}(k) - \bar{C}_n^{(i)}\right)\left(\bar{C}_l^{(j)}(k) - \bar{C}_n^{(j)}\right) - \Sigma_{ij}\right| \\
&+ \frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{k=1}^{b_n} k^2|\Delta_2 w_n(k)|\Bigg[\left|\left(\bar{Y}_l^{(i)}(k) - \bar{C}_l^{(i)}(k)\right)\left(\bar{Y}_l^{(j)}(k) - \bar{C}_l^{(j)}(k)\right)\right| \\
&+ \left|\left(\bar{Y}_l^{(i)}(k) - \bar{C}_l^{(i)}(k)\right)\left(\bar{Y}_n^{(j)} - \bar{C}_n^{(j)}\right)\right| \\
&+ \left|\left(\bar{Y}_l^{(i)}(k) - \bar{C}_l^{(i)}(k)\right)\left(\bar{C}_l^{(j)}(k) - \bar{C}_n^{(j)}\right)\right| + \left|\left(\bar{Y}_n^{(i)} - \bar{C}_n^{(i)}\right)\left(\bar{Y}_n^{(j)} - \bar{C}_n^{(j)}\right)\right| \\
&+ \left|\left(\bar{Y}_n^{(i)} - \bar{C}_n^{(i)}\right)\left(\bar{Y}_l^{(j)}(k) - \bar{C}_l^{(j)}(k)\right)\right| + \left|\left(\bar{Y}_n^{(i)} - \bar{C}_n^{(i)}\right)\left(\bar{C}_l^{(j)}(k) - \bar{C}_n^{(j)}\right)\right| \\
&+ \left|\left(\bar{C}_l^{(i)}(k) - \bar{C}_n^{(i)}\right)\left(\bar{Y}_l^{(j)}(k) - \bar{C}_l^{(j)}(k)\right)\right| \\
&+ \left|\left(\bar{C}_l^{(i)}(k) - \bar{C}_n^{(i)}\right)\left(\bar{Y}_n^{(j)} - \bar{C}_n^{(j)}\right)\right|\Bigg]. \qquad \text{(B.27)}
\end{aligned}
$$

We will show that each of the nine terms in (B.27) goes to 0 with probability 1 as $n \to \infty$. To do that, let us first establish a useful inequality. From (2.4), for any component $i$, and sufficiently large $n$,

$$\left|\sum_{t=1}^{n} Y_t^{(i)} - C^{(i)}(n)\right| < D\psi(n). \qquad \text{(B.28)}$$

1. $\left|\frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{k=1}^{b_n} k^2\Delta_2 w_n(k)\left(\bar{C}_l^{(i)}(k) - \bar{C}_n^{(i)}\right)\left(\bar{C}_l^{(j)}(k) - \bar{C}_n^{(j)}\right) - \Sigma_{ij}\right|$

   Notice that

   $$\frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{k=1}^{b_n} k^2\Delta_2 w_n(k)\left(\bar{C}_l^{(i)}(k) - \bar{C}_n^{(i)}\right)\left(\bar{C}_l^{(j)}(k) - \bar{C}_n^{(j)}\right),$$

   is equivalent to the $ij$th entry in $L\widetilde{\Sigma}_{w,n}L^T$. Then by Corollary B.4 as $n \to \infty$, with probability 1

   $$\left|\frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{k=1}^{b_n} k^2\Delta_2 w_n(k)\left(\bar{C}_l^{(i)}(k) - \bar{C}_n^{(i)}\right)\left(\bar{C}_l^{(j)}(k) - \bar{C}_n^{(j)}\right) - \Sigma_{ij}\right| \to 0\,.$$

2. $\frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{k=1}^{b_n} k^2|\Delta_2 w_n(k)|\left|\left(\bar{Y}_l^{(i)}(k)-\bar{C}_l^{(i)}(k)\right)\left(\bar{Y}_l^{(j)}(k)-\bar{C}_l^{(j)}(k)\right)\right|$

Note that for any component $i$,

$$\begin{aligned}\left|k\left(\bar{Y}_l^{(i)}(k)-\bar{C}_l^{(i)}(k)\right)\right| &= \left|\sum_{t=1}^{k} Y_{l+t}^{(i)} - C^{(i)}(k+l)+C^{(i)}(l)\right| \\ &= \left|\sum_{t=1}^{k+l} Y_t^{(i)} - \sum_{t=1}^{l} Y_t^{(i)} - C^{(i)}(k+l)+C^{(i)}(l)\right| \\ &< \left|\sum_{t=1}^{l+k} Y_t^{(i)} - C^{(i)}(k+l)\right| + \left|\sum_{t=1}^{l} Y_t^{(i)} - C^{(i)}(l)\right| \\ &\le D\psi(l+k)+D\psi(l) \qquad \text{by (B.28)} \\ &\le 2D\psi(n) \qquad \text{since } l+k\le n. \end{aligned} \tag{B.29}$$

By (B.29),

$$\begin{aligned}&\frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{k=1}^{b_n} k^2|\Delta_2 w_n(k)|\left|\left(\bar{Y}_l^{(i)}(k)-\bar{C}_l^{(i)}(k)\right)\left(\bar{Y}_l^{(j)}(k)-\bar{C}_l^{(j)}(k)\right)\right| \\ &< \frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{k=1}^{b_n}|\Delta_2 w_n(k)|\left(2D\psi(n)\right)^2 \\ &= 4D^2\left(\frac{n-b_n+1}{n}\right)\psi(n)^2\sum_{k=1}^{b_n}|\Delta_2 w_n(k)| \\ &\to 0 \quad \text{as } n\to\infty \text{ with probability 1}. \end{aligned}$$

3. $\frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{k=1}^{b_n} k^2|\Delta_2 w_n(k)|\left|\left(\bar{Y}_l^{(i)}(k)-\bar{C}_l^{(i)}(k)\right)\left(\bar{Y}_n^{(j)}-\bar{C}_n^{(j)}\right)\right|$

Note that for any component $i$, using (B.28),

$$\left|\bar{Y}_n^{(i)}-\bar{C}_n^{(i)}\right| = \frac{1}{n}\left|\sum_{t=1}^{n} Y_t^{(i)} - C^{(i)}(n)\right| < \frac{1}{n}D\psi(n)\,. \tag{B.30}$$

By (B.29) and (B.30),

$$\begin{aligned}
&\frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{k=1}^{b_n} k^2|\Delta_2 w_n(k)|\left|\left(\bar{Y}_l^{(i)}(k)-\bar{C}_l^{(i)}(k)\right)\left(\bar{Y}_n^{(j)}-\bar{C}_n^{(j)}\right)\right| \\
&< \frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{k=1}^{b_n} k|\Delta_2 w_n(k)|\left(2D\psi(n)\right)\left(\frac{1}{n}D\psi(n)\right) \\
&= 2D^2\psi(n)^2\left(\frac{n-b_n+1}{n}\right)\frac{1}{n}\sum_{k=1}^{b_n} k|\Delta_2 w_n(k)| \\
&\leq 2D^2\psi(n)^2\left(\frac{n-b_n+1}{n}\right)\frac{b_n}{n}\sum_{k=1}^{b_n}|\Delta_2 w_n(k)| \\
&\to 0 \quad \text{as } n\to\infty \text{ with probability 1}.
\end{aligned}$$

4. Now

$$\begin{aligned}
&\frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{k=1}^{b_n} k^2|\Delta_2 w_n(k)|\left|\left(\bar{Y}_l^{(i)}(k)-\bar{C}_l^{(i)}(k)\right)\left(\bar{C}_l^{(j)}(k)-\bar{C}_n^{(j)}\right)\right| \\
\leq &\frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{k=1}^{b_n} k^2|\Delta_2 w_n(k)|\left|\left(\bar{Y}_l^{(i)}(k)-\bar{C}_l^{(i)}(k)\right)\bar{C}_l^{(j)}(k)\right| \\
&+\frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{k=1}^{b_n} k^2|\Delta_2 w_n(k)|\left|\left(\bar{Y}_l^{(i)}(k)-\bar{C}_l^{(i)}(k)\right)\bar{C}_n^{(j)}\right|.
\end{aligned}$$

We will show that both parts of the sum converge to 0 with probability 1 as $n\to\infty$. Consider the first sum.

$$\begin{aligned}
&\frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{k=1}^{b_n} k^2|\Delta_2 w_n(k)|\left|\left(\bar{Y}_l^{(i)}(k)-\bar{C}_l^{(i)}(k)\right)\bar{C}_l^{(j)}(k)\right| \\
&\leq \frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{k=1}^{b_n} k^2|\Delta_2 w_n(k)|\left|\bar{Y}_l^{(i)}(k)-\bar{C}_l^{(i)}(k)\right|\left|\bar{C}_l^{(j)}(k)\right|
\end{aligned}$$

by (B.2) and (B.29)

$$\leq \frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{k=1}^{b_n} k|\Delta_2 w_n(k)|\,(2D\psi(n))\left(2(1+\epsilon)\sqrt{b_n\Sigma_{ii}}\frac{1}{k}(\log n)^{1/2}\right)$$
$$= \left(\frac{n-b_n+1}{n}\right)4D(1+\epsilon)\sqrt{\Sigma_{ii}b_n\log n}\,\psi(n)\sum_{k=1}^{b_n}|\Delta_2 w_n(k)|$$
$$\to 0 \text{ by Condition 4.2 and (B.25) .}$$

The second part is

$$\frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{k=1}^{b_n} k^2|\Delta_2 w_n(k)|\left|\left(\bar{Y}_l^{(i)}(k)-\bar{C}_l^{(i)}(k)\right)\bar{C}_n^{(j)}\right|$$
$$= \frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{k=1}^{b_n} k^2|\Delta_2 w_n(k)|\left|\bar{Y}_l^{(i)}(k)-\bar{C}_l^{(i)}(k)\right|\left|\bar{C}_n^{(j)}\right|$$

by (B.29) and (B.1)

$$\leq \frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{k=1}^{b_n} k|\Delta_2 w_n(k)|\,(2D\psi(n))\left(\frac{1}{n}(1+\epsilon)\left[2n\Sigma_{ii}\log\log n\right]^{1/2}\right)$$
$$< \left(\frac{n-b_n+1}{n}\right)2\sqrt{2\Sigma_{ii}}\,D(1+\epsilon)\psi(n)\frac{(n\log n)^{1/2}}{n}\sum_{k=1}^{b_n}k|\Delta_2 w_n(k)|$$
$$< \left(\frac{n-b_n+1}{n}\right)2\sqrt{2\Sigma_{ii}}\,D(1+\epsilon)\psi(n)\frac{(\log n)^{1/2}}{n^{1/2}}b_n\sum_{k=1}^{b_n}|\Delta_2 w_n(k)|$$
$$< \left(\frac{n-b_n+1}{n}\right)2\sqrt{2\Sigma_{ii}}\,D(1+\epsilon)\psi(n)(b_n\log n)^{1/2}\frac{b_n^{1/2}}{n^{1/2}}\sum_{k=1}^{b_n}|\Delta_2 w_n(k)|$$
$$\to 0 \text{ by Condition 4.2 and (B.25) .}$$

5. Next,

$$\begin{aligned}
&\frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{k=1}^{b_n} k^2|\Delta_2 w_n(k)|\left|\left(\bar{Y}_n^{(i)}-\bar{C}_n^{(i)}\right)\left(\bar{Y}_n^{(j)}-\bar{C}_n^{(j)}\right)\right| \\
&< \frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{k=1}^{b_n} k^2|\Delta_2 w_n(k)|\frac{1}{n^2}D^2\psi(n)^2 \qquad \text{by (B.30)} \\
&\leq \left(\frac{n-b_n+1}{n}\right) D^2\frac{b_n^2}{n^2}\psi(n)^2\sum_{k=1}^{b_n}|\Delta_2 w_n(k)| \\
&\to 0 \text{ by Condition 4.2 and (B.26) .}
\end{aligned}$$

6.

$$\begin{aligned}
&\frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{k=1}^{b_n} k^2|\Delta_2 w_n(k)|\left|\left(\bar{Y}_n^{(i)}-\bar{C}_n^{(i)}\right)\left(\bar{Y}_l^{(j)}(k)-\bar{C}_l^{(j)}(k)\right)\right| \\
&< \frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{k=1}^{b_n} k|\Delta_2 w_n(k)|\left(\frac{1}{n}D\psi(n)\right)(2D\psi(n)) \\
&< \left(\frac{n-b_n+1}{n}\right) 2D^2\psi(n)^2\frac{1}{n}\sum_{k=1}^{b_n} k|\Delta_2 w_n(k)| \\
&< \left(\frac{n-b_n+1}{n}\right) 2D^2\psi(n)^2\frac{b_n}{n}\sum_{k=1}^{b_n}|\Delta_2 w_n(k)| \\
&\to 0 \text{ by Condition 4.2 and (B.26) .}
\end{aligned}$$

7.

$$\begin{aligned}
&\frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{k=1}^{b_n} k^2|\Delta_2 w_n(k)|\left|\left(\bar{Y}_n^{(i)}-\bar{C}_n^{(i)}\right)\left(\bar{C}_l^{(j)}(k)-\bar{C}_n^{(j)}\right)\right| \\
&\leq \frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{k=1}^{b_n} k^2|\Delta_2 w_n(k)|\left|\left(\bar{Y}_n^{(i)}-\bar{C}_n^{(i)}\right)\bar{C}_l^{(j)}(k)\right|
\end{aligned}$$

$$+\frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{k=1}^{b_n}k^2|\Delta_2 w_n(k)|\left|\left(\bar{Y}_n^{(i)}-\bar{C}_n^{(i)}\right)\bar{C}_n^{(j)}\right|.$$

We will show that each of the two terms goes to 0 with probability 1 as $n\to\infty$.

$$\frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{k=1}^{b_n}k^2|\Delta_2 w_n(k)|\left|\left(\bar{Y}_n^{(i)}-\bar{C}_n^{(i)}\right)\bar{C}_l^{(j)}(k)\right|$$

by (B.2) and (B.30)

$$\begin{aligned}
&< \frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{k=1}^{b_n}k^2|\Delta_2 w_n(k)|\left(\frac{1}{n}D\psi(n)\right)\left(2(1+\epsilon)\sqrt{b_n\Sigma_{ii}}\frac{1}{k}(\log n)^{1/2}\right)\\
&< \left(\frac{n-b_n+1}{n}\right)2D(1+\epsilon)\sqrt{\Sigma_{ii}}\psi(n)\frac{\sqrt{b_n\log n}}{n}\sum_{k=1}^{b_n}k|\Delta_2 w_n(k)|\\
&\le \left(\frac{n-b_n+1}{n}\right)2D(1+\epsilon)\sqrt{\Sigma_{ii}}\frac{b_n}{n}\sqrt{b_n\log n}\,\psi(n)\sum_{k=1}^{b_n}|\Delta_2 w_n(k)|\\
&\to 0 \text{ by Condition 4.2 and (B.25)}.
\end{aligned}$$

For the second term,

$$\begin{aligned}
&\frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{k=1}^{b_n}k^2|\Delta_2 w_n(k)|\left|\left(\bar{Y}_n^{(i)}-\bar{C}_n^{(i)}\right)\bar{C}_n^{(j)}\right|\\
&=\frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{k=1}^{b_n}k^2|\Delta_2 w_n(k)|\left|\left(\bar{Y}_n^{(i)}-\bar{C}_n^{(i)}\right)\right|\left|\bar{C}_n^{(j)}\right|
\end{aligned}$$

by (B.30) and (B.1)

$$< \frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{k=1}^{b_n}k^2|\Delta_2 w_n(k)|\left(\frac{1}{n}D\psi(n)\right)\left(\frac{1}{n}(1+\epsilon)\left[2n\Sigma_{ii}\log\log n\right]^{1/2}\right)$$

$$\leq \left(\frac{n-b_n+1}{n}\right)\sqrt{2\Sigma_{ii}}D(1+\epsilon)\frac{\sqrt{n\log n}\,\psi(n)}{n^2}\sum_{k=1}^{b_n}k^2|\Delta_2 w_n(k)|$$

$$\leq \left(\frac{n-b_n+1}{n}\right)\sqrt{2\Sigma_{ii}}D(1+\epsilon)\frac{b_n^2\sqrt{n\log n}\,\psi(n)}{n^2}\sum_{k=1}^{b_n}|\Delta_2 w_n(k)|$$

$$\leq \left(\frac{n-b_n+1}{n}\right)\sqrt{2\Sigma_{ii}}D(1+\epsilon)\frac{b_n}{n}\frac{b_n^{1/2}}{n^{1/2}}(b_n\log n)^{1/2}\psi(n)\sum_{k=1}^{b_n}|\Delta_2 w_n(k)|$$

$\to 0$ by Condition 4.2 and (B.25) .

8. $$\frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{k=1}^{b_n}k^2|\Delta_2 w_n(k)|\left|\left(\bar{C}_l^{(i)}(k)-\bar{C}_n^{(i)}\right)\left(\bar{Y}_l^{(i)}(k)-\bar{C}_l^{(j)}(k)\right)\right|.$$

This term is the same as term 4 except for a change of components. Thus the same argument can be used to show that it converges to 0 with probability 1 as $n\to\infty$.

9. $$\frac{1}{n}\sum_{l=0}^{n-b_n}\sum_{k=1}^{b_n}k^2|\Delta_2 w_n(k)|\left|\left(\bar{C}_l^{(i)}(k)-\bar{C}_n^{(i)}\right)\left(\bar{Y}_n^{(j)}-\bar{C}_n^{(j)}\right)\right|.$$

This term is the same as term 7 except for a change of components. Thus the same argument can be used to show that it converges to 0 w.p. 1 as $n\to\infty$.

Since each of the nine terms converges to 0 with probability 1, $|\widehat{\Sigma}_{ij}-\Sigma_{ij}|\to 0$ as $n\to\infty$ with probability 1. □

Since we proved that $\Sigma_{SV}=\widehat{\Sigma}_{w,n}+d_n\to\Sigma+0$ as $n\to\infty$ with probability 1, we have the desired result for Theorem 4.1.

**Proof (Proof of Theorem 4.2)**

Since $\mathrm{E}_F\|g\|^{4+\delta}<\infty$ implies $\mathrm{E}_F\|g\|^{2+\delta}<\infty$ and $X$ is a polynomially ergodic Markov chain of order $m\geq(1+\epsilon_1)(1+2/\delta)$ we have from Corollary 2.1 that an SIP holds

such that

$$\left\| \sum_{t=1}^{n} g(X_t) - n\theta - LB(n) \right\| < D\, n^{1/2-\lambda_g}.$$

for some $\lambda_g > 0$ depending on $\epsilon$, $\delta$, and $p$ only.

Since $\mathrm{E}_F\|g\|^{4+\delta} < \infty$ implies $\mathrm{E}_F\|h\|^{2+\delta} < \infty$ and $X$ is a polynomially ergodic Markov chain of order $m \geq (1+\epsilon_1)(1+2/\delta)$ we have from Corollary 2.1 that an SIP holds such that

$$\left\| \sum_{t=1}^{n} h(X_t) - n\theta_h - L_h B(n) \right\| < D_h\, n^{1/2-\lambda_h}.$$

for some $\lambda_h > 0$ depending on $\epsilon$, $\delta$, and $p$ only.

Setting $\lambda = \min\{\lambda_g, \lambda_h\}$ shows that (2.4) and (4.3) hold with

$$\psi(n) = \psi_h(n) = n^{1/2-\lambda}\ .$$

The rest now follows easily from Theorem 4.1. □

## B.3 Proof of Theorem 4.3

We first state the theorem more generally for processes that satisfy a strong invariance principle. Then, the proof of Theorem 4.3 will follow from Theorem B.2 below and Corollary 2.1.

**Theorem B.2** Let (2.4) and Condition 4.5 hold. If $b_n^{-1/2}(\log n)^{1/2}\psi(n) \to 0$ as $n \to \infty$, then $\Sigma_{BM} \to \Sigma$ w.p. 1 as $n \to \infty$. □

The proof of this theorem will be presented in a series of lemmas. We first construct $\widetilde{\Sigma}$, a Brownian motion equivalent of the batch means estimator and show that

$\widetilde{\Sigma}$ converges to the identity matrix with probability 1 as $n$ increases. This result will be critical in proving the theorem.

Let $B(t)$ be a $p$-dimensional standard Brownian motion, and for $i = 1, \ldots, p$, let $B^{(i)}$ denote the $i$th component univariate Brownian motion. For $k = 0, \ldots, a_n - 1$ define

$$\bar{B}_k^{(i)} = \frac{1}{b_n}(B^{(i)}((k+1)b_n) - B^{(i)}(kb_n)) \quad \text{and} \quad \bar{B}^{(i)}(n) = \frac{1}{n}B^{(i)}(n).$$

For $\bar{B}_k = (\bar{B}_k^{(1)}, \ldots, \bar{B}_k^{(p)})^T$ and $\bar{B}(n) = \left(\bar{B}^{(1)}(n), \ldots, \bar{B}^{(p)}(n)\right)^T$, define

$$\widetilde{\Sigma} = \frac{b_n}{a_n - 1}\sum_{k=0}^{a_n-1}[\bar{B}_k - \bar{B}(n)][\bar{B}_k - \bar{B}(n)]^T.$$

Here $\widetilde{\Sigma}$ is the Brownian motion equivalent of $\Sigma_{BM}$ and in Lemma B.15 we will show that $\widetilde{\Sigma}$ converges to the identity matrix with probability 1. But first we state some results we will need.

**Lemma B.15** If Condition 4.5 holds, then $\widetilde{\Sigma} \to I_p$ with probability 1 as $n \to \infty$ where $I_p$ is the $p \times p$ identity matrix. □

**Proof**

For $i, j = 1, \ldots, p$, let $\widetilde{\Sigma}_{ij}$ denote the $(i,j)$th component of $\widetilde{\Sigma}$. For $i = j$, Damerdji (1994) showed that $\widetilde{\Sigma}_{ij} \to 1$ with probability 1 as $n \to \infty$. Thus it is left to show that for $i \neq j$, $\widetilde{\Sigma}_{ij} \to 0$ with probability 1 as $n \to \infty$.

$$\begin{aligned}\widetilde{\Sigma}_{ij} &= \frac{b_n}{a_n - 1}\sum_{k=0}^{a_n-1}\left[\bar{B}_k^{(i)} - \bar{B}^{(i)}(n)\right]\left[\bar{B}_k^{(j)} - \bar{B}^{(j)}(n)\right] \\ &= \frac{b_n}{a_n - 1}\sum_{k=0}^{a_n-1}\left[\bar{B}_k^{(i)}\bar{B}_k^{(j)} - \bar{B}_k^{(i)}\bar{B}^{(j)}(n) - \bar{B}^{(i)}(n)\bar{B}_k^{(j)} + \bar{B}^{(i)}(n)\bar{B}^{(j)}(n)\right].\end{aligned} \tag{B.31}$$

We will show that each of the four terms in (B.31) converges to 0 with probability 1 as $n \to \infty$. But first we note that by the properties of Brownian motion, for all $k = 0, \ldots, a_n - 1$,

$$\begin{aligned} &\bar{B}_k^{(i)} \overset{iid}{\sim} N\left(0, \frac{1}{b_n}\right) \text{ and } \bar{B}_k^{(j)} \overset{iid}{\sim} N\left(0, \frac{1}{b_n}\right) \text{ independently} \\ \Rightarrow &\sqrt{b_n}\bar{B}_k^{(i)} \overset{iid}{\sim} N(0,1) \text{ and } \sqrt{b_n}\bar{B}_k^{(j)} \overset{iid}{\sim} N(0,1) \text{ independently.} \end{aligned} \tag{B.32}$$

1. Naturally by (B.32),

$$X_k := \sqrt{\frac{b_n}{2}}\bar{B}_k^{(i)} + \sqrt{\frac{b_n}{2}}\bar{B}_k^{(j)} \overset{iid}{\sim} N(0,1) \text{ and } Y_k := \sqrt{\frac{b_n}{2}}\bar{B}_k^{(i)} - \sqrt{\frac{b_n}{2}}\bar{B}_k^{(j)} \overset{iid}{\sim} N(0,1)$$

Notice that $AB = (A+B)^2/4 - (A-B)^2/4$. Using $X_k$ as $(A+B)$ and $Y_k$ as $(A-B)$, we can write $b_n\bar{B}_k^{(i)}\bar{B}_k^{(j)}/2$ as a linear combination of two $\chi^2$ random variables. Specifically,

$$\begin{aligned} \frac{b_n}{a_n - 1}\sum_{k=0}^{a_n-1}\bar{B}_k^{(i)}\bar{B}_k^{(j)} &= \frac{2}{a_n-1}\sum_{k=0}^{a_n-1}\sqrt{\frac{b_n}{2}}\bar{B}_k^{(i)}\sqrt{\frac{b_n}{2}}\bar{B}_k^{(j)} \\ &= \frac{1}{2(a_n-1)}\sum_{k=0}^{a_n-1}[X_k^2 - Y_k^2] \\ &= \frac{1}{2(a_n-1)}\sum_{k=0}^{a_n-1}X_k^2 - \frac{1}{2(a_n-1)}\sum_{k=0}^{a_n-1}Y_k^2 \\ &= \frac{a_n}{2(a_n-1)}\frac{X}{a_n} - \frac{a_n}{2(a_n-1)}\frac{Y}{a_n}, \end{aligned} \tag{B.33}$$

where $X = \sum_{k=0}^{a_n-1} X_k^2 \sim \chi^2_{a_n}$ and $Y = \sum_{k=0}^{a_n-1} Y_k^2 \sim \chi^2_{a_n}$ independently.

By Lemma B.2, for all positive integers $c$,

$$\mathrm{E}\left[(X-a_n)^{2c}\right] \leq K a_n^c \Rightarrow \mathrm{E}\left[\left(\frac{X}{a_n}-1\right)^{2c}\right] \leq K\left(\frac{b_n}{n}\right)^c .$$

Thus by Lemma B.3 and Condition 4.5b, $X/a_n \to 1$ with probability 1, as $n \to \infty$. Similarly, $Y/a_n \to 1$ with probability 1, as $n \to \infty$. Using this result in (B.33) and the fact that $a_n/(a_n-1) \to 1$ as $n \to \infty$,

$$\frac{b_n}{a_n-1}\sum_{k=0}^{a_n-1} \bar{B}_k^{(i)}\bar{B}_k^{(j)} \to 0 \text{ w.p. } 1 \text{ as } n \to \infty.$$

2. By the definition of $\bar{B}(n)$ and $\bar{B}_k$,

$$\begin{aligned}\frac{b_n}{a_n-1}\sum_{k=0}^{a_n-1} \bar{B}_k^{(i)}\bar{B}^{(j)}(n) &= \frac{1}{a_n-1}\frac{1}{n}B^{(j)}(n)\sum_{k=0}^{a_n-1} B^{(i)}((k+1)b_n) - B^{(i)}(kb_n) \\ &= \frac{1}{a_n-1}\frac{1}{n}B^{(j)}(n)B^{(i)}(a_nb_n) \\ &= \frac{a_n}{a_n-1}\frac{\sqrt{b_n}}{n}B^{(j)}(n)\frac{\sqrt{b_n}}{a_nb_n}B^{(i)}(a_nb_n)\ . \end{aligned} \tag{B.34}$$

Using properties of Brownian motion,

$$\begin{aligned} & B^{(j)}(n) \sim N(0,n) \quad \text{and} \quad B^{(i)}(a_nb_n) \sim N(0,a_nb_n) \\ \Rightarrow\ & \frac{\sqrt{b_n}}{n}B^{(j)}(n) \overset{d}{\sim} N\left(0,\frac{b_n}{n}\right) \quad \text{and} \quad \frac{\sqrt{b_n}}{a_nb_n}B^{(i)}(a_nb_n) \overset{d}{\sim} N\left(0,\frac{1}{a_n}\right). \end{aligned} \tag{B.35}$$

As $n \to \infty$ both terms in (B.35) tend to Dirac's delta function. Thus as $n \to \infty$.

$$\frac{\sqrt{b_n}}{n}B^{(j)}(n) \to 0 \text{ w.p. } 1 \quad \text{and} \quad \frac{\sqrt{b_n}}{a_nb_n}B^{(i)}(a_nb_n) \to 0 \text{ w.p. } 1. \tag{B.36}$$

Using (B.36) in (B.34),

$$\frac{b_n}{a_n - 1} \sum_{k=0}^{a_n - 1} \bar{B}_k^{(i)} \bar{B}^{(j)}(n) \to 0 \quad \text{w.p. 1 as } n \to \infty.$$

3. A similar argument as above yields,

$$\frac{b_n}{a_n - 1} \sum_{k=0}^{a_n - 1} \bar{B}^{(i)}(n) \bar{B}_k^{(j)} \to 0 \quad \text{w.p. 1 as } n \to \infty.$$

4. By the definition of $\bar{B}^{(i)}(n)$,

$$\begin{aligned} \frac{b_n}{a_n - 1} \sum_{k=0}^{a_n - 1} \bar{B}^{(i)}(n) \bar{B}^{(j)}(n) &= \frac{b_n}{a_n - 1} a_n \frac{1}{n} B^{(i)}(n) \frac{1}{n} B^{(j)}(n) \\ &= \frac{a_n}{a_n - 1} \frac{\sqrt{b_n}}{n} B^{(i)}(n) \frac{\sqrt{b_n}}{n} B^{(j)}(n) \\ &\to 0 \quad \text{w.p. 1 as } n \to \infty \text{ by (B.36).} \end{aligned}$$

Thus each term in (B.31) goes to 0 with probability 1 as $n \to \infty$, yielding $\widetilde{\Sigma} \to I_p$ with probability 1 as $n \to \infty$. □

**Corollary B.5** If Condition 4.5 holds, then for any conformable matrix $L$, as $n \to \infty$, $L\widetilde{\Sigma}L^T \to LL^T$ with probability 1. □

For the rest of the proof, we will require some results regarding Brownian motion. Recall $L$ in (2.4) and set $\Sigma := LL^T$. Define $C(t) = LB(t)$ and if $C^{(i)}(t)$ is the $i$th component of $C(t)$, define

$$\bar{C}_k^{(i)} := \frac{1}{b_n} \left( C^{(i)}((k+1)b_n) - C^{(i)}(kb_n) \right) \quad \text{and} \quad \bar{C}^{(i)}(n) := \frac{1}{n} C^{(i)}(n).$$

Since $C^{(i)}(t) \sim N(0, t\Sigma_{ii})$, where $\Sigma_{ii}$ is the $i$th diagonal of $\Sigma$, $C^{(i)}/\sqrt{\Sigma_{ii}}$ is a 1-dimensional Brownian motion. As a consequence, we have the following corollaries of Lemma B.1.

**Corollary B.6** For all $\epsilon > 0$ and for almost all sample paths there exists $n_0(\epsilon)$ such that for all $n \geq n_0$ and all $i = 1, \ldots, p$

$$\left|C^{(i)}(n)\right| < (1+\epsilon)\left[2\Sigma_{ii} n \log\log n\right]^{1/2},$$

where $\Sigma_{ii}$ is the $i$th diagonal of $\Sigma$. □

**Corollary B.7** For all $\epsilon > 0$ and for almost all sample paths, there exists $n_0(\epsilon)$ such that for all $n \geq n_0$ and all $i = 1, \ldots, p$

$$\left|\bar{C}_k^{(i)}\right| \leq \sqrt{\frac{2\Sigma_{ii}}{b_n}}(1+\epsilon)\left(\log\frac{n}{b_n} + \log\log n\right)^{1/2},$$

where $\Sigma_{ii}$ is the $i$th diagonal of $\Sigma$. □

We finally come to the last leg of the proof, where we will show that for the $(i,j)$th element of $\Sigma_{BM}$, $|\Sigma_{BM,ij} - \Sigma_{ij}| \to 0$ with probability 1 as $n \to \infty$.

Recall that $g_i(X_t)$ be the $i$th component of $g(X_t)$. Define for each $i = 1, \ldots, p$, the process $V_l^{(i)} = g_i(X_l) - \theta_i$ for $l = 1, 2, \ldots$. Further, for $k = 0, \ldots, a_n - 1$ and $j = 1, \ldots, p$ define

$$\bar{V}_k^{(i)} = \frac{1}{b_n}\sum_{l=1}^{b_n} V_{kb_n+l}^{(i)} \quad \text{and} \quad \bar{V}^{(i)}(n) = \frac{1}{n}\sum_{l=1}^{n} V_l^{(i)}.$$

Then

$$\Sigma_{BM,ij} = \frac{b_n}{a_n - 1}\sum_{k=0}^{a_n-1}\left[\bar{V}_k^{(i)} - \bar{V}^{(i)}(n)\right]\left[\bar{V}_k^{(j)} - \bar{V}^{(j)}(n)\right].$$

We will show that $|\Sigma_{BM,ij} - \Sigma_{ij}| \to 0$ w.p. 1 as $n \to \infty$.

$$|\Sigma_{BM,ij} - \Sigma_{ij}|$$

$$
\begin{aligned}
&= \left| \frac{b_n}{a_n - 1} \sum_{k=0}^{a_n-1} \left[ \bar{V}_k^{(i)} - \bar{V}^{(i)}(n) \right] \left[ \bar{V}_k^{(j)} - \bar{V}^{(j)}(n) \right] - \Sigma_{ij} \right| \\
&= \left| \frac{b_n}{a_n - 1} \sum_{k=0}^{a_n-1} \left[ \bar{V}_k^{(i)} - \bar{V}^{(i)}(n) \pm \bar{C}_k^{(i)} \pm \bar{C}^{(i)}(n) \right] \right. \\
&\quad \left. \times \left[ \bar{V}_k^{(j)} - \bar{V}^{(j)}(n) \pm \bar{C}_k^{(j)} \pm \bar{C}^{(j)}(n) \right] - \Sigma_{ij} \right| \\
&= \left| \frac{b_n}{a_n - 1} \sum_{k=0}^{a_n-1} \left[ \left( \bar{V}_k^{(i)} - \bar{C}_k^{(i)} \right) - \left( \bar{V}^{(i)}(n) - \bar{C}^{(i)}(n) \right) + \left( \bar{C}_k^{(i)} - \bar{C}^{(i)}(n) \right) \right] \right. \\
&\qquad \left. \times \left[ \left( \bar{V}_k^{(j)} - \bar{C}_k^{(j)} \right) - \left( \bar{V}^{(j)}(n) - \bar{C}^{(j)}(n) \right) + \left( \bar{C}_k^{(j)} - \bar{C}^{(j)}(n) \right) \right] - \Sigma_{ij} \right| \\
&\leq \left| \frac{b_n}{a_n - 1} \sum_{k=0}^{a_n-1} \left[ \bar{C}_k^{(i)} - \bar{C}^{(i)}(n) \right] \left[ \bar{C}_k^{(j)} - \bar{C}^{(j)}(n) \right] - \Sigma_{ij} \right| \\
&+ \frac{b_n}{a_n - 1} \sum_{k=0}^{a_n-1} \left[ \left| \left( \bar{V}_k^{(i)} - \bar{C}_k^{(i)} \right) \left( \bar{V}_k^{(j)} - \bar{C}_k^{(j)} \right) \right| \right. \\
&+ \left| \left( \bar{V}^{(i)}(n) - \bar{C}^{(i)}(n) \right) \left( \bar{V}^{(j)}(n) - \bar{C}^{(j)}(n) \right) \right| \\
&+ \left| \left( \bar{V}_k^{(i)} - \bar{C}_k^{(i)} \right) \left( \bar{V}^{(j)}(n) - \bar{C}^{(j)}(n) \right) \right| + \left| \left( \bar{V}^{(i)}(n) - \bar{C}^{(i)}(n) \right) \left( \bar{V}_k^{(j)} - \bar{C}_k^{(j)} \right) \right| \\
&+ \left| \left( \bar{V}_k^{(i)} - \bar{C}_k^{(i)} \right) \bar{C}_k^{(j)} \right| + \left| \left( \bar{V}_k^{(j)} - \bar{C}_k^{(j)} \right) \bar{C}_k^{(i)} \right| \\
&+ \left| \left( \bar{V}_k^{(i)} - \bar{C}_k^{(i)} \right) \bar{C}^{(j)}(n) \right| + \left| \left( \bar{V}_k^{(j)} - \bar{C}_k^{(j)} \right) \bar{C}^{(i)}(n) \right| \\
&+ \left| \left( \bar{V}^{(i)}(n) - \bar{C}^{(i)}(n) \right) \bar{C}_k^{(j)} \right| + \left| \left( \bar{V}^{(j)}(n) - \bar{C}^{(j)}(n) \right) \bar{C}_k^{(i)} \right| \\
&+ \left| \left( \bar{V}^{(i)}(n) - \bar{C}^{(i)}(n) \right) \bar{C}^{(j)}(n) \right| + \left| \left( \bar{V}^{(j)}(n) - \bar{C}^{(j)}(n) \right) \bar{C}^{(i)}(n) \right| \Bigg].
\end{aligned}
$$

We will show that each of the 13 terms above tends to 0 w.p. 1 as $n \to \infty$.

1. Notice that,

$$
\frac{b_n}{a_n - 1} \sum_{k=0}^{a_n-1} \left[ \bar{C}_k^{(i)} - \bar{C}^{(i)}(n) \right] \left[ \bar{C}_k^{(j)} - \bar{C}^{(j)}(n) \right],
$$

is the $(i,j)$th entry in $L\widetilde{\Sigma}L^T$. Thus, by Corollary B.5, with probability 1 as $n \to \infty$,

$$\left| \frac{b_n}{a_n - 1} \sum_{k=0}^{a_n - 1} \left[ \bar{C}_k^{(i)} - \bar{C}^{(i)}(n) \right] \left[ \bar{C}_k^{(j)} - \bar{C}^{(j)}(n) \right] - \Sigma_{ij} \right| \to 0.$$

2. By Condition 1

$$\left\| \sum_{l=0}^{n} V_l - LB(n) \right\| < D\psi(n) \text{ w.p. } 1,$$

where $V_l = \left( V_l^{(1)}, \ldots, V_l^{(p)} \right)$. Hence, for components $i$ and $j$

$$\left| \sum_{l=1}^{n} V_l^{(i)} - C^{(i)}(n) \right| < D\psi(n) \quad \text{and} \quad \left| \sum_{l=1}^{n} V_l^{(j)} - C^{(j)}(n) \right| < D\psi(n). \tag{B.37}$$

Note that,

$$\begin{aligned} \left| \bar{V}_k^{(i)} - \bar{C}_k^{(i)} \right| &= \left| \frac{1}{b_n} \left[ \sum_{l=1}^{(k+1)b_n} V_l^{(i)} - C^{(i)}((k+1)b_n) \right] - \frac{1}{b_n} \left[ \sum_{l=1}^{kb_n} V_l^{(i)} - C^{(i)}(kb_n) \right] \right| \\ &\leq \frac{1}{b_n} \left[ \left| \sum_{l=1}^{(k+1)b_n} V_l^{(i)} - C^{(i)}((k+1)b_n) \right| + \left| \sum_{l=1}^{kb_n} V_l^{(i)} - C^{(i)}(kb_n) \right| \right] \\ &\leq \frac{2}{b_n} D\psi(n). \end{aligned} \tag{B.38}$$

Similarly

$$\left| \bar{V}_k^{(j)} - \bar{C}_k^{(j)} \right| \leq \frac{2}{b_n} D\psi(n). \tag{B.39}$$

Thus, using (B.38) and (B.39),

$$\frac{b_n}{a_n - 1} \sum_{k=0}^{a_n - 1} \left| \left( \bar{V}_k^{(i)} - \bar{C}_k^{(i)} \right) \left( \bar{V}_k^{(j)} - \bar{C}_k^{(j)} \right) \right| \leq \frac{b_n}{a_n - 1} a_n \frac{4D^2}{b_n^2} \psi(n)^2$$

$$\leq 4D^2 \frac{a_n}{a_n - 1} \frac{\log n}{b_n} \psi(n)^2$$
$$\to 0 \text{ w.p 1 as } n \to \infty.$$

3. By (B.37), we get

$$\left|\bar{V}^{(i)}(n) - \bar{C}^{(i)}(n)\right| = \frac{1}{n} \left| \sum_{l=1}^{n} V_l^{(i)} - C^{(i)}(n) \right| < D \frac{\psi(n)}{n}. \tag{B.40}$$

Similarly ,

$$\left|\bar{V}^{(i)}(n) - \bar{C}^{(i)}(n)\right| < D \frac{\psi(n)}{n}. \tag{B.41}$$

Then,

$$\frac{b_n}{a_n - 1} \sum_{k=0}^{a_n - 1} \left| \left(\bar{V}^{(i)}(n) - \bar{C}^{(i)}(n)\right) \left(\bar{V}^{(j)}(n) - \bar{C}^{(j)}(n)\right) \right|$$
$$< \frac{b_n}{a_n - 1} a_n D^2 \frac{\psi(n)^2}{n^2}$$
$$= D^2 \frac{a_n}{a_n - 1} \frac{b_n}{n} \frac{b_n}{n} \frac{\psi(n)^2}{b_n}$$
$$< D^2 \frac{a_n}{a_n - 1} \frac{b_n}{n} \frac{b_n}{n} \frac{\psi(n)^2 \log n}{b_n}$$
$$\to 0 \text{ w.p 1 as } n \to \infty \text{ by Condition 4.5a.}$$

4. By (B.38) and (B.41), we have

$$\frac{b_n}{a_n - 1} \sum_{k=0}^{a_n - 1} \left| \left(\bar{V}_k^{(i)} - \bar{C}_k^{(i)}\right) \left(\bar{V}^{(j)}(n) - \bar{C}^{(j)}(n)\right) \right|$$
$$\leq \frac{b_n}{a_n - 1} a_n \left(\frac{2D}{b_n} \psi(n)\right) \left(\frac{D}{n} \psi(n)\right)$$

$$< 2D^2 \frac{a_n}{a_n - 1} \frac{b_n}{n} \frac{\psi(n)^2 \log n}{b_n}$$

$\to 0$ w.p. 1 as $n \to \infty$ by Condition 4.5a.

5. By (B.39) and (B.40), we have

$$\frac{b_n}{a_n - 1} \sum_{k=0}^{a_n - 1} \left| \left( \bar{V}^{(i)}(n) - \bar{C}^{(i)}(n) \right) \left( \bar{V}_k^{(j)} - \bar{C}_k^{(j)} \right) \right|$$
$$\leq \frac{b_n}{a_n - 1} a_n \left( \frac{2D}{b_n} \psi(n) \right) \left( \frac{D}{n} \psi(n) \right)$$
$$< 2D^2 \frac{a_n}{a_n - 1} \frac{b_n}{n} \frac{\psi(n)^2 \log n}{b_n}$$

$\to 0$ w.p. 1 as $n \to \infty$ by Condition 4.5a.

6. By Corollary B.7 and (B.38)

$$\frac{b_n}{a_n - 1} \sum_{k=0}^{a_n - 1} \left| \left( \bar{V}_k^{(i)} - \bar{C}_k^{(i)} \right) \bar{C}_k^{(j)} \right|$$
$$< \frac{b_n}{a_n - 1} a_n \left( \frac{2D}{b_n} \psi(n) \right) \left( \sqrt{\frac{2\Sigma_{ii}}{b_n}} (1 + \epsilon) \left( \log \frac{n}{b_n} + \log \log n \right)^{1/2} \right)$$
$$< 2^{3/2} \Sigma_{ii}^{1/2} D (1 + \epsilon) \frac{a_n}{a_n - 1} \frac{\psi(n)}{\sqrt{b_n}} (2 \log n)^{1/2}$$

$\to 0$ w.p. 1 as $n \to \infty$ by Condition 4.5a.

7. By Corollary B.7 and (B.39)

$$\frac{b_n}{a_n - 1} \sum_{k=0}^{a_n - 1} \left| \left( \bar{V}_k^{(j)} - \bar{C}_k^{(j)} \right) \bar{C}_k^{(i)} \right|$$
$$< \frac{b_n}{a_n - 1} a_n \left( \frac{2D}{b_n} \psi(n) \right) \left( \sqrt{\frac{2\Sigma_{ii}}{b_n}} (1 + \epsilon) \left( \log \frac{n}{b_n} + \log \log n \right)^{1/2} \right)$$

$$< 2^{3/2}\Sigma_{ii}^{1/2} D(1+\epsilon)\frac{a_n}{a_n-1}\frac{\psi(n)}{\sqrt{b_n}}(2\log n)^{1/2}$$

$\to 0$ w.p. 1 as $n \to \infty$ by Condition a.

8. By Corollary B.6 and (B.38)

$$\begin{aligned}
&\frac{b_n}{a_n-1}\sum_{k=0}^{a_n-1}\left|\left(\bar{V}_k^{(i)} - \bar{C}_k^{(i)}\right)\bar{C}^{(j)}(n)\right| \\
&< \frac{b_n}{a_n-1}a_n\left(\frac{2D}{b_n}\psi(n)\right)\left(\frac{1}{n}(1+\epsilon)\left(2\Sigma_{ii}n\log\log n\right)^{1/2}\right) \\
&< 2^{3/2}D\sqrt{\Sigma_{ii}}(1+\epsilon)\frac{a_n}{a_n-1}\frac{\psi(n)(\log n)^{1/2}}{n^{1/2}} \\
&= 2^{3/2}D\sqrt{\Sigma_{ii}}(1+\epsilon)\frac{a_n}{a_n-1}\left(\frac{b_n}{n}\right)^{1/2}\frac{\psi(n)(\log n)^{1/2}}{b_n^{1/2}} \\
&\to 0 \text{ w.p. 1 as } n \to \infty \text{ by Condition 4.5a.}
\end{aligned}$$

9. By Corollary B.6 and (B.39)

$$\begin{aligned}
&\frac{b_n}{a_n-1}\sum_{k=0}^{a_n-1}\left|\left(\bar{V}_k^{(j)} - \bar{C}_k^{(j)}\right)\bar{C}^{(i)}(n)\right| \\
&< \frac{b_n}{a_n-1}a_n\left(\frac{2D}{b_n}\psi(n)\right)\left(\frac{1}{n}(1+\epsilon)\left(2\Sigma_{ii}n\log\log n\right)^{1/2}\right) \\
&< 2^{3/2}D\sqrt{\Sigma_{ii}}(1+\epsilon)\frac{a_n}{a_n-1}\frac{\psi(n)(\log n)^{1/2}}{n^{1/2}} \\
&= 2^{3/2}D\sqrt{\Sigma_{ii}}(1+\epsilon)\frac{a_n}{a_n-1}\left(\frac{b_n}{n}\right)^{1/2}\frac{\psi(n)(\log n)^{1/2}}{b_n^{1/2}} \\
&\to 0 \text{ w.p. 1 as } n \to \infty \text{ by Condition 4.5a.}
\end{aligned}$$

10. By (B.40) and Corollary B.7

$$\begin{aligned}
&\frac{b_n}{a_n-1}\sum_{k=0}^{a_n-1}\left|\left(\bar{V}^{(i)}(n)-\bar{C}^{(i)}(n)\right)\bar{C}_k^{(j)}\right| \\
&< \frac{b_n}{a_n-1}a_n\left(\frac{D}{n}\psi(n)\right)\left(\sqrt{\frac{2\Sigma_{ii}}{b_n}}(1+\epsilon)\left(\log\frac{n}{b_n}+\log\log n\right)^{1/2}\right) \\
&< \sqrt{2\Sigma_{ii}}D(1+\epsilon)\frac{a_n}{a_n-1}\frac{b_n}{n}\psi(n)\left(\frac{2}{b_n}\log n\right)^{1/2} \\
&= 2^{3/2}\Sigma_{ii}^{1/2}D(1+\epsilon)\frac{a_n}{a_n-1}\frac{b_n}{n}\frac{\psi(n)(\log n)^{1/2}}{b_n^{1/2}} \\
&\to 0 \text{ w.p. 1 as } n\to\infty \text{ by Condition 4.5a.}
\end{aligned}$$

11. By (B.41) and Corollary B.7

$$\begin{aligned}
&\frac{b_n}{a_n-1}\sum_{k=0}^{a_n-1}\left|\left(\bar{V}^{(j)}(n)-\bar{C}^{(j)}(n)\right)\bar{C}_k^{(i)}\right| \\
&< \frac{b_n}{a_n-1}a_n\left(D\frac{\psi(n)}{n}\right)\left(\sqrt{\frac{2\Sigma_{ii}}{b_n}}(1+\epsilon)\left(\log\frac{n}{b_n}+\log\log n\right)^{1/2}\right) \\
&< \sqrt{2\Sigma_{ii}}D(1+\epsilon)\frac{a_n}{a_n-1}\frac{b_n}{n}\psi(n)\left(\frac{2}{b_n}\log n\right)^{1/2} \\
&= 2^{3/2}\Sigma_{ii}^{1/2}D(1+\epsilon)\frac{a_n}{a_n-1}\frac{b_n}{n}\frac{\psi(n)(\log n)^{1/2}}{b_n^{1/2}} \\
&\to 0 \text{ w.p. 1 as } n\to\infty \text{ by Condition 4.5a.}
\end{aligned}$$

12. By (B.40) and Corollary B.6

$$\begin{aligned}
&\frac{b_n}{a_n-1}\sum_{k=0}^{a_n-1}\left|\left(\bar{V}^{(i)}(n)-\bar{C}^{(i)}(n)\right)\bar{C}^{(j)}(n)\right| \\
&< \frac{b_n}{a_n-1}a_n\left(\frac{D}{n}\psi(n)\right)\left(\frac{1}{n}(1+\epsilon)(2\Sigma_{ii}n\log\log n)^{1/2}\right)
\end{aligned}$$

$$< \sqrt{2\Sigma_{ii}} D(1+\epsilon) \frac{a_n}{a_n - 1} \frac{b_n}{n} \frac{\psi(n)(\log n)^{1/2}}{n^{1/2}}$$

$$\to 0 \text{ w.p. } 1 \text{ as } n \to \infty \text{ by Condition 4.5a.}$$

13. By (B.41) and Corollary B.6

$$\frac{b_n}{a_n - 1} \sum_{k=0}^{a_n - 1} \left| \left( \bar{V}^{(j)}(n) - \bar{C}^{(j)}(n) \right) \bar{C}^{(i)}(n) \right|$$

$$< \frac{b_n}{a_n - 1} a_n \left( D \frac{\psi(n)}{n} \right) \left( \frac{1}{n} (1+\epsilon)(2\Sigma_{ii} n \log\log n)^{1/2} \right)$$

$$< \sqrt{2\Sigma_{ii}} D(1+\epsilon) \frac{a_n}{a_n - 1} \frac{b_n}{n} \frac{\psi(n)(\log n)^{1/2}}{n^{1/2}}$$

$$\to 0 \text{ w.p. } 1 \text{ as } n \to \infty \text{ by Condition 4.5a.}$$

Thus, each of the 13 terms tends to 0 with probability 1 as $n \to \infty$, giving that $\Sigma_{BM,ij} \to \Sigma_{ij}$ w.p. 1 as $n \to \infty$.

# Appendix C

# Proof of Theorem 5.1

Without loss of generality, we assume $\tau^2 = 1$. The posterior distribution for this Bayesian logistic regression model is,

$$\begin{aligned} f(\beta|y, X) &\propto f(\beta) \prod_{i=1}^{K} f(y_i|X_i, \beta) \\ &\propto e^{-\frac{1}{2}\beta^T\beta} \prod_{i=1}^{K} \left( \frac{1}{1+e^{-X_i\beta}} \right)^{y_i} \left( \frac{e^{-X_i\beta}}{1+e^{-X_i\beta}} \right)^{1-y_i} . \end{aligned} \tag{C.1}$$

For simpler notation we will use $f(\beta)$ to denote the posterior density. Note that the posterior has a moment generating function. Recall that a random walk Metropolis-Hastings algorithm with a multivariate normal proposal distribution is used to sample from the posterior $f(\beta)$. We will use the following result to establish geometric ergodicity of this Markov chain.

**Theorem C.1 (Jarner and Hansen (2000))** Let $m(\beta) = \nabla f(\beta)/\|\nabla f(\beta)\|$ and also let $n(\beta) = \beta/\|\beta\|$. Suppose $f$ on $\mathbb{R}^p$ is super-exponential in that it is positive and has continuous first derivatives such that

$$\lim_{\|\beta\|\to\infty} n(\beta) \cdot \nabla \log f(\beta) = -\infty . \tag{C.2}$$

In addition let the proposal distribution be bounded away from 0 in some region around zero. If

$$\limsup_{\|\beta\|\to\infty} n(\beta)\cdot m(\beta) < 0\,, \tag{C.3}$$

then the random walk Metropolis-Hastings algorithm is geometrically ergodic. □

Note that the multivariate normal proposal distribution is indeed bounded away from zero in some region around zero. We will first show that $f$ is super-exponential. It is easy to see that $f$ has continuous first derivatives and is positive. Next we need to establish (C.2). From (C.1) we see that

$$\begin{aligned}
&\log f(\beta)\\
&= \text{const} - \frac{1}{2}\beta^T\beta - \sum_{i=1}^{K} y_i \log(1+e^{-X_i\beta}) - \sum_{i=1}^{K}(1-y_i)X_i\beta\\
&\quad - \sum_{i=1}^{K}(1-y_i)\log(1+e^{-X_i\beta})\\
&= \text{const} - \frac{1}{2}\beta^T\beta - \sum_{i=1}^{K} \log(1+e^{-X_i\beta}) - \sum_{i=1}^{K}(1-y_i)X_i\beta\\
&= \text{const} - \frac{1}{2}\sum_{j=1}^{p}\beta_j^2 - \sum_{i=1}^{K} \log(1+e^{-\sum_{j=1}^{p} x_{ij}\beta_j}) - \sum_{i=1}^{K}(1-y_i)\sum_{j=1}^{p} x_{ij}\beta_j\,.
\end{aligned}$$

For $l = 1,\ldots,p$

$$\frac{\partial \log f(\beta)}{\partial \beta_l} = -\beta_l + \sum_{i=1}^{K}\frac{x_{il}e^{-X_i\beta}}{1+e^{-X_i\beta}} - \sum_{i=1}^{K}(1-y_i)x_{il}$$

and

$$\beta\cdot\nabla\log f(\beta) = \sum_{j=1}^{p}\left[-\beta_j^2 + \sum_{i=1}^{K} x_{ij}\beta_j\frac{e^{-X_i\beta}}{1+e^{-X_i\beta}} - \sum_{i=1}^{K}(1-y_i)x_{ij}\beta_j\right]$$

$$= -\|\beta\|^2 + \sum_{i=1}^{K} X_i\beta \frac{e^{-X_i\beta}}{1+e^{-X_i\beta}} - \sum_{i=1}^{K}(1-y_i)X_i\beta\,.$$

Hence

$$\frac{\beta}{\|\beta\|}\cdot\nabla \log f(\beta) = -\|\beta\| + \sum_{i=1}^{K}\frac{X_i\beta}{\|\beta\|}\frac{e^{-X_i\beta}}{1+e^{-X_i\beta}} - \sum_{i=1}^{K}(1-y_i)\frac{X_i\beta}{\|\beta\|}.$$

Taking the limit with $\|\beta\| \to \infty$ we obtain

$$\begin{aligned}\lim_{\|\beta\|\to\infty}\frac{\beta}{\|\beta\|}\cdot\nabla \log f(\beta) = &-\lim_{\|\beta\|\to\infty}\|\beta\| + \lim_{\|\beta\|\to\infty}\sum_{i=1}^{K}\frac{X_i\beta}{\|\beta\|}\frac{e^{-X_i\beta}}{1+e^{-X_i\beta}} \\ &- \lim_{\|\beta\|\to\infty}\sum_{i=1}^{K}(1-y_i)\frac{X_i\beta}{\|\beta\|}. \qquad \text{(C.4)}\end{aligned}$$

By the Cauchy-Schwarz inequality we can bound the second term

$$\lim_{\|\beta\|\to\infty}\sum_{i=1}^{K}\frac{X_i\beta}{\|\beta\|}\frac{e^{-X_i\beta}}{1+e^{-X_i\beta}} \le \lim_{\|\beta\|\to\infty}\sum_{i=1}^{K}\frac{|X_i|\|\beta\|}{\|\beta\|}\frac{e^{-X_i\beta}}{1+e^{-X_i\beta}} \le \sum_{i=1}^{K}|X_i|. \qquad \text{(C.5)}$$

For the third term we obtain

$$\begin{aligned}&\lim_{\|\beta\|\to\infty}\sum_{i=1}^{K}(1-y_i)\frac{X_i\beta}{\|\beta\|} \\ &= \lim_{\|\beta\|\to\infty}\sum_{i=1}^{K}(1-y_i)\frac{\sum_{j=1}^{p}x_{ij}\beta_j}{\|\beta\|} \\ &= \sum_{i=1}^{K}(1-y_i)\sum_{j=1}^{p}\lim_{\|\beta\|\to\infty}\frac{x_{ij}\beta_j}{\|\beta\|} \\ &\ge \sum_{i=1}^{K}(1-y_i)\sum_{j=1}^{p}\lim_{\|\beta\|\to\infty}\frac{-|x_{ij}||\beta_j|}{\|\beta\|} \\ &\ge -\sum_{i=1}^{K}(1-y_i)\sum_{j=1}^{p}\lim_{\|\beta\|\to\infty}|x_{ij}| \qquad \text{Since } |\beta_j| \le \|\beta\|\end{aligned}$$

$$= -\sum_{i=1}^{K}(1-y_i)||X_i||_1. \tag{C.6}$$

Using (C.5) and (C.6) in (C.4).

$$\lim_{\|\beta\|\to\infty} \frac{\beta}{\|\beta\|} \cdot \nabla \log f(\beta) \leq -\lim_{\|\beta\|\to\infty} \|\beta\| + \sum_{i=1}^{K}|X_i| + \sum_{i=1}^{K}(1-y_i)||X_i||_1 = -\infty\,.$$

Next we need to establish (C.3). Notice that

$$\begin{aligned} f(\beta) &\propto \exp\left[-\frac{1}{2}\sum_{j=1}^{p}\beta_j^2 - \sum_{i=1}^{K}(1-y_i)\sum_{j=1}^{p}x_{ij}\beta_j - \sum_{i=1}^{K}\log(1+e^{-\sum_{j=1}^{p}x_{ij}\beta_j})\right] \\ &:= e^{C(\beta)} \end{aligned}$$

and hence for $l = 1, \ldots, p$

$$\frac{\partial f(\beta)}{\partial \beta_l} = e^{C(\beta)}\left[-\beta_l - \sum_{i=1}^{K}(1-y_i)x_{il} - \sum_{i=1}^{K}\frac{-x_{il}e^{-X_i\beta}}{1+e^{-X_i\beta}}\right].$$

In order to show the result, we will need evaluate

$$\lim_{\|\beta\|\to\infty} \frac{e^{C(\beta)}\|\beta\|}{\|\nabla f(\beta)\|}.$$

To this end, we will first show that

$$\lim_{\|\beta\|\to\infty} \frac{\|\nabla f(\beta)\|^2}{e^{2C(\beta)}\|\beta\|^2} = 1.$$

We calculate that

$$\begin{aligned} &\|\nabla f(\beta)\|^2 \\ &= e^{2C(\beta)}\sum_{j=1}^{p}\left[-\beta_j - \sum_{i=1}^{K}(1-y_i)x_{ij} + \sum_{i=1}^{K}\frac{x_{ij}e^{-X_i\beta}}{1+e^{-X_i\beta}}\right]^2 \end{aligned}$$

$$
\begin{aligned}
&= e^{2C(\beta)} \sum_{j=1}^{p} \left[ \left( \sum_{i=1}^{K} \frac{x_{ij} e^{-X_i\beta}}{1+e^{-X_i\beta}} \right)^2 + \beta_j^2 + \left( \sum_{i=1}^{K} (1-y_i) x_{ij} \right)^2 \right. \\
&\quad \left. + 2 \sum_{i=1}^{K} (1-y_i) x_{ij} \beta_j - 2 \left( \beta_j + \sum_{i=1}^{K} (1-y_i) x_{ij} \right) \left( \sum_{i=1}^{K} x_{ij} \frac{e^{-X_i\beta}}{1+e^{-X_i\beta}} \right) \right] \\
&= e^{2C(\beta)} \left[ \sum_{j=1}^{p} \left( \sum_{i=1}^{K} \frac{x_{ij} e^{-X_i\beta}}{1+e^{-X_i\beta}} \right)^2 + \|\beta\|^2 + \sum_{j=1}^{p} \left( \sum_{i=1}^{K} (1-y_i) x_{ij} \right)^2 \right. \\
&\quad + 2 \sum_{i=1}^{K} (1-y_i) X_i \beta - 2 \sum_{i=1}^{K} X_i \beta \frac{e^{-X_i\beta}}{1+e^{-X_i\beta}} \\
&\quad \left. - 2 \sum_{j=1}^{p} \left( \sum_{i=1}^{K} (1-y_i) X_i \right) \left( \sum_{i=1}^{K} x_{ij} \frac{e^{-X_i\beta}}{1+e^{-X_i\beta}} \right) \right] \\
&= e^{2C(\beta)} \|\beta\|^2 \left[ \frac{1}{\|\beta\|^2} \sum_{j=1}^{p} \left( \sum_{i=1}^{K} \frac{x_{ij} e^{-X_i\beta}}{1+e^{-X_i\beta}} \right)^2 + 1 + \frac{1}{\|\beta\|^2} \sum_{j=1}^{p} \left( \sum_{i=1}^{K} (1-y_i) x_{ij} \right)^2 \right. \\
&\quad + 2 \sum_{i=1}^{K} (1-y_i) \frac{X_i\beta}{\|\beta\|^2} - 2 \sum_{i=1}^{K} \frac{X_i\beta}{\|\beta\|^2} \frac{e^{-X_i\beta}}{1+e^{-X_i\beta}} \\
&\quad \left. - 2 \frac{1}{\|\beta\|^2} \sum_{j=1}^{p} \left( \sum_{i=1}^{K} (1-y_i) X_i \right) \left( \sum_{i=1}^{K} x_{ij} \frac{e^{-X_i\beta}}{1+e^{-X_i\beta}} \right) \right]
\end{aligned}
$$

and

$$
\begin{aligned}
&\frac{\|\nabla f(\beta)\|^2}{e^{2C(\beta)} \|\beta\|^2} \\
&= \frac{1}{\|\beta\|} \left[ \frac{1}{\|\beta\|} \sum_{j=1}^{p} \left( \sum_{i=1}^{K} \frac{x_{ij} e^{-X_i\beta}}{1+e^{-X_i\beta}} \right)^2 + \frac{1}{\|\beta\|} \sum_{j=1}^{p} \left( \sum_{i=1}^{K} (1-y_i) x_{ij} \right)^2 \right. \\
&\quad + 2 \sum_{i=1}^{K} (1-y_i) \frac{X_i\beta}{\|\beta\|} - 2 \sum_{i=1}^{K} \frac{X_i\beta}{\|\beta\|} \frac{e^{-X_i\beta}}{1+e^{-X_i\beta}} \\
&\quad \left. - 2 \frac{1}{\|\beta\|} \sum_{j=1}^{p} \left( \sum_{i=1}^{K} (1-y_i) X_i \right) \left( \sum_{i=1}^{K} x_{ij} \frac{e^{-X_i\beta}}{1+e^{-X_i\beta}} \right) \right] + 1 \, . \qquad \text{(C.7)}
\end{aligned}
$$

Since $\lim_{\|\beta\| \to \infty} \|\beta\|^{-1} \to 0$, it is left to show that the term in the square brackets is

bounded in the limit. Since $y_i$ and $x_{ij}$ are independent of $\beta$, and $e^{-X_i\beta}/(1+e^{-X_i\beta})$ bounded below by 0 and above by 1, it is only required to show that the third and fourth terms in the square brackets remain bounded in the limit. From (C.6) and the Cauchy-Schwarz inequality

$$-\sum_{i=1}^{K}(1-y_i)||X_i||_1 \leq 2\lim_{\|\beta\|\to\infty}\sum_{i=1}^{K}(1-y_i)\frac{X_i\beta}{\|\beta\|} \leq 2\sum_{i=1}^{K}(1-y_i)|X_i|\,.$$

In addition,

$$\begin{aligned}\sum_{i=1}^{K}\frac{X_i\beta}{\|\beta\|^2}\frac{e^{-X_i\beta}}{1+e^{-X_i\beta}} &\geq -\sum_{i=1}^{K}\sum_{j=1}^{p}\frac{|x_{ij}||\beta_j|}{\|\beta\|}\frac{e^{-X_i\beta}}{1+e^{-X_i\beta}}\\ &\geq -\sum_{i=1}^{K}\sum_{j=1}^{p}\frac{|x_{ij}||\beta_j|}{\|\beta\|}\\ &\geq -\sum_{i=1}^{K}||X_i||_1.\end{aligned}$$

Thus, by the above result and (C.5),

$$-\sum_{i=1}^{K}||x||_1 \leq \lim_{\|\beta\|\to\infty}\sum_{i=1}^{K}\frac{X_i\beta}{\|\beta\|}\frac{e^{-X_i\beta}}{1+e^{-X_i\beta}} \leq \sum_{i=1}^{K}|X_i|.$$

Using these results in (C.7),

$$\begin{aligned}&\lim_{\|\beta\|\to\infty}\frac{\|\nabla f(\beta)\|^2}{e^{2C(\beta)}\|\beta\|^2}\\ &= 1+\lim_{\|\beta\|\to\infty}\frac{1}{\|\beta\|}\left[\frac{1}{\|\beta\|}\sum_{j=1}^{p}\left(\sum_{i=1}^{K}\frac{x_{ij}e^{-X_i\beta}}{1+e^{-X_i\beta}}\right)^2+\frac{1}{\|\beta\|}\sum_{j=1}^{p}\left(\sum_{i=1}^{K}(1-y_i)x_{ij}\right)^2\right.\\ &\quad\left.+2\sum_{i=1}^{K}(1-y_i)\frac{X_i\beta}{\|\beta\|}-2\sum_{i=1}^{K}\frac{X_i\beta}{\|\beta\|}\frac{e^{-X_i\beta}}{1+e^{-X_i\beta}}\right.\end{aligned}$$

$$-2\frac{1}{\|\beta\|}\sum_{j=1}^{p}\left(\sum_{i=1}^{K}(1-y_i)X_i\right)\left(\sum_{i=1}^{K}x_{ij}\frac{e^{-X_i\beta}}{1+e^{-X_i\beta}}\right)\Bigg]$$

$$= 1\,.$$

Next observe that

$$\begin{aligned}\beta\cdot\nabla f(\beta) &= e^{C(\beta)}\sum_{j=1}^{p}\left[-\beta_j^2-\sum_{i=1}^{K}(1-y_i)x_{ij}\beta_j+\sum_{i=1}^{K}x_{ij}\beta_j\frac{e^{-X_i\beta}}{1+e^{-X_i\beta}}\right]\\ &= e^{C(\beta)}\left[-\|\beta\|^2-\sum_{i=1}^{K}(1-y_i)X_i\beta+\sum_{i=1}^{K}X_i\beta\frac{e^{-X_i\beta}}{1+e^{-X_i\beta}}\right].\end{aligned}$$

and hence

$$\frac{\beta}{\|\beta\|}\frac{\nabla f(\beta)}{\|\nabla f(\beta)\|}=\frac{e^{C(\beta)}}{\|\nabla f(\beta)\|}\left[-\|\beta\|-\sum_{i=1}^{K}(1-y_i)\frac{X_i\beta}{\|\beta\|}+\sum_{i=1}^{K}\frac{X_i\beta}{\|\beta\|}\frac{e^{-X_i\beta}}{1+e^{-X_i\beta}}\right].$$

We conclude that

$$\begin{aligned}&\lim_{\|\beta\|\to\infty}\frac{\beta}{\|\beta\|}\frac{\nabla f(\beta)}{\|\nabla f(\beta)\|}\\ &=\lim_{\|\beta\|\to\infty}\frac{e^{C(\beta)}\|\beta\|}{\|\nabla f(\beta)\|}\left[-1-\sum_{i=1}^{K}(1-y_i)\frac{X_i\beta}{\|\beta\|^2}+\sum_{i=1}^{K}\frac{X_i\beta}{\|\beta\|^2}\frac{e^{-X_i\beta}}{1+e^{-X_i\beta}}\right]\\ &=-1\end{aligned}$$

which establishes (C.3).

# Appendix D

# Proofs from Chapter 6

## D.1 BASAD Rejection Sampler

To show the result, we will first integrate $\beta$ out of the posterior distribution.

$$
\begin{aligned}
& f(\beta, \sigma^2, Z \mid y) \\
\propto & f(y \mid \beta, Z, \sigma^2) f(\beta, Z \mid \sigma^2) f(\sigma^2) \\
\propto & (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{(y - X\beta)^T(y - X\beta)}{2\sigma^2}\right\} (\sigma^2)^{-\alpha_1 - 1} \exp\left\{-\frac{\alpha_2}{\sigma^2}\right\} \\
& \times \prod_{i=1}^{p_n} \left((1 - q_n)(2\pi\sigma^2\tau_{0,n}^2)^{-\frac{1}{2}} \exp\left\{-\frac{\beta_i^2}{2\sigma^2\tau_{0,n}^2}\right\}\right)^{1 - Z_i} \\
& \times \left(q_n(2\pi\sigma^2\tau_{1,n}^2)^{-\frac{1}{2}} \exp\left\{-\frac{\beta_i^2}{2\sigma^2\tau_{1,n}^2}\right\}\right)^{Z_i} \\
\propto & \left(\sigma^2\right)^{-\frac{n}{2} - \frac{p_n}{2} - \alpha_1 - 1} \exp\left\{-\frac{\alpha_2}{\sigma^2}\right\} ((1 - q_n)\tau_{0,n}^{-1})^{p_n - \sum Z_i} (q_n\tau_{1,n}^{-1})^{\sum Z_i} \\
& \times \exp\left\{-\frac{(y - X\beta)^T(y - X\beta)}{2\sigma^2}\right\} \exp\left\{-\sum_{i=1}^{p_n} \beta_i^2 \left(\frac{1 - Z_i}{2\sigma^2\tau_{0,n}^2} + \frac{Z_i}{2\sigma^2\tau_{1,n}^2}\right)\right\}
\end{aligned}
$$

Note that,

$$
\left(\frac{1 - Z_i}{2\sigma^2\tau_{0,n}^2} + \frac{Z_i}{2\sigma^2\tau_{1,n}^2}\right) = \frac{1}{2\sigma^2\tau_{0,n}^2}\mathbb{I}_{(Z_i = 0)} + \frac{1}{2\sigma^2\tau_{1,n}^2}\mathbb{I}_{(Z_i = 1)},
$$

and thus if $D_z = \text{Diag}(\tau^2_{z_i,n})$,

$$\sum_{i=1}^{p_n} \beta_i^2 \left( \frac{1 - Z_i}{2\sigma^2 \tau^2_{0,n}} + \frac{Z_i}{2\sigma^2 \tau^2_{1,n}} \right) = \frac{\beta^T D_z^{-1} \beta}{2\sigma^2} .$$

In addition, define $V_z = (X^T X + D_z^{-1})$. Then,

$$\begin{aligned}
& f(\sigma^2, Z \mid y) \\
&= \int f(\beta, \sigma^2, Z \mid y)\, d\beta \\
&\propto \left(\sigma^2\right)^{-\frac{n}{2} - \frac{p_n}{2} - \alpha_1 - 1} \exp\left\{ -\frac{\alpha_2}{\sigma^2} \right\} ((1 - q_n)\tau_{0,n}^{-1})^{p_n - \sum Z_i} (q_n \tau_{1,n}^{-1})^{\sum Z_i} \\
&\quad \int \exp\left\{ -\frac{(y - X\beta)^T (y - X\beta) + \beta^T D_z^{-1} \beta}{2\sigma^2} \right\} d\beta \\
&\propto \left(\sigma^2\right)^{-\frac{n}{2} - \frac{p_n}{2} - \alpha_1 - 1} \exp\left\{ -\frac{\alpha_2}{\sigma^2} \right\} ((1 - q_n)\tau_{0,n}^{-1})^{p_n - \sum Z_i} (q_n \tau_{1,n}^{-1})^{\sum Z_i} \\
&\quad \int \exp\left\{ -\frac{\beta^T (X^T X + D_z^{-1}) \beta - 2 y^T X \beta + y^T y}{2\sigma^2} \right\} d\beta \\
&\propto \left(\sigma^2\right)^{-\frac{n}{2} - \frac{p_n}{2} - \alpha_1 - 1} \exp\left\{ -\frac{\alpha_2}{\sigma^2} \right\} ((1 - q_n)\tau_{0,n}^{-1})^{p_n - \sum Z_i} (q_n \tau_{1,n}^{-1})^{\sum Z_i} \\
&\quad \times \exp\left\{ -\frac{y^T y - y^T X V_z^{-1} X^T y}{2\sigma^2} \right\} \\
&\quad \times \int \exp\left\{ -\frac{\beta^T V_z \beta - 2 y^T X \beta + y^T X V_z^{-1} X^T y}{2\sigma^2} \right\} d\beta \\
&\propto \left(\sigma^2\right)^{-\frac{n}{2} - \frac{p_n}{2} - \alpha_1 - 1} \exp\left\{ -\frac{\alpha_2}{\sigma^2} \right\} ((1 - q_n)\tau_{0,n}^{-1})^{p_n - \sum Z_i} (q_n \tau_{1,n}^{-1})^{\sum Z_i} \\
&\quad \times \exp\left\{ -\frac{y^T (I_n - X V_z^{-1} X^T) y}{2\sigma^2} \right\} \\
&\quad \times \int \exp\left\{ -\frac{(\beta - V_z^{-1} X^T y)^T V_z (\beta - V_z^{-1} X^T y)}{2\sigma^2} \right\} d\beta \\
&\propto \left(\sigma^2\right)^{-\frac{n}{2} - \frac{p_n}{2} - \alpha_1 - 1} \exp\left\{ -\frac{\alpha_2}{\sigma^2} \right\} ((1 - q_n)\tau_{0,n}^{-1})^{p_n - \sum Z_i} (q_n \tau_{1,n}^{-1})^{\sum Z_i} \\
&\quad \times \exp\left\{ -\frac{y^T (I_n - X V_z^{-1} X^T) y}{2\sigma^2} \right\} \left(2\pi\sigma^2\right)^{\frac{p_n}{2}} \det(V_z)^{-\frac{1}{2}} .
\end{aligned}$$

To finish the proof, we find $f(\sigma^2|Z, y)$.

$$f(\sigma^2 \mid Z, y) \propto (\sigma^2)^{-\frac{n}{2}-\alpha_1-1} \exp\left\{-\frac{2\alpha_2 + y^T(I_n - XV_z^{-1}X^T)y}{2\sigma^2}\right\}$$
$$\sigma^2 \mid Z, y \sim IG\left(\alpha_1 + \frac{n}{2}, \frac{2\alpha_2 + y^T(I_n - XV_z^{-1}X^T)y}{2}\right) .$$

It is possible in this case to implement a sequential rejection sampler to sample from the posterior distribution. Notice

$$f(\beta, \sigma^2, Z \mid y) = f(\beta \mid \sigma^2, Z, y) f(\sigma^2 \mid Z, y) f(Z \mid y) .$$

We showed $\sigma^2 \mid Z, y \sim$ Inverse-Gamma$(\alpha_1 + n/2, (2\alpha_2 + y^T(I_n - XV_z^{-1}X^T)y)/2)$. Thus if we can draw independent samples from $f(Z \mid y)$, then we can obtain independent samples from the posterior distribution. We will use a rejections sampler to draw i.i.d samples from $Z \mid y$ using the proposal distribution of independent Bernoullis, that is, $g(Z) = \prod_{i=1}^{p_n} q_n^{Z_i}(1-q_n)^{1-Z_i}$. First let us find the marginal posterior for $Z$, $f(Z|y)$. We know that

$$f(Z \mid y) = \left(\alpha_2 + \frac{y^T(I - XV_z^{-1}X^T)y}{2}\right)^{-(n/2+\alpha_1)} \det(V_Z)^{-1/2}$$
$$\times \left(q_n\tau_{1,n}^{-1}\right)^{\sum Z_i} \left((1-q_n)\tau_{0,n}^{-1}\right)^{p_n - \sum Z_i} .$$

Note that

$$\left(\alpha_2 + \frac{y^T(I - XV_z^{-1}X^T)y}{2}\right)^{-(n/2+\alpha_1)} \leq (\alpha_2)^{-(n/2+\alpha_1)} .$$

Also, by Minskowsi determinant theorem (see Marcus and Minc (1992)),

$$\begin{aligned} \det(X^TX + D_z^{-1})^{1/2} &\geq \det(X^TX)^{1/2} + \det(D_z^{-1})^{1/2} \\ &\geq \det(D_z^{-1})^{1/2} \\ \Rightarrow \det(X^TX + D_z^{-1})^{-1/2} &\leq \det(D_z)^{1/2} . \end{aligned}$$

Let $f(Z \mid y) = k\, q(Z \mid y)$ where $k$ is the unknown normalizing. To implement a rejection sampler, we need to find a constant $C$ such that for all $Z \in \{0,1\}^p$,

$$\frac{q(Z|y)}{g(Z)} \leq C\,.$$

$$\begin{aligned}
&\frac{q(Z|y)}{g(Z)} \\
&= \left(\alpha_2 + \frac{y^T(I - XV_z^{-1}X^T)y}{2}\right)^{-(n/2+\alpha_1)} \det(V_Z)^{-1/2} \left(\tau_{1,n}^{-1}\right)^{\sum Z_i} \left(\tau_{0,n}^{-1}\right)^{p_n - \sum Z_i} \\
&\leq (\alpha_2)^{-(n/2+\alpha_1)} \sqrt{\det(D_Z)} \left(\tau_{1,n}^{-1}\right)^{\sum Z_i} \left(\tau_{0,n}^{-1}\right)^{p_n - \sum Z_i} \\
&= (\alpha_2)^{-(n/2+\alpha_1)} \left(\prod_{i=1}^{p_n} \tau_{0,n}^{1-Z_i} \tau_{1,n}^{Z_i}\right) \left(\tau_{1,n}^{-1}\right)^{\sum Z_i} \left(\tau_{0,n}^{-1}\right)^{p_n - \sum Z_i} \\
&= (\alpha_2)^{-(n/2+\alpha_1)}\,.
\end{aligned}$$

Thus, $C = (\alpha_2)^{-(n/2+\alpha_1)}$ and a rejection sampler can be implemented in theory. However, since the marginal posterior for the $Z$ is not factorizable to the individual components, each update involves accepting or rejecting a $p$ dimensional vector, and for large $p$ (even as small as 3) the acceptance probability is (very) low. Thus, a rejection sampler is not practically feasible.

## D.2 LDA Full Conditionals

In this section we derive the conditional distributions required for the collapsed Gibbs sampler. Recall that notation $n_{ijk}$ refers to the number of times word $j$ has been assigned to topic $k$ in document $i$. Summing over various indices give quantities that

we will find useful.

$$
\begin{aligned}
n_{ijk} &= \sum_{t=1}^{N_i} \mathbb{I}(z_{it} = k, w_{it} = j) \\
&= \text{number of times word } j \text{ is assigned to topic } k \text{ in document } i. \\
n_{i\cdot k} &= \sum_{j=1}^{W} n_{ijk} = \text{number of words assigned topic } k \text{ in document } i. \\
n_{\cdot jk} &= \sum_{i=1}^{D} n_{ijk} = \text{number of times word } j \text{ is assigned topic } k \text{ for all documents.} \\
n_{\cdot\cdot k} &= \sum_{i=1}^{D} \sum_{j=1}^{N_i} n_{ijk} \\
&= \text{number of times topic } k \text{ assigned to any word over all documents.}
\end{aligned}
$$

$$
\begin{aligned}
f(\theta \mid z, w, \phi, a, b) &= \frac{f(w \mid z, \phi) f(z \mid \theta) f(\phi \mid b) f(\theta \mid a)}{f(w \mid a, b)} \\
&= f(z \mid \theta) f(\theta \mid a) \frac{f(w \mid z, \phi) f(\phi \mid b)}{f(w, z \mid a, b)} \\
&\propto f(z \mid \theta) f(\theta \mid a) \\
&= \left( \prod_{i=1}^{D} \prod_{j=1}^{N_i} f(z_{ij} \mid \theta_i) \right) \left( \prod_{i=1}^{D} f(\theta_i \mid a) \right) \\
&= \left( \prod_{i=1}^{D} \prod_{k=1}^{K} \theta_{ik}^{n_{i\cdot k}} \right) \left( \prod_{i=1}^{D} \frac{\Gamma(\sum_{k=1}^{K} a_k)}{\prod_{k=1}^{K} \Gamma(a_k)} \prod_{k=1}^{K} \theta_{ik}^{a_k - 1} \right) \\
&= \prod_{i=1}^{D} \frac{\Gamma(\sum_{k=1}^{K} a_k)}{\prod_{k=1}^{K} \Gamma(a_k)} \prod_{k=1}^{K} \theta_{ik}^{n_{i\cdot k} + a_k - 1} \\
&= \prod_{i=1}^{D} \prod_{k=1}^{K} \theta_{ik}^{n_{i\cdot k + a_k} - 1}.
\end{aligned}
$$

Thus,

$$\theta_i \mid z, w, \phi, a, b \overset{ind}{\sim} \text{Dirichlet}\,(a') \quad \text{where } a'_k = n_{i\cdot k} + a_k. \tag{D.1}$$

Thus, the mean of $\theta$ with respect to the full conditional is

$$g_\theta(z, w) := \mathrm{E}[\theta_{ik} \mid z, w, \phi, a, b] = \frac{n_{i\cdot k} + a_k}{n_{i\cdot\cdot} + \sum_{k=1}^K a_k}.$$

Similarly, we find the full conditional for $\phi$.

$$\begin{aligned} f(\phi \mid z, w, \theta, a, b) &\propto f(w \mid z, \phi) f(\phi \mid b) \\ &= \left( \prod_{i=1}^D \prod_{j=1}^{N_i} \phi_{z_{it} w_{it}} \right) \left( \prod_{k=1}^K \frac{\Gamma(\sum_{t=1}^W b_t)}{\prod_{t=1}^W \Gamma(b_t)} \prod_{t=1}^W \phi_{kt}^{b_t - 1} \right) \\ &\propto \prod_{k=1}^K \left( \prod_{i=1}^D \prod_{j=1}^{N_i} \phi_{kt}^{\mathbb{I}(z_{ij}=k, w_{ij}=t)} \right) \left( \prod_{t=1}^W \phi_{kt}^{b_t - 1} \right) \\ &\propto \prod_{k=1}^K \prod_{t=1}^W \phi_{kt}^{n_{\cdot tk} + b_t - 1}. \end{aligned}$$

Thus,

$$\phi_k \mid z, w, \theta, a, b \sim \text{Dirichlet}(b') \quad \text{where } b'_t = n_{\cdot tk} + b_t - 1. \tag{D.2}$$

As before, the mean of $\phi$ with respect to the full conditional is

$$g_\phi(z, w) := \mathrm{E}[\phi_{kt} \mid z, w, \theta, a, b] = \frac{n_{\cdot tk} + b_t}{n_{\cdot\cdot k} + \sum_{t=1}^W b_t}.$$

For the collapsed Gibbs sampler, we first need to integrate out $\theta$ and $\phi$ from the joint posterior.

$$\begin{aligned} q(z, \theta, \phi \mid w, a, b) &\propto f(w \mid z, \phi) f(\phi \mid b) \cdot f(z \mid \theta) f(\theta \mid a) \\ q(z, \theta \mid w, a, b) &= f(z \mid \theta) f(\theta \mid a) \int f(w \mid z, \phi) f(\phi \mid b)\, d\phi \end{aligned}$$

$$
\begin{aligned}
&\propto f(z \mid \theta) f(\theta \mid a) \int \prod_{k=1}^{K} \prod_{t=1}^{W} \phi_{kt}^{n_{\cdot tk}+b_t-1} \, d\phi \\
&= f(z \mid \theta) f(\theta \mid a) \prod_{k=1}^{K} \frac{\prod_{t=1}^{W} \Gamma(n_{\cdot tk}+b_t)}{\Gamma(\sum_{t=1}^{W}(n_{\cdot tk}+b_t))} \\
&\quad \times \int \frac{\Gamma(\sum_{t=1}^{W}(n_{\cdot tk}+b_t))}{\prod_{t=1}^{W}\Gamma(n_{\cdot tk}+b_t)} \prod_{t=1}^{W} \phi_{kt}^{n_{\cdot tk}+b_t-1} \, d\phi_k \\
&= f(z \mid \theta) f(\theta \mid a) \prod_{k=1}^{K} \frac{\prod_{t=1}^{W} \Gamma(n_{\cdot tk}+b_t)}{\Gamma(\sum_{t=1}^{W}(n_{\cdot tk}+b_t))} \\
q(z \mid w, a, b) &= \prod_{k=1}^{K} \frac{\prod_{t=1}^{W} \Gamma(n_{\cdot tk}+b_t)}{\Gamma(\sum_{t=1}^{W}(n_{\cdot tk}+b_t))} \int f(z \mid \theta) f(\theta \mid a) \, d\theta \\
&\propto \prod_{k=1}^{K} \frac{\prod_{t=1}^{W} \Gamma(n_{\cdot tk}+b_t)}{\Gamma(\sum_{t=1}^{W}(n_{\cdot tk}+b_t))} \int \prod_{i=1}^{D} \prod_{k=1}^{K} \theta_{ik}^{n_{i\cdot k}+a_k-1} d\theta \\
&= \prod_{k=1}^{K} \frac{\prod_{t=1}^{W} \Gamma(n_{\cdot tk}+b_t)}{\Gamma(\sum_{t=1}^{W}(n_{\cdot tk}+b_t))} \prod_{i=1}^{D} \frac{\prod_{k=1}^{K} \Gamma(n_{i\cdot k}+a_k)}{\Gamma(\sum_{k=1}^{K}(n_{i\cdot k}+a_k))} \\
&\quad \times \int \frac{\Gamma(\sum_{k=1}^{K}(n_{i\cdot k}+a_k))}{\prod_{k=1}^{K}\Gamma(n_{i\cdot k}+a_k)} \prod_{k=1}^{K} \theta_{ik}^{n_{i\cdot k}+a_k-1} d\theta_i \\
&= \prod_{k=1}^{K} \frac{\prod_{t=1}^{W} \Gamma(n_{\cdot tk}+b_t)}{\Gamma(\sum_{t=1}^{W}(n_{\cdot tk}+b_t))} \prod_{i=1}^{D} \frac{\prod_{k=1}^{K} \Gamma(n_{i\cdot k}+a_k)}{\Gamma(\sum_{k=1}^{K}(n_{i\cdot k}+a_k))} \\
&= \prod_{k=1}^{K} \frac{\prod_{t=1}^{W} \Gamma(n_{\cdot tk}+b_t)}{\Gamma(n_{\cdot\cdot k}+\sum_{t=1}^{W} b_t)} \prod_{i=1}^{D} \frac{\prod_{k=1}^{K} \Gamma(n_{i\cdot k}+a_k)}{\Gamma(n_{i\cdot\cdot}+\sum_{k=1}^{K} a_k)}
\end{aligned} \tag{D.3}
$$

Equation (D.3) is the joint posterior of the topic assignments to all words over all documents. To implement a Gibbs sampler such that $q(z \mid w, a, b)$ is the invariant distribution, we require full conditionals $q(z_{dw} \mid z_{-dw}, w, a, b)$, where $z_{-dw}$ denotes $z \backslash z_{dw}$. To find the full conditional, we decompose the product into terms that involve the word $dw$ and terms that do not involve the word $dw$. Note that here we use

notation $dw$ to denote word $w$ in document $d$.

$$
\begin{aligned}
q(z_{dw} \mid z_{-dw}, w, a, b) &= \left( \prod_{k=1}^{K} \prod_{t \neq b}^{W} \Gamma(n_{\cdot tk} + b_t) \right) \left( \prod_{k=1}^{K} \frac{\Gamma(n_{\cdot wk} + b_w)}{\Gamma\left(n_{\cdot\cdot k} + \sum_{t=1}^{W} b_t\right)} \right) \\
&\quad \times \left( \prod_{i \neq a}^{D} \frac{\prod_{k=1}^{K} \Gamma(n_{i \cdot k} + a_k)}{\Gamma(n_{i \cdot \cdot} + \sum_{k=1}^{K} a_k)} \right) \left( \frac{\prod_{k=1}^{K} \Gamma(n_{d \cdot k} + a_k)}{\Gamma\left(n_{d \cdot \cdot} + \sum_{k=1}^{K} a_k\right)} \right) \\
&\propto \left( \prod_{k=1}^{K} \frac{\Gamma(n_{\cdot wk} + b_w)}{\Gamma\left(n_{\cdot\cdot k} + \sum_{t=1}^{W} b_t\right)} \right) \left( \frac{\prod_{k=1}^{K} \Gamma(n_{d \cdot k} + a_k)}{\Gamma\left(n_{d \cdot \cdot} + \sum_{k=1}^{K} a_k\right)} \right) \\
&= \prod_{k=1}^{K} \frac{\Gamma(n_{\cdot wk} + b_w)}{\Gamma\left(n_{\cdot\cdot k} + \sum_{t=1}^{W} b_t\right)} \frac{\Gamma(n_{d \cdot k} + a_k)}{\Gamma\left(n_{d \cdot \cdot} + \sum_{k=1}^{K} a_k\right)}
\end{aligned}
$$

We want to know that given $z_{-dw}, w, a, b$ what is the probability that $z_{dw}$ is the topic $k$. Let $n_{-}$ refer to the various counts used, and $n_{-}^{-ab}$ refer to the counts excluding the word $dw$. Note that if word $dw$ is not assigned topic $k$, then $n_{-}^{-dw} = n_{-}$ and if word $dw$ is assigned topic $k$, then $n_{-} = n_{-}^{-dw} + 1$.

$$
\begin{aligned}
q(z_{dw} \mid z_{-dw}, w, a, b) &= \left( \prod_{k \neq z_{dw}}^{K} \frac{\Gamma(n_{\cdot wk} + b_w)}{\Gamma\left(n_{\cdot\cdot k} + \sum_{t=1}^{W} b_t\right)} \frac{\Gamma(n_{d \cdot k} + a_k)}{\Gamma\left(n_{d \cdot \cdot} + \sum_{k=1}^{K} a_k\right)} \right) \\
&\quad \times \left( \frac{\Gamma(n_{\cdot w z_{dw}} + b_w)}{\Gamma\left(n_{\cdot\cdot z_{dw}} + \sum_{t=1}^{W} b_t\right)} \frac{\Gamma(n_{d \cdot z_{dw}} + a_{z_{dw}})}{\Gamma\left(n_{d \cdot \cdot} + \sum_{k=1}^{K} a_k\right)} \right) \\
&\propto \left( \frac{\Gamma(n_{\cdot b z_{dw}}^{-dw} + b_w + 1)}{\Gamma\left(n_{\cdot\cdot z_{dw}}^{-dw} + \sum_{t=1}^{W} b_t + 1\right)} \frac{\Gamma(n_{d \cdot z_{dw}}^{-dw} + a_{z_{dw}} + 1)}{\Gamma\left(n_{w \cdot \cdot}^{-dw} + \sum_{k=1}^{K} a_k + 1\right)} \right) \\
&= \frac{n_{\cdot w z_{dw}}^{-dw} + b_w}{n_{\cdot\cdot z_{dw}} + \sum_{t=1}^{W} b_t} \frac{n_{d \cdot z_{dw}}^{-dw} + a_{z_{dw}}}{n_{d \cdot \cdot}^{-dw} + \sum_{k=1}^{K} a_k} \\
&\propto \frac{\left(n_{\cdot w z_{dw}}^{-dw} + b_w\right)\left(n_{d \cdot z_{dw}}^{-dw} + a_{z_{dw}}\right)}{n_{\cdot\cdot z_{dw}} + \sum_{t=1}^{W} b_t}
\end{aligned}
$$

Thus for the $j$th word in the $i$th document, the Gibbs sampler updates the topic assignment according to

$$p(z_{dw} = k \mid z_{-dw}, w, a, b) = \frac{\frac{\left(n_{\cdot wk}^{-dw} + b_w\right)\left(n_{d\cdot k}^{-dw} + a_k\right)}{n_{\cdot\cdot k} + \sum_{t=1}^{W} b_t}}{\sum_{k=1}^{K} \frac{\left(n_{\cdot wk}^{-dw} + b_j\right)\left(n_{d\cdot k}^{-dw} + a_k\right)}{n_{\cdot\cdot k} + \sum_{t=1}^{W} b_t}}.$$

## D.3 Preliminaries

In general, $\mathrm{E}_k$ represents expectation with respect to the Markov chain transition density, $k$ for that problem. Expectations with respect to a full conditional of a variable are denoted by $\mathrm{E}_{(\cdot)}$. The index 0 on variables denotes starting values for the Markov chain.

Below are some properties of known distributions that will be used often.

- If $1/X \sim$ Inverse-Gaussian$(a, b)$, then $\mathrm{E}[X] = 1/a + 1/b$.
- If $X \sim N_p(\mu, \Sigma)$, then $\mathrm{E}[XX^T] = \Sigma + \mu\mu^T$.
- If $X \sim$ Inverse-Gamma$(a, b)$, then $\mathrm{E}[X] = b/(a-1)$.
- If $X \sim$ Inverse-Gamma$(a, b)$, then $\mathrm{E}[1/X] = a/b$.

We present some results that will used in the proofs of geometric ergodicity for all three samplers.

**Lemma D.1** Let $y, X$, and $\beta$ be the observed $n \times 1$ response, the $n \times p$ matrix of covariates and the $p \times 1$ vector of regression coefficients. Let $\Sigma$ be the $p \times p$ positive definite matrix such that for $\sigma^2 > 0$,

$$\beta \sim N_p\left((X^TX + \Sigma^{-1})^{-1}X^Ty, \sigma^2(X^TX + \Sigma^{-1})^{-1}\right),$$

then

$$\mathrm{E}\left[(y - X\beta)^T(y - X\beta) + \beta^T\Sigma^{-1}\beta\right] \leq y^Ty + p\sigma^2. \qquad \square$$

**Proof**

This result was used in Khare and Hobert (2013) for a particular $\Sigma$. The proof here is similar but is presented for completeness.

$$\begin{aligned}
&\mathrm{E}\left[(y - X\beta)^T(y - X\beta) + \beta^T\Sigma^{-1}\beta\right] \\
&= y^Ty + \mathrm{E}\left[\beta^TX^TX\beta - 2y^TX\beta + \beta^T\Sigma^{-1}\beta\right] \\
&= y^Ty - 2y^TX\,\mathrm{E}\left[\beta\right] + \mathrm{E}\left[\beta^T(X^TX + \Sigma^{-1})\beta\right] \\
&= y^Ty - 2y^TX(X^TX + \Sigma^{-1})^{-1}X^Ty + \mathrm{E}\left[tr(\beta^T(X^TX + \Sigma^{-1})\beta)\right] \\
&= y^Ty - 2y^TX(X^TX + \Sigma^{-1})^{-1}X^Ty + \mathrm{E}\left[tr((X^TX + \Sigma^{-1})\beta\beta^T)\right] \\
&= y^Ty - 2y^TX(X^TX + \Sigma^{-1})^{-1}X^Ty + \ tr\left((X^TX + \Sigma^{-1})\mathrm{E}\left[\beta\beta^T\right]\right) \\
&= y^Ty - 2y^TX(X^TX + \Sigma^{-1})^{-1}X^Ty + \ tr\left(\sigma^2(X^TX + \Sigma^{-1})(X^TX + \Sigma^{-1})^{-1}\right) \\
&\quad + \ tr\left((X^TX + \Sigma^{-1})(X^TX + \Sigma^{-1})^{-1}X^Tyy^TX(X^TX + \Sigma^{-1})^{-1}\right) \\
&= y^Ty - 2y^TX(X^TX + \Sigma^{-1})^{-1}X^Ty + p\sigma^2 + \ tr\left(y^TX(X^TX + \Sigma^{-1})^{-1}X^Ty\right) \\
&= y^Ty - y^TX(X^TX + \Sigma^{-1})^{-1}X^Ty + p\sigma^2 \\
&\leq y^Ty + p\sigma^2. \qquad \square
\end{aligned}$$

**Lemma D.2** For $\alpha = (\alpha_1, \ldots, \alpha_p)$ and $\delta = (\delta_1, \ldots, \delta_p)$,

$$\frac{\sum_{i=1}^p \alpha_i^2}{\sum_{i=1}^p \alpha_i^2/\delta_i^2} \leq \sum_{i=1}^p \delta_i^2. \qquad \square$$

**Proof**

$$\frac{\sum_{i=1}^{p}\alpha_i^2}{\sum_{i=1}^{p}\alpha_i^2/\delta_i^2} = \frac{\sum_{i=1}^{p}\frac{\alpha_i^2}{\delta_i^2}\delta_i^2}{\sum_{i=1}^{p}\alpha_i^2/\delta_i^2} \leq \frac{\sum_{i=1}^{p}\frac{\alpha_i^2}{\delta_i^2}\left(\sum_{i=1}^{p}\delta_i^2\right)}{\sum_{i=1}^{p}\alpha_i^2/\delta_i^2} = \sum_{i=1}^{p}\delta_i^2. \qquad \square$$

**Lemma D.3** For $\lambda^2, a^2, \sigma^2 > 0$, if the random variable $X$ has the probability density function $f(x)$ such that

$$f(x) \propto x^{-1/2}\exp\left\{-\frac{\lambda^2 x}{2} - \frac{a^2}{2\sigma^2 x}\right\},$$

then $1/X \sim$ Inverse-Gaussian distribution with mean parameter $\sqrt{\lambda^2\sigma^2/a^2}$ and scale parameter $\lambda^2$. $\square$

**Proof**

It is given that,

$$f(x) \propto x^{-\frac{1}{2}}\exp\left\{-\frac{\lambda^2 x}{2} - \frac{a^2}{2\sigma^2 x}\right\}.$$

For the change of variable $z = 1/x$

$$\begin{aligned} f(z) &\propto z^{-2}z^{\frac{1}{2}}\exp\left\{-\frac{\lambda^2}{2z} - \frac{a^2 z}{2\sigma^2}\right\} \\ &= z^{-\frac{3}{2}}\exp\left\{-\frac{\lambda^2\sigma^2 + a^2 z^2}{2\sigma^2 z}\right\} \\ &= z^{-\frac{3}{2}}\exp\left\{-\frac{a^2\left(\frac{\lambda^2\sigma^2}{a^2} + z^2\right)}{2\sigma^2 z}\right\} \end{aligned}$$

$$= \exp\left\{-\sqrt{\frac{\lambda^2 a^2}{\sigma^2}}\right\} z^{-\frac{3}{2}} \exp\left\{-\frac{a^2\left(\frac{\lambda^2\sigma^2}{a^2} - 2\sqrt{\frac{\lambda^2\sigma^2}{a^2}}z + z^2\right)}{2\sigma^2 z}\right\}$$

$$\propto z^{-\frac{3}{2}} \exp\left\{-\frac{\lambda^2\left(\frac{\lambda^2\sigma^2}{a^2} - 2\sqrt{\frac{\lambda^2\sigma^2}{a^2}}z + z^2\right)}{2\frac{\lambda^2\sigma^2}{a^2}z}\right\}.$$

Thus, $Z \sim$ Inverse-Gaussian with mean parameter $\sqrt{\lambda^2\sigma^2/a^2}$ and scale parameter $\lambda^2$. □

**Lemma D.4** If $1/X \sim$ Inverse-Gaussian with mean parameter $\sqrt{\lambda^2\sigma^2/a^2}$ and scale parameter $\lambda^2$ and $a^2 \leq d^2$ for some $d^2 > 0$, then

$$f(x) \geq \exp\left\{-\sqrt{\frac{\lambda^2 d^2}{\sigma^2}}\right\} q(x),$$

where $f(x)$ is the probability density function of $X$ and $q(x)$ is the probability density function of the reciprocal Inverse-Gaussian distribution with mean parameter $\sqrt{\lambda^2\sigma^2/d^2}$ and scale parameter $\lambda^2$. □

**Proof**

By Lemma D.3, we have

$$f(x)$$

$$= \sqrt{\frac{\lambda^2}{2\pi}}\,(x)^{-\frac{1}{2}} \exp\left\{-\frac{\lambda^2\left(\frac{\lambda^2\sigma^2}{a^2} - 2\sqrt{\frac{\lambda^2\sigma^2}{a^2}}\frac{1}{x} + \frac{1}{x^2}\right)}{2\frac{\lambda^2\sigma^2}{a^2}\frac{1}{x}}\right\}$$

$$= \exp\left\{\sqrt{\frac{\lambda^2 a^2}{\sigma^2}}\right\} \sqrt{\frac{\lambda^2}{2\pi}}\,(x)^{-\frac{1}{2}} \exp\left\{-\frac{\lambda^2\left(\frac{\lambda^2\sigma^2}{a^2} + \frac{1}{x^2}\right)}{2\frac{\lambda^2\sigma^2}{a^2}\frac{1}{x}}\right\}$$

$$\geq \sqrt{\frac{\lambda^2}{2\pi}}\,(x)^{-\frac{1}{2}} \exp\left\{ -\frac{\lambda^2\left(\frac{\lambda^2\sigma^2}{a^2}+\frac{1}{x^2}\right)}{2\frac{\lambda^2\sigma^2}{a^2}\frac{1}{x}} \right\}$$

$$= \sqrt{\frac{\lambda^2}{2\pi}}\,(x)^{-\frac{1}{2}} \exp\left\{ -\frac{\lambda^2\left(\lambda^2\sigma^2+\frac{a^2}{x^2}\right)}{2\lambda^2\sigma^2\frac{1}{x}} \right\}$$

$$\geq \sqrt{\frac{\lambda^2}{2\pi}}\,(x)^{-\frac{1}{2}} \exp\left\{ -\frac{\lambda^2\left(\lambda^2\sigma^2+\frac{d^2}{x^2}\right)}{2\lambda^2\sigma^2\frac{1}{x}} \right\}$$

$$= \exp\left\{ -\sqrt{\frac{\lambda^2 d^2}{\sigma^2}} \right\} \exp\left\{ \sqrt{\frac{\lambda^2 d^2}{\sigma^2}} \right\} \sqrt{\frac{\lambda^2}{2\pi}}\,(x)^{-\frac{1}{2}} \exp\left\{ -\frac{\lambda^2\left(\frac{\lambda^2\sigma^2}{d^2}+\frac{1}{x^2}\right)}{2\frac{\lambda^2\sigma^2}{d^2}\frac{1}{x}} \right\}$$

$$= \exp\left\{ -\sqrt{\frac{\lambda^2 d^2}{\sigma^2}} \right\} \sqrt{\frac{\lambda^2}{2\pi}}\,(x)^{-\frac{1}{2}} \exp\left\{ -\frac{\lambda^2\left(\frac{\lambda^2\sigma^2}{d^2}-2\sqrt{\frac{\lambda^2\sigma^2}{d^2}}\frac{1}{x}+\frac{1}{x^2}\right)}{2\frac{\lambda^2\sigma^2}{d^2}\frac{1}{x}} \right\}$$

$$= \exp\left\{ -\sqrt{\frac{\lambda^2 d^2}{\sigma^2}} \right\} q(x). \qquad \square$$

**Lemma D.5** Let $y, X$, and $\beta$ be the observed $n \times 1$ response, the $n \times p$ matrix of covariates and the $p \times 1$ vector of regression coefficients respectively. Let $\Sigma$ be thae $p \times p$ positive definite matrix. Then,

$$(y - X\beta)^T(y - X\beta) + \beta^T\Sigma^{-1}\beta \geq y^T y - y^T X\left(X^T X + \Sigma^{-1}\right)^{-1} X^T y. \qquad \square$$

**Proof**

The proof mainly requires completing the square in the following way.

$$
\begin{aligned}
&(y - X\beta)^T(y - X\beta) + \beta^T\Sigma^{-1}\beta \\
&= y^Ty - 2y^TX\beta + \beta^T(X^TX + \Sigma^{-1})\beta \\
&= y^Ty - 2y^TX(X^TX + \Sigma^{-1})(X^TX + \Sigma^{-1})^{-1}\beta + \beta^T(X^TX + \Sigma^{-1})\beta \\
&\quad + y^TX(X^TX + \Sigma^{-1})^{-1}(X^TX + \Sigma^{-1})(X^TX + \Sigma^{-1})^{-1}X^Ty \\
&\quad - y^TX(X^TX + \Sigma^{-1})^{-1}(X^TX + \Sigma^{-1})(X^TX + \Sigma^{-1})^{-1}X^Ty \\
&= y^Ty - y^TX(X^TX + \Sigma^{-1})^{-1}X^Ty \\
&\quad + (\beta - (X^TX + \Sigma^{-1})^{-1}X^Ty)^T(X^TX + \Sigma^{-1})(\beta - (X^TX + \Sigma^{-1})^{-1}X^Ty) \\
&\geq y^Ty - y^TX(X^TX + \Sigma^{-1})^{-1}X^Ty. \qquad \square
\end{aligned}
$$

## D.4 Bayesian Fused Lasso Prior

In this section we show that the Bayesian fused lasso prior is proper and that the resulting posterior is the one desired.

First note that $\det(\Sigma_\beta) = \det\left(\Sigma_\beta^{-1}\right)^{-1}$. We decompose $\Sigma_\beta^{-1}$ into

$$\Sigma_\beta^{-1} = L_1 + L_2 \tag{D.4}$$

where

$$L_1 = \text{diag}\left(\frac{1}{2\tau_1^2}, \frac{1}{2\tau_2^2}, \ldots, \frac{1}{2\tau_p^2}\right) \text{ and,}$$

$$L_2 = \begin{bmatrix} \frac{1}{2\tau_1^2} + \frac{1}{w_1^2} & -\frac{1}{w_1^2} & 0 & \dots & 0 \\ -\frac{1}{w_1^2} & \frac{1}{2\tau_2^2} + \frac{1}{w_1^2} + \frac{1}{w_2^2} & -\frac{1}{w_2^2} & \dots & 0 \\ 0 & -\frac{1}{w_2^2} & \frac{1}{2\tau_3^2} + \frac{1}{w_2^2} + \frac{1}{w_3^2} & \dots & 0 \\ \dots & \dots & \dots & \ddots & \dots \\ 0 & 0 & \dots & \frac{1}{2\tau_{p-1}^2} + \frac{1}{w_{p-2}^2} + \frac{1}{w_{p-1}^2} & -\frac{1}{w_{p-1}^2} \\ 0 & 0 & \dots & -\frac{1}{w_{p-1}^2} & \frac{1}{2\tau_p^2} + \frac{1}{w_{p-1}^2} \end{bmatrix}. \tag{D.5}$$

The diagonal matrix $L_1$ is clearly positive definite. The tridiagonal matrix $L_2$ is also positive definite since $L_2$ is real symmetric, has positive diagonals, and is strictly diagonally dominant (see Theorem 1.2 in Andelić and Da Fonseca (2011)). Here the condition of strict diagonal dominance is satisfied since

$$\frac{1}{2\tau_i^2} + \frac{1}{w_{i-1}^2} + \frac{1}{w_i^2} > \frac{1}{w_{i-1}^2} + \frac{1}{w_i^2}.$$

Thus,

$$\det(\Sigma_\beta^{-1}) = \det(L_1 + L_2) \geq \det(L_1) + \det(L_2) \geq \det(L_1) = \prod_{i=1}^{p} \left( \frac{1}{2\tau_i^2} \right)$$

$$\Rightarrow \det\left(\Sigma_\beta^{-1}\right)^{-1/2} \leq \prod_{i=1}^{p} \left(2\tau_i^2\right)^{1/2}.$$

Thus, the joint prior on $(\tau^2, w^2)$ satisfies,

$$\begin{aligned} \pi(\tau^2, w^2) &\propto \det\left(\Sigma_\beta\right)^{1/2} \left( \prod_{i=1}^{p} \left(\tau_i^2\right)^{-1/2} e^{-\lambda_1 \tau_i^2/2} \right) \left( \prod_{i=1}^{p-1} \left(w_i^2\right)^{-1/2} e^{-\lambda_2 w_i^2/2} \right) \\ &\leq \prod_{i=1}^{p} \left(2\tau_i^2\right)^{1/2} \left( \prod_{i=1}^{p} \left(\tau_i^2\right)^{-1/2} e^{-\lambda_1 \tau_i^2/2} \right) \left( \prod_{i=1}^{p-1} \left(w_i^2\right)^{-1/2} e^{-\lambda_2 w_i^2/2} \right) \end{aligned}$$

$$= 2^{p/2} \left( \prod_{i=1}^{p} e^{-\lambda_1 \tau_i^2/2} \right) \left( \prod_{i=1}^{p-1} \left(w_i^2\right)^{-1/2} e^{-\lambda_2 w_i^2/2} \right).$$

The term above is the product of the density of $p$ exponentials and the density of $p-1$ Gamma distributions. Thus, the prior is proper.

Next we demonstrate that our choice of prior in the Bayesian fused lasso leads to the Laplace prior in (6.13). First we expand $\beta^T \Sigma_\beta^{-1} \beta$.

$$\begin{aligned}
&\beta^T \Sigma_\beta^{-1} \beta \\
&= \begin{bmatrix} \beta_1 \left(\frac{1}{\tau_1^2} + \frac{1}{w_1^2}\right) - \frac{\beta_2}{w_1^2} \\ -\frac{\beta_1}{w_1^2} + \beta_2 \left(\frac{1}{\tau_2^2} + \frac{1}{w_1^2} + \frac{1}{w_2^2}\right) - \frac{\beta_3}{w_2^2} \\ -\frac{\beta_2}{w_2^2} + \beta_3 \left(\frac{1}{\tau_3^2} + \frac{1}{w_2^2} + \frac{1}{w_3^2}\right) - \frac{\beta_4}{w_3^2} \\ \vdots \\ -\frac{\beta_{p-1}}{w_{p-1}^2} + \beta_p \left(\frac{1}{\tau_p^2} + \frac{1}{w_{p-1}^2}\right) \end{bmatrix}^T \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_p \end{bmatrix} \\
&= \beta_1^2 \left(\frac{1}{\tau_1^2} + \frac{1}{w_1^2}\right) - \frac{\beta_1\beta_2}{w_1^2} - \frac{\beta_1\beta_2}{w_1^2} + \beta_2^2 \left(\frac{1}{\tau_2^2} + \frac{1}{w_1^2} + \frac{1}{w_2^2}\right) - \frac{\beta_2\beta_3}{w_2^2} \\
&\quad - \frac{\beta_2\beta_3}{w_2^2} + \beta_3^2 \left(\frac{1}{\tau_3^2} + \frac{1}{w_2^2} + \frac{1}{w_3^2}\right) - \frac{\beta_3\beta_4}{w_3^2} + \cdots - \frac{\beta_{p-1}\beta_p}{w_{p-1}^2} + \beta_p^2 \left(\frac{1}{\tau_p^2} + \frac{1}{w_{p-1}^2}\right) \\
&= \sum_{i=1}^{p} \frac{\beta_i^2}{\tau_i^2} + \frac{\beta_1^2 + \beta_2^2 - 2\beta_1\beta_2}{w_1^2} + \frac{\beta_2^2 + \beta_3^2 - 2\beta_2\beta_3}{w_2^2} + \cdots + \frac{\beta_p^2 + \beta_{p-1}^2 - 2\beta_p\beta_{p-1}}{w_{p-1}^2} \\
&= \sum_{i=1}^{p} \frac{\beta_i^2}{\tau_i^2} + \sum_{i=1}^{p-1} \frac{(\beta_{i+1} - \beta_i)^2}{w_i^2}. \qquad \text{(D.6)}
\end{aligned}$$

Using (D.6),

$$\begin{aligned}
&\pi(\beta \mid \sigma^2) \\
&\propto \int_{\mathbb{R}_+^p} \int_{\mathbb{R}_+^{p-1}} (2\pi\sigma^2)^{-\frac{p}{2}} \det\left(\Sigma_\beta^{-1}\right)^{1/2} \exp\left\{ -\frac{\beta^T \Sigma_\beta^{-1} \beta}{2\sigma^2} \right\}
\end{aligned}$$

$$
\begin{aligned}
&\times \quad \det\left(\Sigma_\beta\right)^{1/2} \left(\prod_{i=1}^{p} \left(\tau_i^2\right)^{-1/2} e^{-\lambda_1 \tau_i^2/2}\right) \left(\prod_{i=1}^{p-1} \left(w_i^2\right)^{-1/2} e^{-\lambda_2 w_i^2/2}\right) dw^2 d\tau^2 \\
&\propto \int_{\mathbb{R}_+^p} \prod_{i=1}^{p} \left(\tau_i^2\right)^{-1/2} \exp\left\{-\frac{\lambda_1 \tau_i^2}{2} - \frac{\beta_i^2}{2\sigma^2 \tau_i^2}\right\} d\tau^2 \\
&\quad \times \int_{\mathbb{R}_+^{p-1}} \prod_{i=1}^{p-1} \left(w_i^2\right)^{-1/2} \exp\left\{-\frac{\lambda_2 w_i^2}{2} - \frac{\left(\beta_{i+1} - \beta_i\right)^2}{2\sigma^2 w_i^2}\right\} dw^2 \\
&= \exp\left\{-\frac{\lambda_1}{\sigma} \sum_{i=1}^{p} |\beta_i| - \frac{\lambda_2}{\sigma} \sum_{i=1}^{p-1} |\beta_{i+1} - \beta_i|\right\} \\
&\quad \times \int_{\mathbb{R}_+^p} \prod_{i=1}^{p} \left(\tau_i^2\right)^{-1/2} \exp\left\{-\frac{\lambda_1 \tau_i^2}{2} - \frac{\beta_i^2}{2\sigma^2 \tau_i^2} + \frac{\lambda_1}{\sigma} |\beta_i|\right\} d\tau^2 \\
&\quad \times \int_{\mathbb{R}_+^{p-1}} \prod_{i=1}^{p-1} \left(w_i^2\right)^{-1/2} \exp\left\{-\frac{\lambda_2 w_i^2}{2} - \frac{\left(\beta_{i+1} - \beta_i\right)^2}{2\sigma^2 w_i^2} + \frac{\lambda_2}{\sigma} |\beta_{i+1} - \beta_i|\right\} dw^2 \\
&\propto \exp\left\{-\frac{\lambda_1}{\sigma} \sum_{i=1}^{p} |\beta_i| - \frac{\lambda_2}{\sigma} \sum_{i=1}^{p-1} |\beta_{i+1} - \beta_i|\right\},
\end{aligned}
$$

where the last equality is due to the integrands being the densities of the reciprocal of Inverse-Gaussian distributions; see Lemma D.3.

## D.5 Proof of Theorem 6.1

We will establish geometric ergodicity of the three variable Gibbs sampler for the BFL by establishing a drift condition and an associated minorization condition. We first establish the drift condition. Consider the drift function

$$
V_{BFL}(\beta, \tau^2, w^2, \sigma^2) = (y - X\beta)^T (y - X\beta) + \beta^T \Sigma_\beta^{-1} \beta + \frac{\lambda_1^2}{4} \sum_{i=1}^{p} \tau_i^2 + \frac{\lambda_2^2}{4} \sum_{i=1}^{p-1} w_i^2. \quad \text{(D.7)}
$$

Then $V_{BFL} : \mathbb{R}^p \times \mathbb{R}_+^p \times \mathbb{R}_+^{p-1} \times \mathbb{R}_+ \to [0, \infty)$. To establish the drift condition we need to show that there exists a $0 < \phi_{BFL} < 1$ and $L_{BFL} > 0$ such that,

$$\mathrm{E}_k \left[ V(\beta, \tau^2, w^2, \sigma^2) \,|\, \beta_0, \tau_0^2, w_0^2, \sigma_0^2 \right] \leq \phi_{BFL} V_{BFL}(\beta_0, \tau_0^2, w_0^2, \sigma_0^2) + L_{BFL},$$

for every $(\beta_0, \tau_0^2, w_0^2, \sigma_0^2) \in \mathbb{R}^p \times \mathbb{R}_+^p \times \mathbb{R}_+^{p-1} \times \mathbb{R}_+$.

$$\begin{aligned}
&\mathrm{E}_k \left[ V_{BFL}(\beta, \tau^2, w^2, \sigma^2) \;\mid\; \beta_0, \tau_0^2, w_0^2, \sigma_0^2 \right] \\
&= \int V_{BFL}(\beta, \tau^2, w^2, \sigma^2) f(\sigma^2 \mid \beta_0, \tau_0^2, w_0^2, y) f(\tau^2, w^2 \mid \beta_0, \sigma^2, y) \\
&\quad \times f(\beta \mid \tau^2, w^2, \sigma^2, y) d\beta \, d\tau^2 \, dw^2 \, d\sigma^2 \\
&= \int_{\mathbb{R}_+} f(\sigma^2 \mid \beta_0, \tau_0^2, w_0^2, y) \int_{\mathbb{R}_+^p \times \mathbb{R}_+^{p-1}} f(\tau^2, w^2 \mid \beta_0, \sigma^2, y) \\
&\quad \times \int_{\mathbb{R}^p} V_{BFL}(\beta, \tau^2, w^2, \sigma^2) f(\beta \mid \tau^2, w^2, \sigma^2, y) d\beta \, d\tau^2 \, dw^2 \, d\sigma^2 \\
&= \mathrm{E}_{\sigma^2} \left[ \mathrm{E}_{\tau^2, w^2} \left[ \mathrm{E}_\beta \left[ V_{BFL}(\beta, \tau^2, w^2, \sigma^2) \;\mid\; \tau^2, w^2, \sigma^2, y \right] \;\mid\; \beta_0, \sigma^2, y \right] \;\mid\; \beta_0, \tau_0^2, w_0^2, y \right].
\end{aligned}$$

We will evaluate the expectations sequentially, starting with the innermost expectation. By Lemma D.1,

$$\begin{aligned}
&\mathrm{E} \left[ V_{BFL}(\beta, \tau^2, w^2, \sigma^2) \mid \tau^2, w^2, \sigma^2, y \right] \\
&= \mathrm{E} \left[ (y - X\beta)^T (y - X\beta) + \beta^T \Sigma_\beta^{-1} \beta + \frac{\lambda_1^2}{4} \sum_{i=1}^{p} \tau_i^2 + \frac{\lambda_2^2}{4} \sum_{i=1}^{p-1} w_i^2 \mid \tau^2, w^2, \sigma^2, y \right] \\
&\leq y^T y + \frac{\lambda_1^2}{4} \sum_{i=1}^{p} \tau_i^2 + \frac{\lambda_2^2}{4} \sum_{i=1}^{p-1} w_i^2 + p\sigma^2.
\end{aligned}$$

Next we move on to the expectation with respect to the full conditional of $\tau^2, w^2$.

$$\begin{aligned}
&\mathrm{E}_{\tau^2, w^2} \left[ \mathrm{E}_\beta \left[ V_{BFL}(\beta, \tau^2, w^2, \sigma^2) \;\mid\; \tau^2, w^2, \sigma^2, y \right] \;\mid\; \beta_0, \sigma^2, y \right] \\
&\leq \mathrm{E}_{\tau^2, w^2} \left[ y^T y + \frac{\lambda_1^2}{4} \sum_{i=1}^{p} \tau_i^2 + \frac{\lambda_2^2}{4} \sum_{i=1}^{p-1} w_i^2 + p\sigma^2 \mid \beta_0, \sigma^2, y \right]
\end{aligned}$$

$$= y^T y + p\sigma^2 + \frac{\lambda_1^2}{4}\sum_{i=1}^{p} \mathrm{E}_{\tau^2}\left[\tau_i^2 \mid \beta_0, \sigma^2, y\right] + \frac{\lambda_2^2}{4}\sum_{i=1}^{p-1} \mathrm{E}_{w^2}\left[w_i^2 \mid \beta_0, \sigma^2, y\right]$$
$$= y^T y + p\sigma^2 + \frac{\lambda_1^2}{4}\sum_{i=1}^{p}\left[\sqrt{\frac{\beta_{0,i}^2}{\lambda_1^2\sigma^2}} + \frac{1}{\lambda_1^2}\right] + \frac{\lambda_2^2}{4}\sum_{i=1}^{p-1}\left[\sqrt{\frac{(\beta_{0,i+1}-\beta_{0,i})^2}{\lambda_2^2\sigma^2}} + \frac{1}{\lambda_2^2}\right],$$

where the last equality is using the properties of the Inverse-Gaussian distribution mentioned in Appendix D.3. Recall that for $a, b > 0$, $2ab \leq a^2 + b^2$.

$$\begin{aligned}
&\mathrm{E}_{\tau^2,w^2}\left[\mathrm{E}_{\beta}\left[V_{BFL}(\beta, \tau^2, w^2, \sigma^2) \mid \tau^2, w^2, \sigma^2, y\right] \mid \beta_0, \sigma^2, y\right] \\
&\leq y^T y + p\sigma^2 + \frac{\lambda_1^2}{4}\sum_{i=1}^{p}\left[\frac{\beta_{0,i}^2}{2\sigma^2(n+p+2\alpha)} + \frac{(n+p+2\alpha)}{2\lambda_1^2} + \frac{1}{\lambda_1^2}\right] \\
&\quad + \frac{\lambda_2^2}{4}\sum_{i=1}^{p-1}\left[\frac{(\beta_{0,i+1}-\beta_{0,i})^2}{2\sigma^2(n+p+2\alpha)} + \frac{(n+p+2\alpha)}{2\lambda_2^2} + \frac{1}{\lambda_2^2}\right] \\
&= y^T y + p\sigma^2 + \frac{p}{4}\left(2 + (n+p+2\alpha)\right) + \frac{\lambda_1^2}{8(n+p+2\alpha)}\sum_{i=1}^{p}\frac{\beta_{0,i}^2}{\sigma^2} \\
&\quad + \frac{\lambda_2^2}{8(n+p+2\alpha)}\sum_{i=1}^{p-1}\frac{(\beta_{0,i+1}-\beta_{0,i})^2}{\sigma^2}.
\end{aligned}$$

Finally, the last expectation,

$$\begin{aligned}
&\mathrm{E}_{\sigma^2}\left[\mathrm{E}_{\tau^2,w^2}\left[\mathrm{E}_{\beta}\left[V_{BFL}(\beta, \tau^2, w^2, \sigma^2) \mid \tau^2, w^2, \sigma^2, y\right] \mid \beta_0, \sigma^2, y\right] \mid \beta_0, \tau_0^2, w_0^2, y\right] \\
&\leq y^T y + \frac{p}{4}\left(n+p+2\alpha+2\right) + p\mathrm{E}_{\sigma^2}[\sigma^2 \mid \beta_0, \tau_0^2, w_0^2, y] \\
&\quad + \frac{\lambda_1^2}{8(n+p+2\alpha)}\sum_{i=1}^{p}\mathrm{E}_{\sigma^2}\left[\frac{\beta_{0,i}^2}{\sigma^2} \mid \beta_0, \tau_0^2, w_0^2, y\right] \\
&\quad + \frac{\lambda_2^2}{8(n+p+2\alpha)}\sum_{i=1}^{p-1}\mathrm{E}_{\sigma^2}\left[\frac{(\beta_{0,i+1}-\beta_{0,i})^2}{\sigma^2} \mid \beta_0, \tau_0^2, w_0^2, y\right] \\
&\leq y^T y + \frac{p}{4}\left(n+p+2\alpha+2\right) + p\frac{(y-X\beta_0)^T(y-X\beta_0) + \beta_0^T\Sigma_{\beta_0}^{-1}\beta_0 + 2\xi}{n+p+2\alpha-2}
\end{aligned}$$

$$
\begin{aligned}
&+ \frac{\lambda_1^2}{8(n+p+2\alpha)} \sum_{i=1}^{p} \frac{(n+p+2\alpha)\beta_{0,i}^2}{(y-X\beta_0)^T(y-X\beta_0)+\beta_0^T\Sigma_{\beta_0}^{-1}\beta_0+2\xi} \\
&+ \frac{\lambda_2^2}{8(n+p+2\alpha)} \sum_{i=1}^{p-1} \frac{(n+p+2\alpha)(\beta_{0,i+1}-\beta_{0,i})^2}{(y-X\beta_0)^T(y-X\beta_0)+\beta_0^T\Sigma_{\beta_0}^{-1}\beta_0+2\xi} \\
&\leq y^Ty + p\frac{(y-X\beta_0)^T(y-X\beta_0)+\beta_0^T\Sigma_{\beta_0}^{-1}\beta_0+2\xi}{n+p+2\alpha-2} \\
&\quad + \frac{p}{4}(n+p+2\alpha+2) + \frac{\lambda_1^2}{8}\frac{\sum_{i=1}^{p}\beta_{0,i}^2}{\beta_0^T\Sigma_{\beta_0}^{-1}\beta_0} + \frac{\lambda_2^2}{8}\frac{\sum_{i=1}^{p-1}(\beta_{0,i+1}-\beta_{0,i})^2}{\beta_0^T\Sigma_{\beta_0}^{-1}\beta_0}
\end{aligned}
$$

Using (D.6),

$$
\beta_0^T\Sigma_{\beta_0}^{-1}\beta_0 \geq \sum_{i=1}^{p}\frac{\beta_{0,i}^2}{\tau_{0,i}^2} \quad \text{and} \quad \beta_0^T\Sigma_{\beta_0}^{-1}\beta_0 \geq \sum_{i=1}^{p-1}\frac{(\beta_{0,i+1}-\beta_{0,i})^2}{w_{0,i}^2}.
$$

$$
\begin{aligned}
&\mathrm{E}_{\sigma^2}\left[\mathrm{E}_{\tau^2,w^2}\left[\mathrm{E}_{\beta}\left[V_{BFL}(\beta,\tau^2,w^2,\sigma^2) \mid \tau^2,w^2,\sigma^2,y\right] \mid \beta_0,\sigma^2,y\right] \mid \beta_0,\tau_0^2,w_0^2,y\right] \\
&\leq y^Ty + p\frac{(y-X\beta_0)^T(y-X\beta_0)+\beta_0^T\Sigma_{\beta_0}^{-1}\beta_0+2\xi}{n+p+2\alpha-2} \\
&\quad + \frac{p}{4}(n+p+2\alpha+2) + \frac{\lambda_1^2}{8}\frac{\sum_{i=1}^{p}\beta_{0,i}^2}{\sum_{i=1}^{p}\beta_{0,i}^2/\tau_{0,i}^2} + \frac{\lambda_2^2}{8}\frac{\sum_{i=1}^{p-1}(\beta_{0,i+1}-\beta_{0,i})^2}{\sum_{i=1}^{p-1}(\beta_{0,i+1}-\beta_{0,i})^2/w_{0,i}^2}.
\end{aligned}
$$

By Lemma D.2,

$$
\begin{aligned}
&\mathrm{E}_{\sigma^2}\left[\mathrm{E}_{\tau^2,w^2}\left[\mathrm{E}_{\beta}\left[V_{BFL}(\beta,\tau^2,w^2,\sigma^2) \mid \tau^2,w^2,\sigma^2,y\right] \mid \beta_0,\sigma^2,y\right] \mid \beta_0,\tau_0^2,w_0^2,y\right] \\
&\leq y^Ty + \frac{p}{4}(n+p+2\alpha+2) + \frac{2p\xi}{n+p+2\alpha-2} \\
&\quad + \frac{p}{n+p+2\alpha-2}\left((y-X\beta_0)^T(y-X\beta_0)+\beta_0^T\Sigma_{\beta_0}^{-1}\beta_0\right) \\
&\quad + \frac{\lambda_1^2}{8}\sum_{i=1}^{p}\tau_{0,i}^2 + \frac{\lambda_2^2}{8}\sum_{i=1}^{p-1}w_{0,i}^2 \\
&\leq \phi_{BFL}V(\beta_0,\tau_0^2,w_0^2,\sigma_0^2) + L_{BFL},
\end{aligned}
$$

where

$$\phi_{BFL} = \max\left\{\frac{p}{n+p+2\alpha-2}, \frac{1}{2}\right\} < 1 \text{ for } n \geq 3, \tag{D.8}$$

and

$$L_{BFL} = y^T y + \frac{p}{2}(n+p+2\alpha+2) + \frac{2p\xi}{n+p+2\alpha-2}. \tag{D.9}$$

Thus we have established the drift condition and turn our attention to the minorization condition. To establish a one-step minorization, we need to show that for all sets $C_d$ defined as $C_d = \{(\beta, \tau^2, w^2, \sigma^2) : V_{BFL}(\beta, \tau^2, w^2, \sigma^2) \leq d\}$ there exists an $\epsilon > 0$ and a density $q$ such that for all $(\beta_0, \tau_0^2, w_0^2, \sigma_0^2) \in C_d$

$$k_{BFL}(\beta, \tau^2, w^2, \sigma^2 \mid \beta_0, \tau_0^2, w_0^2, \sigma_0^2) \geq \epsilon\, q(\beta, \tau^2, w^2, \sigma^2).$$

To establish this condition, recall that,

$$\begin{aligned} k_{BFL}(\beta, \tau^2, w^2, \sigma^2 \mid \beta_0, \tau_0^2, w_0^2, \sigma_0^2) = {} & f(\beta \mid \tau^2, w^2, \sigma^2, y)\; f(\tau^2, w^2 \mid \beta_0, \sigma^2, y) \\ & \times f(\sigma^2 \mid \beta_0, \tau_0^2, w_0^2, y). \end{aligned}$$

By our definition of the drift function, for all $(\beta_0, \tau_0^2, w_0^2, \sigma_0^2) \in C_d$ the following relation holds due to (D.6),

$$\begin{aligned} d &\geq (y - X\beta_0)^T(y - X\beta_0) + \beta_0^T \Sigma_{\beta_0}^{-1} \beta_0 + \frac{\lambda_1^2}{4}\sum_{i=1}^{p} \tau_{0,i}^2 + \frac{\lambda_2^2}{4}\sum_{i=1}^{p-1} w_{0,i}^2 \\ d &\geq (y - X\beta_0)^T(y - X\beta_0) + \sum_{i=1}^{p} \frac{\beta_{0,i}^2}{\tau_{0,i}^2} + \sum_{i=1}^{p-1} \frac{(\beta_{0,i+1} - \beta_{0,i})^2}{w_{0,i}^2} \\ &\quad + \frac{\lambda_1^2}{4}\sum_{i=1}^{p} \tau_{0,i}^2 + \frac{\lambda_2^2}{4}\sum_{i=1}^{p-1} w_{0,i}^2. \end{aligned}$$

Using the above and Lemma D.2, for each $\beta_{0,j}$

$$\beta_{0,j}^2 \leq \sum_{i=1}^{p} \beta_{0,i}^2 \leq \left(\sum_{i=1}^{p} \tau_{0,i}^2\right)\left(\sum_{i=1}^{p} \frac{\beta_{0,i}^2}{\tau_{0,i}^2}\right) \leq \frac{4d^2}{\lambda_1^2} := d_1^2, \tag{D.10}$$

and similarly for each $i = 1, \ldots, p-1$

$$(\beta_{0,j+1} - \beta_{0,j})^2 \leq \sum_{i=1}^{p-1} (\beta_{0,i+1} - \beta_{0,i})^2 \leq \left(\sum_{i=1}^{p-1} w_{0,i}^2\right)\left(\sum_{i=1}^{p-1} \frac{(\beta_{0,i} - \beta_{0,i})^2}{w_{0,i}^2}\right) \leq \frac{4d^2}{\lambda_2^2} := d_2^2. \tag{D.11}$$

With these bounds involving $\beta_0$ and using Lemma D.4

$$\begin{aligned}
&f(\tau^2, w^2 \mid \beta_0, \sigma^2, y) \\
&= f(\tau^2 \mid \beta_0, \sigma^2, y)\, f(w^2 \mid \beta_0, \sigma^2, y) \\
&\geq \prod_{i=1}^{p} \exp\left\{-\sqrt{\frac{\lambda_1^2 d_1^2}{\sigma^2}}\right\} q_i(\tau_i^2 \mid \sigma^2) \prod_{i=1}^{p-1} \exp\left\{-\sqrt{\frac{\lambda_2^2 d_2^2}{\sigma^2}}\right\} h_i(w_i^2 \mid \sigma^2) \\
&= \exp\left\{-p\sqrt{\frac{\lambda_1^2 d_1^2}{\sigma^2}} - p\sqrt{\frac{\lambda_2^2 d_2^2}{\sigma^2}}\right\} \left[\prod_{i=1}^{p} q_i(\tau_i^2 \mid \sigma^2)\right] \left[\prod_{i=1}^{p-1} h_i(w_i^2 \mid \sigma^2)\right]
\end{aligned}$$

Since for $a, b \geq 0, 2ab \leq a^2 + b^2$,

$$\geq \exp\left\{-\frac{1}{2} - \frac{p^2\lambda_2^2 d_2^2}{2\sigma^2} - \frac{1}{2} - \frac{p^2\lambda_1^2 d_1^2}{2\sigma^2}\right\} \left[\prod_{i=1}^{p} q_i(\tau_i^2 \mid \sigma^2)\right] \left[\prod_{i=1}^{p-1} h_i(w_i^2 \mid \sigma^2)\right] \tag{D.12}$$

where $q_i$ is the density of the reciprocal of an Inverse-Gaussian distribution with parameters $\sqrt{\lambda_1^2\sigma^2/d_1^2}$ and $\lambda_1^2$ and $h_i$ is the density of the reciprocal of an Inverse-Gaussian distribution with parameters $\sqrt{\lambda_2^2\sigma^2/d_2^2}$ and $\lambda_2^2$.

By Lemma D.5

$$(y - X\beta_0)^T(y - X\beta_0) + \beta_0^T\Sigma_{\beta_0}^{-1}\beta_0 \geq y^Ty - y^TX\left(X^TX + \Sigma_{\beta_0}^{-1}\right)^{-1}X^Ty$$

Recall the decomposition $\Sigma_{\beta_0}^{-1} = L_{0,1} + L_{0,2}$ in (D.4); here the 0 in the index indicates $\tau_0^2$ and $w_0^2$ entries. Here $L_{0,1}$ is the diagonal matrix with entries $1/(2\tau_{0,i}^2)$. Then since

$$\begin{aligned} &y^TX(X^TX + L_{0,1} + L_{0,2})X^Ty \geq y^TX(X^TX + L_{0,1})X^Ty \\ \Rightarrow\ &y^TX(X^TX + L_{0,1} + L_{0,2})^{-1}X^Ty \leq y^TX(X^TX + L_{0,1})^{-1}X^Ty \end{aligned}$$

Using the above and the fact that for each $i = 1, \ldots, p$, $2\tau_{0,i}^2 \leq 8d/\lambda_1^2$

$$\begin{aligned} (y - X\beta_0)^T(y - X\beta_0) + \beta_0^T\Sigma_{\beta_0}^{-1}\beta_0 &\geq y^Ty - y^TX\left(X^TX + \Sigma_{\beta}^{-1}\right)^{-1}X^Ty \\ &\geq y^Ty - y^TX(X^TX + L_{0,1})^{-1}X^Ty \\ &\geq y^Ty - y^TX\left(X^TX + \frac{\lambda_1^2}{8d}I_p\right)^{-1}X^Ty \end{aligned} \tag{D.13}$$

Using (D.13) and the fact that for $(\beta_0, \tau_0^2, w_0^2, \sigma_0^2) \in C_d$, $(y - X\beta_0)^T(y - X\beta_0) + \beta_0^T\Sigma_{\beta_0}^{-1}\beta_0 \leq d$,

$$\begin{aligned} &\exp\left\{-\frac{1}{2} - \frac{p^2\lambda_2^2d_2^2}{2\sigma^2} - \frac{1}{2} - \frac{p^2\lambda_1^2d_1^2}{2\sigma^2}\right\} f(\sigma^2 \mid \beta_0, \tau_0^2, w_0^2, y) \\ &= \exp\left\{-1 - \frac{p^2\lambda_2^2d_2^2}{2\sigma^2} - \frac{p^2\lambda_1^2d_1^2}{2\sigma^2}\right\} \frac{\left(\frac{(y-X\beta_0)^T(y-X\beta_0)+\beta_0^T\Sigma_{\beta_0}^{-1}\beta_0+2\xi}{2}\right)^{\frac{n+p}{2}+\alpha}}{\Gamma\left(\frac{n+p}{2}+\alpha\right)} \\ &\quad \times \left(\sigma^2\right)^{-\frac{n+p}{2}-\alpha-1}\exp\left\{-\frac{(y - X\beta_0)^T(y - X\beta_0) + \beta_0^T\Sigma_{\beta_0}^{-1}\beta_0 + 2\xi}{2\sigma^2}\right\} \\ &\geq e^{-1}\left(\frac{y^Ty - y^TX(X^TX + \lambda_1^2(8d)^{-1}I_p)^{-1}X^Ty + 2\xi}{2}\right)^{\frac{n+p}{2}+\alpha}\frac{1}{\Gamma\left(\frac{n+p}{2}+\alpha\right)} \end{aligned}$$

$$
\begin{aligned}
&\times \left(\sigma^2\right)^{-\frac{n+p}{2}-\alpha-1} \exp\left\{-\frac{d+2\xi+p^2\lambda_2^2 d_2^2+p^2\lambda_1^2 d_1^2}{2\sigma^2}\right\} \\
&= e^{-1}\left(\frac{y^Ty - y^TX(X^TX+\lambda_1^2(8d)^{-1}I_p)^{-1}X^Ty+2\xi}{d+2\xi+p^2\lambda_2^2 d_2^2+p^2\lambda_1^2 d_1^2}\right)^{\frac{n+p}{2}+\alpha} \\
&\times \frac{\left(\frac{d+2\xi+p^2\lambda_2^2 d_2^2+p^2\lambda_1^2 d_1^2}{2}\right)^{\frac{n+p}{2}+\alpha}}{\Gamma\left(\frac{n+p}{2}+\alpha\right)} \\
&\times \left(\sigma^2\right)^{-\frac{n+p}{2}-\alpha-1} \exp\left\{-\frac{d+2\xi+p^2\lambda_2^2 d_2^2+p^2\lambda_1^2 d_1^2}{2\sigma^2}\right\} \\
&= e^{-1}\left(\frac{y^Ty - y^TX(X^TX+\lambda_1^2(8d)^{-1}I_p)^{-1}X^Ty+2\xi}{d+2\xi+p^2\lambda_2^2 d_2^2+p^2\lambda_1^2 d_1^2}\right)^{\frac{n+p}{2}+\alpha} q(\sigma^2), \qquad \text{(D.14)}
\end{aligned}
$$

where $q(\sigma^2)$ is the density of the Inverse-Gamma distribution with parameters, $(n+p)/2+\alpha$ and $d+2\xi+p^2\lambda_2^2 d_2^2+p^2\lambda_1^2 d_1^2$.

Finally, using (D.12) and (D.14),

$$
\begin{aligned}
&k_{BFL}(\beta,\tau^2,w^2,\sigma^2 \mid \beta_0,\tau_0^2,w_0^2,\sigma_0^2) \\
\geq\ &\epsilon\, f(\beta \mid \tau^2,w^2,\sigma^2,y)\, q(\sigma^2)\left[\prod_{i=1}^{p} q_i(\tau_i^2 \mid \sigma^2)\right]\left[\prod_{i=1}^{p-1} h_i(w_i^2 \mid \sigma^2)\right], \qquad \text{(D.15)}
\end{aligned}
$$

where

$$
\epsilon = e^{-1}\left(\frac{y^Ty - y^TX(X^TX+\lambda_1(8d)^{-1}I_p)^{-1}X^Ty+2\xi}{d+2\xi+p^2\lambda_2^2 d_2^2+p^2\lambda_1^2 d_1^2}\right)^{\frac{n+p}{2}+\alpha}. \qquad \text{(D.16)}
$$

Thus the minorization condition is satisfied. This completes the proof of Theorem 6.1.

### D.5.1 Starting Values

$$
V_{BFL}(\beta,\tau^2,w^2,\sigma^2) = (y-X\beta)^T(y-X\beta) + \beta^T\Sigma_\beta^{-1}\beta + \frac{\lambda_1^2}{4}\sum_{i=1}^{p}\tau_i^2 + \frac{\lambda_2^2}{4}\sum_{i=1}^{p-1} w_i^2.
$$

Starting value $(\beta_0, \tau_0^2, w_0^2, \sigma_0^2)$ is chosen such that

$$(\beta_0, \tau_0^2, w_0^2, \sigma_0^2) = \arg\min \ \ V_{BFL}(\beta, \tau^2, w^2, \sigma^2).$$

We will find the minimum by using profiling out $\tau^2$ and $w^2$. By (D.6)

$$\frac{\partial V_{BFL}}{\partial \tau_{0,i}^2} = 0 \Rightarrow = -\frac{\beta_{0,i}^2}{\tau_{0,i}^4} + \frac{\lambda_1^2}{4} = 0 \Rightarrow \tau_{0,i}^2 = \sqrt{\frac{4\beta_{0,i}^2}{\lambda_1^2}}$$
$$\frac{\partial V_{BFL}}{\partial w_{0,i}^2} = 0 \Rightarrow = -\frac{(\beta_{0,i+1} - \beta_{0,i})^2}{w_{0,i}^4} + \frac{\lambda_2^2}{4} = 0 \Rightarrow w_{0,i}^2 = \sqrt{\frac{4(\beta_{0,i+1} - \beta_{0,i})^2}{\lambda_2^2}}$$

The $\beta_0$ that minimizes $V_{BFL}$ is,

$$\begin{aligned}\beta_0 &= \arg\min_{\beta \in \mathbb{R}^p} \left\{ (y - X\beta)^T (y - X\beta) + \sum_{i=1}^{p} \frac{\lambda_1 \beta_i^2}{2\sqrt{\beta_i^2}} + \sum_{i=1}^{p-1} \frac{\lambda_2 (\beta_{i+1} - \beta_i)^2}{2\sqrt{(\beta_{i+1} - \beta_i)^2}} \right. \\ &\quad \left. + \sum_{i=1}^{p} \frac{\lambda_1^2}{4} \sqrt{\frac{4\beta_i^2}{\lambda_1^2}} + \frac{\lambda_2^2}{4} \sum_{i=1}^{p-1} \sqrt{\frac{4(\beta_{i+1} - \beta_i)^2}{\lambda_2^2}} \right\} \\ &= \arg\min_{\beta \in \mathbb{R}^p} \left\{ (y - X\beta)^T (y - X\beta) + \lambda_1 \sum_{i=1}^{p} |\beta_i| + \lambda_2 \sum_{i=1}^{p-1} |\beta_{i+1} - \beta_i| \right\},\end{aligned}$$

which equivalent to the fused lasso solution. Thus, a reasonable starting value is $\beta_0$ being the fused lasso estimate, $\tau_{0,i}^2 = 2|\beta_{0,i}|/\lambda_1$ and $w_{0,i}^2 = 2|\beta_{0,i+1} - \beta_{0,i}|/\lambda_2$.

## D.6 Proof of Theorem 6.2

We will prove geometric ergodicity for the three variable Gibbs sampler for the BGL by establishing a drift condition and an associated minorization condition.

Let

$$V_{BGL}(\beta, \tau^2, \sigma^2) = (y - X\beta)^T (y - X\beta) + \beta^T D_\tau^{-1} \beta + \frac{\lambda^2}{4} \sum_{k=1}^{K} \tau_k^2. \tag{D.17}$$

Then $V_{BGL} : \mathbb{R}^p \times \mathbb{R}_+^K \times \mathbb{R}_+ \to \mathbb{R}_+$. To establish the drift condition we need to show that there exists a $0 < \phi_{BGL} < 1$ and $L_{BGL} > 0$ such that,

$$\mathrm{E}_k \left[ V_{BGL}(\beta, \tau^2, \sigma^2) \,|\, \beta_0, \tau_0^2, \sigma_0^2 \right] \leq \phi_{BGL} V_{BGL}(\beta_0, \tau_0^2, \sigma_0^2) + L_{BGL},$$

for every $(\beta_0, \tau_0^2, \sigma_0^2) \in \mathbb{R}^p \times \mathbb{R}_+^K \times \mathbb{R}_+$. Just as in the proof for BFL,

$$\begin{aligned}
&\mathrm{E}_k \left[ V_{BGL}(\beta, \tau^2, \sigma^2) \;\mid\; \beta_0, \tau_0^2, \sigma_0^2 \right] \\
&= \mathrm{E}_{\sigma^2} \left[ \mathrm{E}_{\tau^2} \left[ \mathrm{E}_{\beta} \left[ V_{BGL}(\beta, \tau^2, \sigma^2) \;\mid\; \tau^2, \sigma^2, y \right] \;\mid\; \beta_0, \sigma^2, y \right] \;\mid\; \beta_0, \tau_0^2, y \right].
\end{aligned}$$

We will evaluate the expectations sequentially, starting with the innermost expectation. By Lemma D.1

$$\begin{aligned}
&\mathrm{E}_{\beta} \left[ V_{BGL}(\beta, \tau^2, \sigma^2) \mid \tau^2, \sigma^2 \right] \\
&= \mathrm{E}_{\beta} \left[ (y - X\beta)^T (y - X\beta) + \beta^T D_\tau^{-1} \beta + \frac{\lambda^2}{4} \sum_{k=1}^{K} \tau_k^2 \mid \tau^2, \sigma^2 \right] \\
&\leq y^T y + \frac{\lambda^2}{4} \sum_{k=1}^{K} \tau_k^2 + p\sigma^2.
\end{aligned}$$

Next we move on to the expectation with respect to the full conditional of $\tau^2$.

$$\begin{aligned}
&\mathrm{E}_{\tau^2} \left[ \mathrm{E}_{\beta} \left[ V_{BGL}(\beta, \tau^2, \sigma^2) \;\mid\; \tau^2, \sigma^2, y \right] \;\mid\; \beta_0, \sigma^2, y \right] \\
&\quad \leq \mathrm{E}_{\tau^2} \left[ y^T y + \frac{\lambda^2}{4} \sum_{k=1}^{K} \tau_k^2 + p\sigma^2 \mid \beta_0, \sigma^2, y \right] \\
&= y^T y + p\sigma^2 + \frac{\lambda^2}{4} \sum_{k=1}^{K} \mathrm{E}_{\tau^2} \left[ \tau_k^2 \mid \beta_0, \sigma^2, y \right] \\
&= y^T y + p\sigma^2 + \frac{\lambda^2}{4} \sum_{k=1}^{K} \left[ \sqrt{\frac{\beta_{0,G_k}^T \beta_{0,G_k}}{\lambda^2 \sigma^2}} + \frac{1}{\lambda^2} \right].
\end{aligned}$$

Let $M = \max\{m_1, \ldots, m_K\}$. Then,

$$
\begin{aligned}
&\mathrm{E}_{\tau^2}\left[\mathrm{E}_{\beta}\left[V_{BGL}(\beta, \tau^2, \sigma^2) \mid \tau^2, \sigma^2, y\right] \mid \beta_0, \sigma^2, y\right] \\
&\leq y^T y + p\sigma^2 + \frac{\lambda^2}{4}\sum_{k=1}^{K}\left[\sqrt{\frac{M(n+p+2\alpha)\beta_{0,G_k}^T\beta_{0,G_k}}{M(n+p+2\alpha)\lambda^2\sigma^2}} + \frac{1}{\lambda^2}\right] \\
&\leq y^T y + p\sigma^2 + \frac{\lambda^2}{4}\sum_{k=1}^{K}\left[\frac{\beta_{0,G_k}^T\beta_{0,G_k}}{2\sigma^2 M(n+p+2\alpha)} + \frac{M(n+p+2\alpha)}{2\lambda^2} + \frac{1}{\lambda^2}\right] \\
&\leq y^T y + p\sigma^2 + \frac{p}{4}\left(1 + \frac{M(n+p+2\alpha)}{2}\right) + \frac{\lambda^2\sum_{k=1}^{K}\beta_{0,G_k}^T\beta_{0,G_k}}{8\sigma^2 M(n+p+2\alpha)}
\end{aligned}
$$

Finally, the last expectation,

$$
\begin{aligned}
&\mathrm{E}_{\sigma^2}\left[\mathrm{E}_{\tau^2}\left[\mathrm{E}_{\beta}\left[V_{BGL}(\beta, \tau^2, \sigma^2) \mid \tau^2, \sigma^2, y\right] \mid \beta_0, \sigma^2, y\right] \mid \beta_0, \tau_0^2, y\right] \\
&\leq y^T y + \frac{p}{4}\left(1 + \frac{M(n+p+2\alpha)}{2}\right) \\
&\quad + \frac{\lambda^2\sum_{k=1}^{K}\beta_{0,G_k}^T\beta_{0,G_k}}{8M(n+p+2\alpha)}\,\mathrm{E}_{\sigma^2}\left[\frac{1}{\sigma^2} \mid \beta_0, \tau_0^2, y\right] + p\mathrm{E}_{\sigma^2}\left[\sigma^2 \mid \beta_0, \tau_0^2, y\right] \\
&= y^T y + \frac{p}{4}\left(1 + \frac{M(n+p+2\alpha)}{2}\right) \\
&\quad + \frac{\lambda^2\sum_{k=1}^{K}\beta_{0,G_k}^T\beta_{0,G_k}}{8M(n+p+2\alpha)}\,\frac{(n+p+2\alpha)}{(y-X\beta_0)^T(y-X\beta_0) + \beta_0^T D_{\tau_0}^{-1}\beta_0 + 2\xi} \\
&\quad + p\frac{(y-X\beta_0)^T(y-X\beta_0) + \beta_0^T D_{\tau_0}^{-1}\beta_0 + 2\xi}{n+p+2\alpha-2} \\
&= y^T y + \frac{p}{4}\left(1 + \frac{M(n+p+2\alpha)}{2}\right) \\
&\quad + \frac{\lambda^2\sum_{k=1}^{K}\beta_{0,G_k}^T\beta_{0,G_k}}{8M\left((y-X\beta_0)^T(y-X\beta_0) + \beta_0^T D_{\tau_0}^{-1}\beta_0 + 2\xi\right)} \\
&\quad + p\frac{(y-X\beta_0)^T(y-X\beta_0) + \beta_0^T D_{\tau_0}^{-1}\beta_0 + 2\xi}{n+p+2\alpha-2} \\
&\leq y^T y + \frac{p}{4}\left(1 + \frac{M(n+p+2\alpha)}{2}\right) + \frac{\lambda^2}{8M}\left(\frac{\sum_{k=1}^{K}\beta_{0,G_k}^T\beta_{0,G_k}}{\beta_0^T D_{\tau_0}^{-1}\beta_0}\right)
\end{aligned}
$$

$$+ p\frac{(y - X\beta_0)^T(y - X\beta_0) + \beta_0^T D_{\tau_0}^{-1}\beta_0 + 2\xi}{n + p + 2\alpha - 2}$$

Recall that,

$$D_{\tau_0} = \text{diag}(\underbrace{\tau_{0,1}^2, \ldots, \tau_{0,1}^2}_{m_1}, \underbrace{\tau_{0,2}^2, \ldots, \tau_{0,2}^2}_{m_2}, \ldots, \underbrace{\tau_{0,K}^2, \ldots, \tau_{0,K}^2}_{m_K}),$$

Let the diagonals of $D_{\tau_0}$ be $\tau_{0,*i}^2$ for $i = 1, \ldots, p$. Then $\beta_0^T D_{\tau_0}^{-1}\beta_0 = \sum_{i=1}^p \beta_{0,i}^2/\tau_{0,*i}^2$ and $\sum_{i=1}^p \tau_{0,*i}^2 \leq M\sum_{k=1}^K \tau_{0,k}^2$. By Lemma D.2,

$$\begin{aligned}
&\text{E}_{\sigma^2}\left[\text{E}_{\tau^2}\left[\text{E}_{\beta}\left[V_{BGL}(\beta, \tau^2, \sigma^2) \mid \tau^2, \sigma^2, y\right] \mid \beta_0, \sigma^2, y\right] \mid \beta_0, \tau_0^2, y\right] \\
&\leq y^T y + \frac{p}{4}\left(1 + \frac{M(n + p + 2\alpha)}{2}\right) + \frac{\lambda^2}{8M}\sum_{i=1}^p \tau_{0,*i}^2 \\
&\quad + p\frac{(y - X\beta_0)^T(y - X\beta_0) + \beta_0^T D_{\tau_0}^{-1}\beta_0 + 2\xi}{n + p + 2\alpha - 2} \\
&\leq y^T y + \frac{p}{4}\left(1 + \frac{M(n + p + 2\alpha)}{2}\right) + \frac{\lambda^2}{8}\sum_{k=1}^K \tau_{0,k}^2 \\
&\quad + p\frac{(y - X\beta_0)^T(y - X\beta_0) + \beta_0^T D_{\tau_0}^{-1}\beta_0 + 2\xi}{n + p + 2\alpha - 2} \\
&= \frac{p}{n + p + 2\alpha - 2}\left((y - X\beta_0)^T(y - X\beta_0) + \beta_0^T D_{\tau_0}^{-1}\beta_0\right) + \frac{1}{2}\left(\frac{\lambda^2}{4}\sum_{k=1}^K \tau_{0,k}^2\right) \\
&\quad + y^T y + \frac{p}{4}\left(1 + \frac{M(n + p + 2\alpha)}{2}\right) + \frac{2p\xi}{n + p + 2\alpha - 2} \\
&\leq \phi_{BGL} V_{BGL}(\beta_0, \tau_0^2, \sigma_0^2) + L_{BGL},
\end{aligned}$$

where

$$\phi_{BGL} = \max\left\{\frac{p}{n + p + 2\alpha - 2}, \frac{1}{2}\right\} < 1 \text{ for } n \geq 3, \tag{D.18}$$

and

$$L_{BGL} = y^T y + \frac{p}{4}\left(1 + \frac{M(n+p+2\alpha)}{2}\right) + \frac{2p\xi}{n+p+2\alpha-2}. \tag{D.19}$$

This establishes the drift condition and we now turn our attention to the minorization condition. For $d > 0$, define $C_d = \{(\beta, \tau^2, \sigma^2) : V_{BGL}(\beta, \tau^2, \sigma^2) \leq d\}$. We need to show that for all sets $C_d$ there exists an $\epsilon > 0$ and a density $q$ such that for all $(\beta_0, \tau_0^2, \sigma_0^2) \in C_d$

$$k_{BGL}(\beta, \tau^2, \sigma^2 \mid \beta_0, \tau_0^2, \sigma_0^2) \geq \epsilon\, q(\beta, \tau^2, \sigma^2).$$

To establish this condition, we recall that,

$$k_{BGL}(\beta, \tau^2, \sigma^2 \mid \beta_0, \tau_0^2, \sigma_0^2) = f(\beta \mid \tau^2, \sigma^2, y)\; f(\tau^2 \mid \beta_0, \sigma^2, y)\; f(\sigma^2 \mid \beta_0, \tau_0^2, y).$$

By our definition of the drift function, for all $(\beta_0, \tau_0^2, \sigma_0^2) \in C_d$ the following relation holds,

$$(y - X\beta_0)^T(y - X\beta_0) + \beta_0^T D_{\tau_0}^{-1}\beta_0 + \frac{\lambda^2}{4}\sum_{k=1}^{K}\tau_{0,k}^2 \leq d.$$

The above implies that each of $\beta_0^T D_{\tau_0}^{-1}\beta_0$ and $(\lambda^2/4)\sum_{k=1}^{K}\tau_{0,k}^2$ is less than or equal to $d$, so

$$\begin{aligned}
&\left(\beta_0^T D_{\tau_0}^{-1}\beta_0\right)\left(\sum_{k=1}^{K}\tau_{0,k}^2\right) \leq \frac{4d^2}{\lambda^2} \\
&\Rightarrow \sum_{k=1}^{K}\beta_{0,G_k}^T\beta_{0,G_k} \leq \left(\sum_{k=1}^{K}\frac{\beta_{0,G_k}^T\beta_{0,G_k}}{\tau_{0,k}^2}\right)\left(\sum_{k=1}^{K}\tau_{0,k}^2\right) \leq \frac{4d^2}{\lambda^2} \quad \text{by Lemma D.2} \\
&\Rightarrow \beta_{0,G_k}^T\beta_{0,G_k} \leq \frac{4d^2}{\lambda^2} := d_1^2 \quad \text{for all } k = 1, \ldots, K.
\end{aligned} \tag{D.20}$$

By Lemma D.4,

$$f(\tau^2 \mid \beta_0, \sigma^2, y) \geq \prod_{k=1}^{K} \exp\left\{-\sqrt{\frac{\lambda^2 d_1^2}{\sigma^2}}\right\} q_k(\tau_k^2 \mid \sigma^2)$$
$$\geq \exp\left\{-\frac{1}{2} - \frac{K^2\lambda^2 d_1^2}{2\sigma^2}\right\} \prod_{k=1}^{K} q_k(\tau_k^2 \mid \sigma^2), \tag{D.21}$$

where $q_k$ is the density of the reciprocal of an Inverse-Gaussian distribution with parameters $\sqrt{\lambda^2\sigma^2/d_1^2}$ and $\lambda^2$.

Now, since for each $i = 1, \ldots, p$, $\tau_{0,i}^2 \leq 4d/\lambda^2$, by Lemma D.5

$$(y - X\beta_0)^T(y - X\beta_0) + \beta_0^T D_{\tau_0}^{-1}\beta_0 \geq y^T y - y^T X \left(X^T X + D_{\tau_0}^{-1}\right)^{-1} X^T y$$
$$\geq y^T y - y^T X \left(X^T X + \frac{\lambda^2}{4d} I_p\right)^{-1} X^T y$$

Using the above relation, we arrive at the following,

$$\exp\left\{-\frac{1}{2} - \frac{K^2\lambda^2 d_1^2}{2\sigma^2}\right\} f(\sigma^2 \mid \beta_0, \tau_0^2, y)$$
$$= \exp\left\{-\frac{1}{2} - \frac{K^2\lambda^2 d_1^2}{2\sigma^2}\right\} \frac{\left(\frac{(y-X\beta_0)^T(y-X\beta_0)+\beta_0^T D_{\tau_0}^{-1}\beta_0+2\xi}{2}\right)^{\frac{n+p}{2}+\alpha}}{\Gamma\left(\frac{n+p}{2}+\alpha\right)} \left(\sigma^2\right)^{-\frac{n+p}{2}-\alpha-1}$$
$$\times \exp\left\{-\frac{(y - X\beta_0)^T(y - X\beta_0) + \beta_0^T D_{\tau_0}^{-1}\beta_0 + 2\xi}{2\sigma^2}\right\}$$
$$\geq e^{-\frac{1}{2}} \left(\frac{y^T y - y^T X(X^T X + \lambda^2(4d)^{-1} I_p)^{-1} X^T y + 2\xi}{2}\right)^{\frac{n+p}{2}+\alpha} \frac{1}{\Gamma\left(\frac{n+p}{2}+\alpha\right)}$$
$$\times \left(\sigma^2\right)^{-\frac{n+p}{2}-\alpha-1} \exp\left\{-\frac{d + 2\xi + K^2\lambda^2 d_1^2}{2\sigma^2}\right\}$$
$$= e^{-\frac{1}{2}} \left(\frac{y^T y - y^T X(X^T X + \lambda^2(4d)^{-1} I_p)^{-1} X^T y + 2\xi}{d + 2\xi + K^2\lambda^2 d_1^2}\right)^{\frac{n+p}{2}+\alpha}$$
$$\times \frac{\left(\frac{d + 2\xi + K^2\lambda^2 d_1^2}{2}\right)^{\frac{n+p}{2}+\alpha}}{\Gamma\left(\frac{n+p}{2}+\alpha\right)} \left(\sigma^2\right)^{-\frac{n+p}{2}-\alpha-1} \exp\left\{-\frac{d + 2\xi + K^2\lambda^2 d_1^2}{2\sigma^2}\right\}$$

$$= e^{-\frac{1}{2}} \left( \frac{y^T y - y^T X (X^T X + \lambda^2 (4d)^{-1} I_p)^{-1} X^T y + 2\xi}{d + 2\xi + K^2 \lambda^2 d_1^2} \right)^{\frac{n+p}{2}+\alpha} q(\sigma^2), \tag{D.22}$$

where $q(\sigma^2)$ is the density of the Inverse-Gamma distribution with parameters, $(n + p)/2 + \alpha$ and $d + 2\xi + K^2\lambda^2 d_1^2$.

Finally, using (D.21) and (D.22),

$$k_{BGL}(\beta, \tau^2, \sigma^2 \mid \beta_0, \tau_0^2, \sigma_0^2) \geq \ \epsilon \ f(\beta \mid \tau^2, \sigma^2, y) \ q(\sigma^2) \prod_{k=1}^{K} q_k(\tau^2 \mid \sigma^2), \tag{D.23}$$

where

$$\epsilon = e^{-\frac{1}{2}} \left( \frac{y^T y - y^T X (X^T X + \lambda^2 (4d)^{-1} I_p)^{-1} X^T y + 2\xi}{d + 2\xi + 4K^2 d^2} \right)^{\frac{n+p}{2}+\alpha}. \tag{D.24}$$

### D.6.1 Starting Values

$$V_{BGL}(\beta, \tau^2, \sigma^2) = (y - X\beta)^T (y - X\beta) + \beta^T D_\tau^{-1} \beta + \frac{\lambda^2}{4} \sum_{k=1}^{K} \tau_k^2.$$

The starting value $(\beta_0, \tau_0^2, \sigma_0^2)$ is chosen such that,

$$(\beta_0, \tau_0^2, \sigma_0^2) = \arg\min \ \ V_{BGL}(\beta, \tau^2, \sigma^2).$$

We will solve the minimization problem by profiling. We proceed by first differentiating with respect to $\tau^2$ and then with respect to $\beta$.

$$\frac{\partial V_{BGL}}{\partial \tau_{0,k}^2} = 0 \Rightarrow = -\frac{\beta_{0,G_k}^T \beta_{0,G_k}}{\tau_{0,k}^4} + \frac{\lambda^2}{4} = 0 \Rightarrow \tau_{0,k}^2 = \sqrt{\frac{4\beta_{0,G_k}^T \beta_{0,G_k}}{\lambda^2}}$$

Thus, the $\beta_0$ that minimizes $V_{BGL}$ is then,

$$\beta_0 = \arg\min_{\beta \in \mathbb{R}^p} \ (y - X\beta)^T (y - X\beta) + \sum_{k=1}^{K} \frac{\lambda \ \beta_{G_k}^T \beta_{G_k}}{2\sqrt{\beta_{G_k}^T \beta_{G_k}}} + \frac{\lambda^2}{4} \sum_{k=1}^{K} \sqrt{\frac{4\beta_{G_k}^T \beta_{G_k}}{\lambda^2}}$$

$$= \arg\min_{\beta \in \mathbb{R}^p} \ (y - X\beta)^T (y - X\beta) + \lambda \sum_{k=1}^{K} \sqrt{\beta_{G_k}^T \beta_{G_k}},$$

which equivalent to the group lasso solution. Thus a reasonable starting value for the Markov chain is $\beta_0$ being the group lasso estimate and $\tau_{0,k}^2 = 2\sqrt{\beta_{0,G_k}^T \beta_{0,G_k}}/\lambda$.

## D.7 Proof of Theorem 6.3

Consider the drift function

$$V_{BSGL}(\beta, \tau^2, \gamma^2, \sigma^2) = (y - X\beta)^T (y - X\beta) + \beta^T V_{\tau,\gamma}^{-1} \beta + \frac{\lambda_1^2}{4} \sum_{k=1}^{K} \tau_k^2 + \frac{\lambda_2^2}{4} \sum_{k=1}^{K} \sum_{j=1}^{m_k} \gamma_{k,j}^2. \tag{D.25}$$

Then $V_{BSGL} : \mathbb{R}^p \times \mathbb{R}_+^K \times \mathbb{R}_+^p \times \mathbb{R}_+ \to \mathbb{R}_+$. The BSGL Gibbs sampler MTD is

$$\begin{aligned} &k_{BSGL}(\beta, \tau^2, \gamma^2, \sigma^2 \mid \beta_0, \tau_0^2, \gamma_0^2, \sigma_0^2) \\ &= f(\beta \mid \tau^2, \gamma^2, \sigma^2, y) \ f(\tau^2, \gamma^2 \mid \beta_0, \sigma^2, y) \ f(\sigma^2 \mid \beta_0, \tau_0^2, \gamma_0^2, y). \end{aligned}$$

Just as before

$$\begin{aligned} &\mathrm{E}\left[V_{BSGL}(\beta, \tau^2, \gamma^2, \sigma^2) \mid \tau^2, \gamma^2, \sigma^2\right] \\ &= \mathrm{E}_{\sigma^2}\left[\mathrm{E}_{\tau^2,\gamma^2}\left[\mathrm{E}_{\beta}\left[V_{BSGL}(\beta, \tau^2, \gamma^2, \sigma^2) \ \mid \ \tau^2, \gamma^2, \sigma^2, y\right] \ \mid \ \beta_0, \sigma^2, y\right] \ \mid \ \beta_0, \tau_0^2, \gamma_0^2, y\right] \end{aligned}$$

In order to establish the drift condition, we will evaluate the expectations sequentially, starting with the innermost expectation. By Lemma D.1

$$\mathrm{E}\left[V_{BSGL}(\beta, \tau^2, \gamma^2, \sigma^2) \mid \tau^2, \gamma^2, \sigma^2\right]$$

$$= \mathrm{E}\left[(y - X\beta)^T(y - X\beta) + \beta^T V_{\tau,\gamma}^{-1}\beta + \frac{\lambda_1^2}{4}\sum_{k=1}^{K}\tau_k^2 + \frac{\lambda_2^2}{4}\sum_{k=1}^{K}\sum_{j=1}^{m_k}\gamma_{k,j}^2 \mid \tau^2, \gamma^2, \sigma^2\right]$$
$$\leq y^T y + \frac{\lambda_1^2}{4}\sum_{k=1}^{K}\tau_k^2 + \frac{\lambda_2^2}{4}\sum_{k=1}^{K}\sum_{j=1}^{m_k}\gamma_{k,j}^2 + p\sigma^2.$$

Next we move on to the expectation with respect to the full conditional of $\tau^2, \gamma^2$.

$$\mathrm{E}_{\tau^2,\gamma^2}\left[\mathrm{E}_\beta\left[V_{BSGL}(\beta, \tau^2, \gamma^2, \sigma^2) \mid \tau^2, \gamma^2, \sigma^2, y\right] \mid \beta_0, \sigma^2, y\right]$$
$$\leq \mathrm{E}_{\tau^2,\gamma^2}\left[y^T y + \frac{\lambda_1^2}{4}\sum_{k=1}^{K}\tau_k^2 + \frac{\lambda_2^2}{4}\sum_{k=1}^{K}\sum_{j=1}^{m_k}\gamma_{k,j}^2 + p\sigma^2 \mid \beta_0, \sigma^2, y\right]$$
$$= y^T y + p\sigma^2 + \frac{\lambda_1^2}{4}\sum_{k=1}^{K}\mathrm{E}_{\tau^2}\left[\tau_k^2 \mid \beta_0, \sigma^2, y\right] + \frac{\lambda_2^2}{4}\sum_{k=1}^{K}\sum_{j=1}^{m_k}\mathrm{E}_{\gamma^2}\left[\gamma_{k,j}^2 \mid \beta_0, \sigma^2, y\right]$$
$$= y^T y + p\sigma^2 + \frac{\lambda_1^2}{4}\sum_{k=1}^{K}\left[\sqrt{\frac{\beta_{0,G_k}^T\beta_{0,G_k}}{\lambda_1^2\sigma^2}} + \frac{1}{\lambda_1^2}\right] + \frac{\lambda_2^2}{4}\sum_{k=1}^{K}\sum_{j=1}^{m_k}\left[\sqrt{\frac{\beta_{0,k,j}^2}{\lambda_2^2\sigma^2}} + \frac{1}{\lambda_2^2}\right].$$

Define $M = \max\{m_1, \ldots, m_K\}$. In addition, define

$$A = \left(1 + \frac{\lambda_1^2}{\lambda_2^2} + \frac{\lambda_2^2}{\lambda_1^2}\right)(n + p + 2\alpha).$$

Then,

$$\mathrm{E}_{\tau^2,\gamma^2}\left[\mathrm{E}_\beta\left[V_{BSGL}(\beta, \tau^2, \sigma^2) \mid \tau^2, \sigma^2, y\right] \mid \beta_0, \sigma^2, y\right]$$
$$\leq y^T y + p\sigma^2 + \frac{\lambda_1^2}{4}\sum_{k=1}^{K}\left[\frac{\beta_{0,G_k}^T\beta_{0,G_k}}{2\sigma^2 AM} + \frac{AM}{2\lambda_1^2} + \frac{1}{\lambda_1^2}\right]$$
$$+ \frac{\lambda_2^2}{4}\sum_{k=1}^{K}\sum_{j=1}^{m_k}\left[\frac{\beta_{0,k,j}^2}{2\sigma^2 AM} + \frac{AM}{2\lambda_2^2} + \frac{1}{\lambda_2^2}\right]$$
$$\leq y^T y + p\sigma^2 + \frac{p}{4}(2 + AM) + \frac{\lambda_1^2}{8AM}\sum_{k=1}^{K}\frac{\beta_{0,G_k}^T\beta_{0,G_k}}{\sigma^2} + \frac{\lambda_2^2}{8AM}\sum_{k=1}^{K}\sum_{j=1}^{m_k}\frac{\beta_{0,k,j}^2}{\sigma^2}$$
$$= y^T y + p\sigma^2 + \frac{p}{4}(2 + AM) + \left[\frac{\lambda_1^2 + \lambda_2^2}{8AM}\right]\frac{\beta_0^T\beta_0}{\sigma^2}.$$

Finally, the last expectation,

$$
\begin{aligned}
&\mathrm{E}_{\sigma^2}\left[\mathrm{E}_{\tau^2,\gamma^2}\left[\mathrm{E}_{\beta}\left[V_{BSGL}(\beta,\tau^2,\gamma^2,\sigma^2) \mid \tau^2,\gamma^2,\sigma^2,y\right] \mid \beta_0,\sigma^2,y\right] \mid \beta_0,\tau_0^2,\gamma_0^2,y\right] \\
&\leq y^Ty + \frac{p}{4}(2+AM) + \left[\frac{\lambda_1^2+\lambda_2^2}{8AM}\right] \mathrm{E}_{\sigma^2}\left[\frac{\beta_0^T\beta_0}{\sigma^2} \mid \beta_0,\tau_0^2,\gamma_0^2,y\right] \\
&\quad + p\mathrm{E}_{\sigma^2}\left[\sigma^2 \mid \beta_0,\tau_0^2,\gamma_0^2,y\right] \\
&= y^Ty + \frac{p}{4}(2+AM) + \left[\frac{\lambda_1^2+\lambda_2^2}{8AM}\right] \frac{(n+p+2\alpha)\beta_0^T\beta_0}{(y-X\beta_0)^T(y-X\beta_0)+\beta^TV_{\tau_0,\gamma_0}^{-1}\beta_0+2\xi} \\
&\quad + p\frac{(y-X\beta_0)^T(y-X\beta_0)+\beta_0^TV_{\tau_0,\gamma_0}^{-1}\beta_0+2\xi}{n+p+2\alpha-2} \\
&= y^Ty + \frac{p}{4}(2+AM) + p\frac{(y-X\beta_0)^T(y-X\beta_0)+\beta_0^TV_{\tau_0,\gamma_0}^{-1}\beta+2\xi}{n+p+2\alpha-2} \\
&\quad + (\lambda_1^2+\lambda_2^2)\left[8M\left(1+\frac{\lambda_1^2}{\lambda_2^2}+\frac{\lambda_2^2}{\lambda_1^2}\right)\right]^{-1}\frac{\beta_0^T\beta_0}{(y-X\beta_0)^T(y-X\beta_0)+\beta_0^TV_{\tau_0,\gamma_0}^{-1}\beta_0+2\xi} \\
&\leq y^Ty + \frac{p}{4}(2+AM) + (\lambda_1^2+\lambda_2^2)\left[8M\left(1+\frac{\lambda_1^2}{\lambda_2^2}+\frac{\lambda_2^2}{\lambda_1^2}\right)\right]^{-1}\left(\frac{\beta_0^T\beta_0}{\beta_0^TV_{\tau_0,\gamma_0}^{-1}\beta_0}\right) \\
&\quad + p\frac{(y-X\beta_0)^T(y-X\beta_0)+\beta_0^TV_{\tau_0,\gamma_0}^{-1}\beta_0+2\xi}{n+p+2\alpha-2}
\end{aligned}
$$

Let $v_{0,i}$ denote the diagonals of $V_{\tau_0,\gamma_0}$. Then by Lemma D.2,

$$
\frac{\beta_0^T\beta_0}{\beta_0^TV_{\tau_0,\gamma_0}^{-1}\beta_0} \leq \sum_{i=1}^{p} v_{0,i}
$$

Now note that the harmonic mean of positive numbers is less than their arithmetic mean. Using this property,

$$
\begin{aligned}
\sum_{i=1}^{p} v_{0,i} &= \sum_{k=1}^{K}\sum_{j=1}^{m_k}\left(\frac{1}{\tau_{0,k}^2}+\frac{1}{\gamma_{0,k,j}^2}\right)^{-1} = \frac{1}{2}\sum_{k=1}^{K}\sum_{j=1}^{m_k} 2\left(\frac{1}{\tau_{0,k}^2}+\frac{1}{\gamma_{0,k,j}^2}\right)^{-1} \\
&\leq \frac{1}{2}\sum_{k=1}^{K}\sum_{j=1}^{m_k}\frac{\tau_{0,k}^2+\gamma_{0,k,j}^2}{2} \leq \frac{M}{4}\sum_{k=1}^{K}\tau_{0,k}^2 + \frac{1}{4}\sum_{k=1}^{K}\sum_{j=1}^{m_k}\gamma_{0,k,j}^2.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\mathrm{E}_{\sigma^2}\left[\mathrm{E}_{\tau^2}\left[\mathrm{E}_{\beta}\left[V_{BSGL}(\beta,\tau^2,\sigma^2)\mid \tau^2,\sigma^2,y\right]\mid \beta_0,\sigma^2,y\right]\mid \beta_0,\tau_0^2,y\right]\\
&\leq y^Ty+\frac{p}{4}\left(2+AM\right)\\
&\quad+\ (\lambda_1^2+\lambda_2^2)\left[8M\left(1+\frac{\lambda_1^2}{\lambda_2^2}+\frac{\lambda_2^2}{\lambda_1^2}\right)\right]^{-1}\left(\frac{M}{4}\sum_{k=1}^{K}\tau_{0,k}^2+\frac{1}{4}\sum_{k=1}^{K}\sum_{j=1}^{m_k}\gamma_{0,k,j}^2\right)\\
&\quad+p\frac{(y-X\beta_0)^T(y-X\beta_0)+\beta_0^TV_{\tau_0,\gamma_0}^{-1}\beta_0+2\xi}{n+p+2\alpha-2}\\
&\leq y^Ty+\frac{p}{4}\left(2+AM\right)+\ \left(1+\frac{\lambda_2^2}{\lambda_1^2}\right)\left[8\left(1+\frac{\lambda_1^2}{\lambda_2^2}+\frac{\lambda_2^2}{\lambda_1^2}\right)\right]^{-1}\left(\frac{\lambda_1^2}{4}\sum_{k=1}^{K}\tau_{0,k}^2\right)\\
&\quad+\left(1+\frac{\lambda_1^2}{\lambda_2^2}\right)\left[8M\left(1+\frac{\lambda_1^2}{\lambda_2^2}+\frac{\lambda_2^2}{\lambda_1^2}\right)\right]^{-1}\left(\frac{\lambda_2^2}{4}\sum_{k=1}^{K}\sum_{j=1}^{m_k}\gamma_{0,k,j}^2\right)\\
&\quad+p\frac{(y-X\beta_0)^T(y-X\beta_0)+\beta_0^TV_{\tau_0,\gamma_0}^{-1}\beta_0}{n+p+2\alpha-2}\\
&\leq y^Ty+\frac{p}{4}\left(2+AM\right)+\frac{2p\xi}{n+p+2\alpha-2}\\
&\quad+\frac{p}{n+p+2\alpha-2}\left[(y-X\beta_0)^T(y-X\beta_0)+\beta_0^TV_{\tau_0,\gamma_0}^{-1}\beta_0\right]\\
&\quad+\left(1+\frac{\lambda_2^2}{\lambda_1^2}\right)\left[8\left(1+\frac{\lambda_1^2}{\lambda_2^2}+\frac{\lambda_2^2}{\lambda_1^2}\right)\right]^{-1}\left(\frac{\lambda_1^2}{4}\sum_{k=1}^{K}\tau_{0,k}^2\right)\\
&\quad+\left(1+\frac{\lambda_1^2}{\lambda_2^2}\right)\left[8M\left(1+\frac{\lambda_1^2}{\lambda_2^2}+\frac{\lambda_2^2}{\lambda_1^2}\right)\right]^{-1}\left(\frac{\lambda_2^2}{4}\sum_{k=1}^{K}\sum_{j=1}^{m_k}\gamma_{0,k,j}^2\right)\\
&\leq \phi_{BSGL}\ V_{BSGL}(\beta_0,\tau_0^2,\gamma_0^2,\sigma_0^2)+L_{BSGL},
\end{aligned}
$$

where for $n\geq 3$

$$
\phi_{BSGL}=\max\left\{\frac{p}{n+p+2\alpha-2},\frac{\left(1+\frac{\lambda_2^2}{\lambda_1^2}\right)}{8\left(1+\frac{\lambda_1^2}{\lambda_2^2}+\frac{\lambda_2^2}{\lambda_1^2}\right)},\frac{\left(1+\frac{\lambda_1^2}{\lambda_2^2}\right)}{8M\left(1+\frac{\lambda_1^2}{\lambda_2^2}+\frac{\lambda_2^2}{\lambda_1^2}\right)}\right\}<1, \tag{D.26}
$$

and

$$L_{BSGL} = y^T y + \frac{p}{4}(2 + AM) + \frac{2p\xi}{n + p + 2\alpha - 2}. \tag{D.27}$$

This establishes the drift condition and we now turn our attention to the minorization condition. For $d > 0$, define $C_d = \{(\beta, \tau^2, \gamma^2, \sigma^2) : V(\beta, \tau^2, \gamma^2, \sigma^2) \leq d\}$. To establish a one-step minorization, we need to show that for all sets $C_d$, there exists an $\epsilon > 0$ and a density $q$ such that for all $(\beta_0, \tau_0^2, \gamma_0^2, \sigma_0^2) \in C_d$

$$k_{BSGL}(\beta, \tau^2, \gamma^2, \sigma^2 \mid \beta_0, \tau_0^2, \gamma_0^2, \sigma_0^2) \geq \epsilon q(\beta, \tau^2, \gamma^2, \sigma^2).$$

To establish this condition, we recall that,

$$\begin{aligned}
&k_{BSGL}(\beta, \tau^2, \gamma^2, \sigma^2 \mid \beta_0, \tau_0^2, \gamma_0^2, \sigma_0^2) \\
&= f(\beta \mid \tau^2, \gamma^2, \sigma^2, y)\ f(\tau^2, \gamma^2 \mid \beta_0, \sigma^2, y)\ f(\sigma^2 \mid \beta_0, \tau_0^2, \gamma_0^2, y).
\end{aligned}$$

By our definition of the drift function, for all $(\beta_0, \tau_0^2, \gamma_0^2, \sigma_0^2) \in C_d$ the following relation holds,

$$\begin{aligned}
d &\geq (y - X\beta_0)^T (y - X\beta_0) + \beta_0^T V_{\tau_0, \gamma_0}^{-1} \beta_0 + \frac{\lambda_1^2}{4} \sum_{k=1}^{K} \tau_{0,k}^2 + \frac{\lambda_2^2}{4} \sum_{k=1}^{K} \sum_{j=1}^{m_k} \gamma_{0,k,j}^2 \\
d &\geq (y - X\beta_0)^T (y - X\beta_0) + \sum_{k=1}^{K} \sum_{j=1}^{m_k} \beta_{0,k,j}^2 \left( \frac{1}{\tau_{0,k}^2} + \frac{1}{\gamma_{0,k,j}^2} \right) + \frac{\lambda_1^2}{4} \sum_{k=1}^{K} \tau_{0,k}^2 \\
&\quad + \frac{\lambda_2^2}{4} \sum_{k=1}^{K} \sum_{j=1}^{m_k} \gamma_{0,k,j}^2 \\
d &\geq (y - X\beta_0)^T (y - X\beta_0) + \sum_{k=1}^{K} \frac{\beta_{0,G_k}^T \beta_{0,G_k}}{\tau_{0,k}^2} + \sum_{k=1}^{K} \sum_{j=1}^{m_k} \frac{\beta_{0,k,j}^2}{\gamma_{0,k,j}^2} + \frac{\lambda_1^2}{4} \sum_{k=1}^{K} \tau_{0,k}^2 \\
&\quad + \frac{\lambda_2^2}{4} \sum_{k=1}^{K} \sum_{j=1}^{m_k} \gamma_{0,k,j}^2.
\end{aligned}$$

Using the above and following on the lines of (D.20) we get

$$\beta_{0,G_k}^T \beta_{0,G_k} \leq \frac{4d^2}{\lambda_1^2} := d_1^2 \quad \text{for all } k = 1, \ldots, K., \tag{D.28}$$

and similarly for each $k = 1, \ldots, K$ and $j = 1, \ldots, m_k$

$$\beta_{0,k,j}^2 \leq \frac{4d^2}{\lambda_2^2} := d_2^2. \tag{D.29}$$

Using Lemma D.4 and (D.28)

$$\begin{aligned}
&f(\tau^2, \gamma^2 \mid \beta_0, \sigma^2, y) \\
&= f(\tau^2 \mid \beta_0, \sigma^2, y)\, f(\gamma^2 \mid \beta_0, \sigma^2, y) \\
&\geq \prod_{k=1}^{K} \exp\left\{-\sqrt{\frac{\lambda_1^2 d_1^2}{2\sigma^2}}\right\} q_k(\tau_k^2 \mid \sigma^2) \left(\prod_{k=1}^{K}\prod_{j=1}^{m_k} \exp\left\{-\sqrt{\frac{\lambda_2^2 d_2^2}{2\sigma^2}}\right\} q_{k,j}(\gamma_{k,j}^2 \mid \sigma^2)\right) \\
&\geq \exp\left\{-\frac{1}{2} - \frac{p^2\lambda_2^2 d_2^2}{2\sigma^2} - \frac{1}{2} - \frac{K^2\lambda_1^2 d_1^2}{2\sigma^2}\right\} \prod_{k=1}^{K}\left[q_k(\tau_k^2 \mid \sigma^2) \prod_{j=1}^{m_k} q_{k,j}(\gamma_{k,j}^2 \mid \sigma^2)\right],
\end{aligned} \tag{D.30}$$

where $q_k(\tau_k^2 \mid \sigma^2)$ is the density of the reciprocal of an Inverse-Gaussian distribution with parameters $\sqrt{\lambda_1^2\sigma^2/d_1^2}$ and $\lambda_1^2$ and where $q_{k,j}(\gamma_{k,j}^2 \mid \sigma^2)$ is the density of the reciprocal of an Inverse-Gaussian distribution with parameters $\sqrt{\lambda_2^2\sigma^2/d_2^2}$ and $\lambda_2^2$.

In addition, since each $\tau_{0,k}^2 \leq 4d/\lambda_1^2$ and each $\gamma_{0,k,j}^2 \leq 4d/\lambda_2^2$, so

$$\left(\frac{1}{\tau_{0,k}^2} + \frac{1}{\gamma_{0,k,j}^2}\right)^{-1} \leq \left(\frac{\lambda_1^2}{4d} + \frac{\lambda_2^2}{4d}\right)^{-1} := d_3.$$

By Lemma D.5

$$(y - X\beta_0)^T(y - X\beta_0) + \beta_0^T V_{\tau_0,\gamma_0}^{-1}\beta_0 = y^T y - y^T X(X^T X + V_{\tau_0,\gamma_0}^{-1})^{-1} X^T y$$

$$\geq y^T y - y^T X \left( X^T X + \frac{1}{d_3} I_p \right)^{-1} X^T y \tag{D.31}$$

Using (D.31),

$$\begin{aligned}
&\exp\left\{ -\frac{1}{2} - \frac{p^2\lambda_2^2 d_2^2}{2\sigma^2} - \frac{1}{2} - \frac{K^2\lambda_1^2 d_1^2}{2\sigma^2} \right\} f(\sigma^2 \mid \beta_0, \tau_0^2, \gamma_0^2, y) \\
&= \exp\left\{ -1 - \frac{p^2\lambda_2^2 d_2^2}{2\sigma^2} - \frac{K^2\lambda_1^2 d_1^2}{2\sigma^2} \right\} \frac{\left( \frac{(y-X\beta_0)^T(y-X\beta_0) + \beta_0^T V_{\tau_0,\gamma_0}^{-1}\beta_0 + 2\xi}{2} \right)^{\frac{n+p}{2}+\alpha}}{\Gamma\left(\frac{n+p}{2}+\alpha\right)} \\
&\quad \times \left(\sigma^2\right)^{-\frac{n+p}{2}-\alpha-1} \exp\left\{ -\frac{(y-X\beta_0)^T(y-X\beta_0) + \beta_0^T V_{\tau_0,\gamma_0}^{-1}\beta_0 + 2\xi}{2\sigma^2} \right\} \\
&\geq e^{-1} \left( \frac{y^T y - y^T X (X^T X + d_3^{-1} I_p)^{-1} X^T y + 2\xi}{2} \right)^{\frac{n+p}{2}+\alpha} \frac{1}{\Gamma\left(\frac{n+p}{2}+\alpha\right)} \\
&\quad \times \left(\sigma^2\right)^{-\frac{n+p}{2}-\alpha-1} \exp\left\{ -\frac{d + 2\xi + p^2\lambda_2^2 d_2^2 + K^2\lambda_1^2 d_1^2}{2\sigma^2} \right\} \\
&= e^{-1} \left( \frac{y^T y - y^T X (X^T X + d_3^{-1} I_p)^{-1} X^T y + 2\xi}{d + 2\xi + p^2\lambda_2^2 d_2^2 + K^2\lambda_1^2 d_1^2} \right)^{\frac{n+p}{2}+\alpha} \\
&\quad \times \frac{\left( \frac{d + 2\xi + p^2\lambda_2^2 d_2^2 + K^2\lambda_1^2 d_1^2}{2} \right)^{\frac{n+p}{2}+\alpha}}{\Gamma\left(\frac{n+p}{2}+\alpha\right)} \\
&\quad \times \left(\sigma^2\right)^{-\frac{n+p}{2}-\alpha-1} \exp\left\{ -\frac{d + 2\xi + p^2\lambda_2^2 d_2^2 + K^2\lambda_1^2 d_1^2}{2\sigma^2} \right\} \\
&= e^{-1} \left( \frac{y^T y - y^T X (X^T X + d_3^{-1} I_p)^{-1} X^T y + 2\xi}{d + 2\xi + p^2\lambda_2^2 d_2^2 + K^2\lambda_1^2 d_1^2} \right)^{\frac{n+p}{2}+\alpha} q(\sigma^2),
\end{aligned} \tag{D.32}$$

where $q(\sigma^2)$ is the density of the Inverse-Gamma distribution with parameters, $(n + p)/2 + \alpha$ and $d + 2\xi + p^2\lambda_2^2 d_2^2 + K^2\lambda_1^2 d_1^2$.

Finally, using (D.21) and (D.22),

$$k_{BSGL}(\beta, \tau^2, \gamma^2, \sigma^2 \mid \beta_0, \tau_0^2, \gamma_0^2, \sigma_0^2)$$

$$\geq \epsilon\, f(\beta \mid \tau^2, \gamma^2, \sigma^2, y)\, q(\sigma^2) \prod_{k=1}^{K} \left[ q_k(\tau_k^2 \mid \sigma^2) \prod_{j=1}^{m_k} q_{k,j}(\gamma_{k,j}^2 \mid \sigma^2) \right],$$

where

$$\epsilon = e^{-1} \left( \frac{y^T y - y^T X (X^T X + d_3^{-1} I_p)^{-1} X^T y + 2\xi}{d + 2\xi + p^2 \lambda_2^2 d_2^2 + K^2 \lambda_1^2 d_1^2} \right)^{\frac{n+p}{2}+\alpha}. \tag{D.33}$$

## D.7.1 Starting Values

The starting values satisfy,

$$(\beta_0, \tau_0^2, \gamma_0^2, \sigma_0^2) = \arg\min \;\; V_{BSGL}(\beta, \tau^2, \gamma^2, \sigma^2).$$

To minimize $V_{BSGL}$,

$$\frac{\partial V_{BSGL}}{\partial \tau_{0,k}^2} = 0 \Rightarrow = -\frac{\beta_{0,G_k}^T \beta_{0,G_k}}{\tau_{0,k}^4} + \frac{\lambda_1^2}{4} = 0 \Rightarrow \tau_{0,k}^2 = \sqrt{\frac{4\beta_{0,G_k}^T \beta_{0,G_k}}{\lambda_1^2}}$$
$$\frac{\partial V_{BSGL}}{\partial \gamma_{0,k,j}^2} = 0 \Rightarrow = -\frac{\beta_{0,k,j}^2}{\gamma_{0,k,j}^4} + \frac{\lambda_2^2}{4} = 0 \Rightarrow \gamma_{0,k,j}^2 = \sqrt{\frac{4\beta_{0,k,j}^2}{\lambda_2^2}}$$

$$\begin{aligned}
\beta_0 = \arg\min_{\beta \in \mathbb{R}^p} \Bigg\{ & (y - X\beta)^T (y - X\beta) + \sum_{k=1}^{K} \frac{\lambda_1 \beta_{G_k}^T \beta_{G_k}}{2\sqrt{\beta_{G_k}^T \beta_{G_k}}} + \sum_{k=1}^{K} \sum_{j=1}^{m_k} \frac{\lambda_2 \beta_{k,j}^2}{2\sqrt{\beta_{k,j}^2}} \\
& + \sum_{k=1}^{K} \frac{\lambda_1^2}{4} \sqrt{\frac{4\beta_{G_k}^T \beta_{G_k}}{\lambda_1^2}} + \frac{\lambda_2^2}{4} \sum_{k=1}^{K} \sum_{j=1}^{m_k} \sqrt{\frac{4\beta_{k,j}^2}{\lambda_2^2}} \Bigg\} \\
= \arg\min_{\beta \in \mathbb{R}^p} \; & (y - X\beta)^T (y - X\beta) + \lambda_1 \sum_{k=1}^{K} \sqrt{\beta_{G_k}^T \beta_{G_k}} + \lambda_2 \sum_{k=1}^{K} \sum_{j=1}^{m_k} |\beta_{k,j}|,
\end{aligned}$$

which corresponds to the sparse group lasso solutions. Thus a reasonable starting value for the Markov chain is $\beta_0$ being the sparse group lasso estimate, $\tau_{0,k}^2 = 2\sqrt{\beta_{0,G_k}^T \beta_{0,G_k}}/\lambda_1$ and $\gamma_{0,k}^2 = 2|\beta_{0,k,j}|/\lambda_2$.

# References


Acosta, F., Huber, M. L., and Jones, G. L. (2015). Markov chain Monte Carlo with linchpin variables. *Preprint*.

Andelić, M. and Da Fonseca, C. (2011). Sufficient conditions for positive definiteness of tridiagonal matrices revisited. *Positivity*, 15:155–159.

Anderson, T. W. (1971). *The Statistical Analysis of Time Series.* John Wiley & Son, New York.

Andrews, D. W. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, 59:817–858.

Atchadé, Y. F. (2011). Kernel estimators of asymptotic variance for adaptive Markov chain Monte Carlo. *The Annals of Statistics*, 39:990–1011.

Atchadé, Y. F. (2016). Markov chain Monte Carlo confidence intervals. *Bernoulli*, 22:1808–1838.

Athreya, K. B., Doss, H., and Sethuraman, J. (1996). On the convergence of the Markov chain simulation method. *The Annals of Statistics*, 24:69–100.

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*, 25:468–474.

Bednorz, W. and Łatuszyński, 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.

Berkes, I. and Philipp, W. (1979). Approximation theorems for independent and weakly dependent random vectors. *The Annals of Probability*, 7:29–54.

Bhattacharya, A., Chakraborty, A., and Mallick, B. K. (2015). Fast sampling with Gaussian scale-mixture priors in high-dimensional regression. *arXiv preprint arXiv:1506.04778*.

Billingsley, P. (1968). *Convergence of Probability Measures*. Wiley, New York.

Billingsley, P. (2008). *Probability and Measure*. John Wiley & Sons, New York.

Blei, D. M. and Lafferty, J. D. (2009). Topic models. *Text Mining: Classification, Clustering, and Applications*, 10:34.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

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

Charnes, J. M. (1995). Analyzing multivariate output. In *Proceedings of the 27th conference on Winter simulation*, pages 201–208. IEEE Computer Society.

Chen, D.-F. R. and Seila, A. F. (1987). Multivariate inference in stationary simulation using batch means. In *Proceedings of the 19th conference on Winter simulation*, pages 302–304. ACM.

Chow, Y. S. and Teicher, H. (1978). *Probability Theory*. Springer-Verlag, New York.

Csörgő, M. and Révész, P. (1981). *Strong Approximations in Probability and Statistics*. Academic Press.

Dai, N. and Jones, G. (2016). Multivariate initial sequence estimators in Markov chain Monte Carlo. *arXiv*.

Damerdji, H. (1991). Strong consistency and other properties of the spectral variance estimator. *Management Science*, 37:1424–1440.

Damerdji, H. (1994). Strong consistency of the variance estimator in steady-state simulation output analysis. *Mathematics of Operations Research*, 19:494–512.

Damerdji, H. (1995). Mean-square consistency of the variance estimator in steady-state simulation output analysis. *Operations Research*, 43:282–291.

De Jong, R. M. (2000). A strong consistency proof for heteroskedasticity and autocorrelation consistent covariance matrix estimators. *Econometric Theory*, 16:262–268.

Dehling, H. and Philipp, W. (1982). Almost sure invariance principles for weakly dependent vector-valued random variables. *The Annals of Probability*, 10:689–701.

Doss, H. and Hobert, J. P. (2010). Estimation of Bayes factors in a class of hierarchical random effects models using a geometrically ergodic MCMC algorithm. *Journal of Computational and Graphical Statistics*, 19:295–312.

Drummond, A. J., Ho, S. Y., Phillips, M. J., and Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. *PLoS Biology*, 4:699.

Eberlein, E. (1986). On strong invariance principles under dependence assumptions. *The Annals of Probability*, 14:260–270.

Einmahl, U. (1989). Extensions of results of Komlós, Major, and Tusnády to the multivariate case. *Journal of Multivariate Analysis*, 28:20–68.

Finley, A., Banerjee, S., and Gelfand, A. (2015). spbayes for large univariate and multivariate point-referenced spatio-temporal data models. *Journal of Statistical Software*, 63:1–28.

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., Hughes, J., and Vats, D. (2015). *mcmcse: Monte Carlo Standard Errors for MCMC*. Riverside, CA and Minneapolis, MN. R package version 1.1-2.

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.

Franklin, J. N. (2012). *Matrix theory*. Courier Corporation.

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

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

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

Geyer, C. J. (1992). Practical Markov chain Monte Carlo (with discussion). *Statistical Science*, 7:473–511.

Geyer, C. J. (1995). Conditioning in Markov chain Monte Carlo. *Journal of Computational and Graphical Statistics*, 4:148–154.

Giordano, R., Broderick, T., and Jordan, M. (2015). Linear response methods for accurate covariance estimates from mean field variational Bayes. In *Advances in Neural Information Processing Systems*, pages 1441–1449.

Glynn, P. W. and Iglehart, D. L. (1990). Simulation output analysis using standardized time series. *Mathematics of Operations Research*, 15:1–16.

Glynn, P. W. and Whitt, W. (1991). Estimating the asymptotic variance with batch means. *Operations Research Letters*, 10:431–435.

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. (2016). A practical sequential stopping rule for high-dimensional Markov chain Monte Carlo. *Journal of Computational and Graphical Statistics*, 25:684–700.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101:5228–5235.

Horvath, L. (1984). Strong approximation of extended renewal processes. *The Annals of Probability*, 12:1149–1166.

Ibragimov, I. A. and Linnik, Y. V. (1971). *Independent and Stationary Sequences of Random Variables*. Walters-Noordhoff, The Netherlands.

Ishwaran, H. and Rao, J. S. (2005). Spike and slab variable selection: frequentist and Bayesian strategies. *Annals of Statistics*, pages 730–773.

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

Johnson, A. A. and Jones, G. L. (2015). Geometric ergodicity of random scan Gibbs samplers for hierarchical one-way random effects models. *Journal of Multivariate Analysis*, 140:325–342.

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.

Jones, G. L. and Hobert, J. P. (2004). Sufficient burn-in for Gibbs samplers for a hierarchical random effects model. *The Annals of Statistics*, 32:784–817.

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*, 52:93–100.

Kendall, M. G. and Stuart, A. (1963). *The Advanced Theory of Statistics*, volume 1. Charles Griffen & Company Limited, London, second edition.

Khare, K. and Hobert, J. P. (2012). Geometric ergodicity of the Gibbs sampler for Bayesian quantile regression. *Journal of Multivariate Analysis*, 112:108–116.

Khare, K. and Hobert, J. P. (2013). Geometric ergodicity of the Bayesian lasso. *Electronic Journal of Statistics*, 7:2150–2163.

Komlós, J., Major, P., and Tusnády, G. (1975). An approximation of partial sums of independent RV'-s, and the sample DF. I. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 32:111–131.

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.

Kuelbs, J. and Philipp, W. (1980). Almost sure invariance principles for partial sums of mixing B-valued random variables. *The Annals of Probability*, 8:1003–1036.

Kyung, M., Gill, J., Ghosh, M., and Casella, G. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis*, 5:369–411.

Liu, J. S. (2008). *Monte Carlo Strategies in Scientific Computing*. Springer, New York.

Marcus, M. and Minc, H. (1992). *A survey of matrix theory and matrix inequalities*, volume 14. Courier Corporation.

Merlevède, F. and Rio, E. (2015). Strong approximation for additive functionals of geometrically ergodic Markov chains. *Electronic Journal of Probability*, 20.

Meyn, S. P. and Tweedie, R. L. (2009). *Markov Chains and Stochastic Stability*. Cambridge University Press.

Muñoz, D. F. and Glynn, P. W. (2001). Multivariate standardized time series for steady-state simulation output analysis. *Operations Research*, 49:413–422.

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

Narisetty, N. N. and He, X. (2014). Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics*, 42:789–817.

Osborne, B. G., Fearn, T., Miller, A. R., and Douglas, S. (1984). Application of near infrared reflectance spectroscopy to the compositional analysis of biscuits and biscuit doughs. *Journal of the Science of Food and Agriculture*, 35:99–105.

Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, 103:681–686.

Pazhayidam George, C. (2015). *Latent Dirichlet Allocation: Hyperparameter selection and applications to electronic discovery*. PhD thesis, University of Florida, Gainesville.

Philipp, W. and Stout, W. F. (1975). *Almost Sure Invariance Principles for Partial Sums of Weakly Dependent Random Variables*, volume 161. American Mathematical Society.

Robert, C. P. and Casella, G. (2013). *Monte Carlo Statistical Methods*. Springer, New York.

Roberts, G. O., Gelman, A., and Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, 7:110–120.

Roberts, G. O. and Rosenthal, J. S. (2001). Markov chains and de-initializing processes. *Scandinavian Journal of Statistics*, 28:489–504.

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

Roy, V. and Chakraborty, S. (2016). Selection of tuning parameters, solution paths and standard errors for Bayesian lassos. *Bayesian Analysis,* to appear.

Seila, A. F. (1982). Multivariate estimation in regenerative simulation. *Operations Research Letters*, 1:153–156.

SenGupta, A. (1987). Tests for standardized generalized variances of multivariate normal populations of possibly different dimensions. *Journal of Multivariate Analysis*, 23:209–219.

Strassen, V. (1964). An invariance principle for the law of the iterated logarithm. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 3:211–226.

Tjøstheim, D. (1990). Non-linear time series and Markov chains. *Advances in Applied Probability*, 22:587–611.

Vats, D. (2016). Geometric ergodicity of Gibbs samplers in Bayesian penalized regression models. *arXiv preprint arXiv:1609.04057*.

Vats, D., Flegal, J. M., and Jones, G. L. (2016a). Multivariate output analysis for Markov chain Monte Carlo. *arXiv preprint arXiv:1512.07713*.

Vats, D., Flegal, J. M., and Jones, G. L. (2016b). Strong consistency of multivariate spectral variance estimators in Markov chain Monte Carlo. *Bernoulli,* to appear.

Whittle, P. (1960). Bounds for the moments of linear and quadratic forms in independent variables. *Theory of Probability & Its Applications*, 5:302–305.

Wilks, S. S. (1932). Certain generalizations in the analysis of variance. *Biometrika*, pages 471–494.

Wu, W. B. (2007). Strong invariance principles for dependent random variables. *The Annals of Probability*, 35:2294–2320.

Xu, X. and Ghosh, M. (2015). Bayesian variable selection and estimation for group lasso. *Bayesian Analysis*, 10:909–936.

Zaitsev, A. Y. (1998). Multidimensional version of the results of Komlós and Tusnády for vectors with finite exponential moments. *ESAIM: Probability and Statistics*, 2:41–108.