Monte Carlo Standard Errors for Markov Chain Monte Carlo

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

James Marshall Flegal

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

Galin L. Jones and
Glen D. Meeden, Advisers

July 2008

Acknowledgements

I would like to thank my adviser Galin Jones for his guidance, friendship, encouragement, and patience throughout my studies at the University of Minnesota. Beginning my first day of class through final corrections on my thesis, Galin challenged me to always continue to learn. Fortunately, I've come a long way since that first day, and I look forward to continuing the journey. I would also like to thank my other committee members Glen Meeden, Charlie Geyer, and Jim Hodges for their research guidance during my final two years. Without their generous support, this would not have been possible.

I also owe a debt of gratitude to the faculty, staff, and fellow students in the School of Statistics during my time at Minnesota. In particularly, I would like to thank Gary Oehlert, the Director of Graduate Studies who saw potential in an unusual application to the School of Statistics.

I an extremely grateful to the Kalamazoo Foundation, specifically all of the contributors to the Clarence L. Remynse Scholarship whose aid was greatly appreciated.

Finally, I am thankful to both friends and family whose support helped me get through the good and bad times during my years in Minnesota. I am especially grateful to Amanda Isaacson for her daily support and encourage, particularly during the final days writing my thesis.

Thank you, endless gratitude is due to all of you.

Abstract

Markov chain Monte Carlo (MCMC) is a method of producing a correlated sample to estimate characteristics of a target distribution. A fundamental question is how long should the simulation be run? One method to address this issue is to run the simulation until the width of a confidence interval for the quantity of interest is below a user-specified value. The use of this fixed-width methods requires an estimate of the Monte Carlo standard error (MCSE). This dissertation begins by discussing why MCSEs are important, how they can be easily calculated in MCMC and how they can be used to decide when to stop the simulation. The use of MCSEs is then compared to a popular alternative in the context of multiple examples.

This dissertation continues by discussing the relevant Markov chain theory with particular attention paid to the conditions and definitions needed to establish a Markov chain central limit theorem. Estimating MCSEs requires estimating the associated asymptotic variance. I introduce several techniques for estimating MCSEs: batch means, overlapping batch means, regeneration, subsampling and spectral variance estimation. Asymptotic properties useful in MCMC settings are established for these variance estimators. Specifically, I established conditions under which the estimator for the asymptotic variance in a Markov chain central limit theorem is strongly consistent. Strong consistency ensures that confidence intervals formed will be asymptotically valid. In addition, I established conditions to ensure mean-square consistency for the estimators using batch means and overlapping batch means. Mean-square con-

sistency is useful in choosing an optimal batch size for MCMC practitioners.

Several approaches have been introduced dealing with the special case of estimating ergodic averages and their corresponding standard errors. Surprisingly, very little attention has been given to characteristics of the target distribution that cannot be represented as ergodic averages. To this end, I proposed use of subsampling methods as a means for estimating the qth quantile of the posterior distribution. Finally, the finite sample properties of subsampling are examined.

Contents

List of Tables

List of Figures

Chapter 1

Motivation: Can We Trust the Third Significant Figure?

Current reporting of results based on Markov chain Monte Carlo computations could be improved. In particular, a measure of the accuracy of the resulting estimates is rarely reported. Thus we have little ability to objectively assess the quality of the reported estimates. We address this issue in that we discuss why Monte Carlo standard errors are important, how they can be easily calculated in Markov chain Monte Carlo and how they can be used to decide when to stop the simulation. We compare their use to a popular alternative in the context of two examples.

The content of this chapter is primarily contained in Flegal et al. (2008) and serves as an introduction to the problem of interest. The results here are expanded and formalized in subsequent chapters.

1.1 Introduction

Hoaglin and Andrews (1975) consider the general problem of what information should be included in publishing computation-based results. The goal of their suggestions was "...to make it easy for the reader to make reasonable assessments of the numerical quality of the results." In particular, Hoaglin and Andrews suggested that it is crucial

to report some notion of the accuracy of the results and, for Monte Carlo studies this should include estimated standard errors. However, in settings where Markov chain Monte Carlo (MCMC) is used there is a culture of rarely reporting such information. For example, we looked at the issues published in 2006 of *Journal of the American Statistical Association*, *Biometrika* and *Journal of the Royal Statistical Society, Series B*. In these journals we found 39 papers that used MCMC. Only 3 of them directly addressed the Monte Carlo error in the reported estimates. Thus it is apparent that the readers of the other papers have little ability to objectively assess the quality of the reported estimates. We attempt to address this issue in that we discuss why Monte Carlo standard errors are important, how they can be easily calculated in MCMC settings and compare their use to a popular alternative.

Simply put, MCMC is a method for using a computer to generate data and subsequently using standard large sample statistical methods to estimate fixed, unknown quantities of a given target distribution. (Thus, we object to calling it 'Bayesian Computation'.) That is, it is used to produce a point estimate of some characteristic of a target distribution π having support X. The most common use of MCMC is to estimate $E_\pi g := \int_{\mathsf{X}} g(x)\,\pi(dx)$ where g is a real-valued, π-integrable function on X.

Suppose that $X = \{X_1, X_2, X_3, \dots\}$ is an aperiodic, irreducible, positive Harris recurrent Markov chain with state space X and invariant distribution π (for definitions see Section 2.1). In this case X is *Harris ergodic*. Typically, estimating $E_\pi g$ is natural since an appeal to the Ergodic Theorem implies that if $E_\pi|g| < \infty$ then, with probability 1,

$$\bar{g}_n := \frac{1}{n}\sum_{i=1}^{n} g(X_i) \to E_\pi g \quad \text{as } n \to \infty.$$

The MCMC method entails constructing a Markov chain X satisfying the regularity conditions described above and then simulating X for a finite number of steps, say n, and using $\bar{g}_n$ to estimate $E_\pi g$. The popularity of MCMC largely is due to the ease

with which such an X can be simulated (Chen et al., 2000; Robert and Casella, 1999; Liu, 2001).

An obvious question is when should we stop the simulation? That is, how large should n be? Or, when is $\bar{g}_n$ a good estimate of $E_\pi g$? In a given application we usually have an idea about how many significant figures we want in our estimate but how should this be assessed? Responsible statisticians and scientists want to do the right thing but output analysis in MCMC has become a muddled area with often conflicting advice and dubious terminology. This leaves many in a position where they feel forced to rely on intuition, folklore and heuristics. We believe this often leads to some poor practices: (A) Stopping the simulation too early, (B) Wasting potentially useful samples, and, most importantly, (C) Providing no notion of the quality of $\bar{g}_n$ as an estimate of $E_\pi g$. In this thesis we focus on issue (C) but touch briefly on (A) and (B).

The rest of this chapter is organized as follows. In Section 1.2 we briefly introduce some basic concepts from the theory of Markov chains. In Section 1.3 we consider estimating the Monte Carlo error of $\bar{g}_n$. Then Section 1.4 covers two methods for stopping the simulation and compares them in a toy example. In Section 1.5 the two methods are compared again in a realistic spatial model for a data set on wheat crop flowering dates in North Dakota. We close with some final remarks in Section 1.6.

1.2 Markov Chain Basics

Suppose that $X = \{X_1, X_2, X_3 \dots\}$ is a Harris ergodic Markov chain with state space X and invariant distribution π. For $n \in \mathbb{N} := \{1, 2, 3, \dots\}$ let $P^n(x, \cdot)$ be the n-step Markov transition kernel; that is, for $x \in \mathsf{X}$ and a measurable set A, $P^n(x, A) = \Pr(X_{n+i} \in A \mid X_i = x)$. An extremely useful property of X is that the

chain will converge to the invariant distribution. Specifically,

$$\|P^n(x,\cdot) - \pi(\cdot)\| \downarrow 0 \quad \text{as } n \to \infty,$$

where the left-hand side is the *total variation* distance between $P^n(x,\cdot)$ and $\pi(\cdot)$. (This is stronger than convergence in distribution.) The Markov chain X is *geometrically ergodic* if there exists a constant $0 < t < 1$ and a function $M : \mathsf{X} \to \mathbb{R}^+$ such that for any $x \in \mathsf{X}$,

$$\|P^n(x,\cdot) - \pi(\cdot)\| \leq M(x)\, t^n \tag{1.1}$$

for $n \in \mathbb{N}$. If $M(x)$ is bounded, then X is *uniformly ergodic*. Thus uniform ergodicity implies geometric ergodicity. However, as one might imagine, finding M and t directly is often quite difficult in realistic settings.

There has been a substantial amount of effort devoted to establishing (1.1) in MCMC settings. For example, Hobert and Geyer (1998), Johnson and Jones (2008), Jones and Hobert (2004), Marchev and Hobert (2004), Mira and Tierney (2002), Robert (1995), Roberts and Polson (1994), Roberts and Rosenthal (1999), Rosenthal (1995, 1996), Roy and Hobert (2007), and Tierney (1994) examined Gibbs samplers while Christensen et al. (2001), Douc et al. (2004), Fort and Moulines (2000), Fort and Moulines (2003), Geyer (1999), Jarner and Hansen (2000), Jarner and Roberts (2002), Meyn and Tweedie (1994), and Mengersen and Tweedie (1996) considered Metropolis-Hastings algorithms.

1.3 Monte Carlo Error

A Monte Carlo approximation is not exact. The number $\bar{g}_n$ is not the exact value of the integral we are trying to approximate. It is off by some amount, the *Monte Carlo error*, $\bar{g}_n - E_\pi g$. How large is the Monte Carlo error? Unfortunately, we can never

know unless we know $E_\pi g$.

We don't know the Monte Carlo error, but we can get a handle on its sampling distribution. That is, assessing the Monte Carlo error can be accomplished by estimating the variance of the asymptotic distribution of $\bar{g}_n$. Under regularity conditions, the Markov chain X and function g will admit a CLT. That is,

$$\sqrt{n}(\bar{g}_n - E_\pi g) \xrightarrow{d} \mathrm{N}(0, \sigma_g^2) \tag{1.2}$$

as $n \to \infty$ where $\sigma_g^2 := \mathrm{var}_\pi\{g(X_1)\} + 2\sum_{i=2}^{\infty} \mathrm{cov}_\pi\{g(X_1), g(X_i)\}$; the subscript π means that the expectations are calculated assuming $X_1 \sim \pi$. The CLT holds for *any* initial distribution when either (i) X is geometrically ergodic and $E_\pi|g|^{2+\delta} < \infty$ for some $\delta > 0$ or (ii) X is uniformly ergodic and $E_\pi g^2 < \infty$. These are not the only sufficient conditions for a CLT but are among the most straightforward to state; the interested reader is pointed to the summaries provided by Jones (2004) and Roberts and Rosenthal (2004).

Given a CLT we can assess the Monte Carlo error in $\bar{g}_n$ by estimating the variance, σ_g^2. That is, we can calculate and report an estimate of σ_g^2, say $\hat{\sigma}_g^2$ that will allow us to assess the accuracy of the point estimate. There have been many techniques introduced for estimating σ_g^2; see, among others, Bratley et al. (1987), Fishman (1996), Geyer (1992), Glynn and Iglehart (1990), Glynn and Whitt (1991), Mykland et al. (1995) and Roberts (1996). For example, regenerative simulation, batch means and spectral variance estimators all can be appropriate in MCMC settings. In this chapter, we will consider only one of the available methods, namely non-overlapping batch means. We chose this method is because it is easy to implement and can enjoy desirable theoretical properties. However, overlapping batch means has a reputation of sometimes being more efficient than non-overlapping batch means (a relationship investigated in Chapter 3).

1.3.1 Batch Means

In non-overlapping batch means the output is broken into blocks of equal size. Suppose the algorithm is run for a total of $n = ab$ iterations (hence $a = a_n$ and $b = b_n$ are implicit functions of n) and define

$$\bar{Y}_j := \frac{1}{b} \sum_{i=(j-1)b+1}^{jb} g(X_i) \qquad \text{for } j = 1, \ldots, a \, .$$

The batch means estimate of σ_g^2 is

$$\hat{\sigma}_g^2 = \frac{b}{a-1} \sum_{j=1}^{a} (\bar{Y}_j - \bar{g}_n)^2 \, . \tag{1.3}$$

Batch means is attractive because it is easy to implement (and it is available in some software, e.g. WinBUGS) but some authors encourage caution in its use (Roberts, 1996). In particular, we believe careful use is warranted since (1.3), in general, is not a consistent estimator of σ_g^2. On the other hand, Jones et al. (2006) showed that if the batch size and the number of batches are allowed to increase as the overall length of the simulation increases by setting $b_n = \lfloor n^\theta \rfloor$ and $a_n = \lfloor n/b_n \rfloor$, then $\hat{\sigma}_g^2 \to \sigma_g^2$ with probability 1 as $n \to \infty$. Throughout this thesis, we consider the case consistent batch means (BM) rather than the usual fixed number of batches version. The regularity conditions require that X be geometrically ergodic, $E_\pi |g|^{2+\delta+\epsilon} < \infty$ for some $\delta > 0$ and some $\epsilon > 0$ and $(1 + \delta/2)^{-1} < \theta < 1$; often $\theta = 1/2$ (i.e., $b_n = \lfloor \sqrt{n} \rfloor$ and $a_n = \lfloor n/b_n \rfloor$) is a convenient choice that works well in applications. Note that the only practical difference between BM and standard batch means is that the batch number and size are chosen as functions of the overall run length, n.

Using BM to get an estimate of the Monte Carlo standard error (MCSE) of $\bar{g}_n$, say $\hat{\sigma}_g/\sqrt{n}$, we can form an asymptotically valid confidence interval for $E_\pi g$. The

half-width of the interval is given by

$$t_{a_n-1}\,\frac{\hat{\sigma}_g}{\sqrt{n}} \tag{1.4}$$

where t_{a_n-1} is an appropriate quantile from Student's t distribution with a_n-1 degrees of freedom.

1.3.2 How Many Significant Figures?

The title of the chapter contains a rhetorical question; we don't always care about the *third* significant figure. But we should care about how many significant figures there are in our estimates. Assessing the Monte Carlo error through (1.4) gives us a tool to do this. For example, suppose $\bar{g}_n = 0.02$, then there is exactly one significant figure in the estimate, namely the "2", but how confident are we about it? Letting h_α denote the half width given in (1.4) of a $(1-\alpha)100\%$ interval, we would trust the one significant figure in our estimate if $0.02 \pm h_\alpha \subseteq [0.015, 0.025)$ since otherwise values such as $E_\pi g = 0.01$ or $E_\pi g = 0.03$ are plausible through rounding.

More generally, we can use (1.4) to assess how many significant figures we have in our estimates. This is illustrated in the following toy example that will be used several times throughout the rest of this thesis.

Toy Example

Let $Y_1, \ldots, Y_K$ be i.i.d. $\mathrm{N}(\mu, \lambda)$ and let the prior for (μ, λ) be proportional to $1/\sqrt{\lambda}$. The posterior density is characterized by

$$\pi(\mu, \lambda | y) \propto \lambda^{-\frac{K+1}{2}} \exp\left\{-\frac{1}{2\lambda}\sum_{j=1}^{K}(y_j - \mu)^2\right\} \tag{1.5}$$

where $y = (y_1, \ldots, y_K)^T$. It is easy to check that this posterior is proper as long as $K \geq 3$ and we assume this throughout. Using the Gibbs sampler to make draws from (1.5) requires the full conditional densities, $f(\mu|\lambda, y)$ and $f(\lambda|\mu, y)$, which are as follows:

$$\begin{aligned} \mu|\lambda, y &\sim \mathrm{N}(\bar{y}, \lambda/K) \,, \\ \lambda|\mu, y &\sim \mathrm{IG}\left(\frac{K-1}{2}, \frac{(K-1)s^2 + K(\bar{y}-\mu)^2}{2}\right) \,, \end{aligned}$$

where $\bar{y}$ is the sample mean and $(K-1)s^2 = \sum(y_i - \bar{y})^2$. (We say $W \sim \mathrm{IG}(\alpha, \beta)$ if its density is proportional to $w^{-(\alpha+1)}e^{-\beta/w}I(w > 0)$.) Consider estimating the posterior means of μ and λ. It is easy to prove that $E(\mu|y) = \bar{y}$ and $E(\lambda|y) = (K-1)s^2/(K-4)$ for $K > 4$. Thus we do not need MCMC to estimate these quantities but we will ignore this and use the output of a Gibbs sampler to estimate $E(\mu|y)$ and $E(\lambda|y)$.

Consider the Gibbs sampler that updates λ then μ; that is, letting (λ', μ') denote the current state and (λ, μ) denote the future state, the transition looks like $(\lambda', \mu') \to (\lambda, \mu') \to (\lambda, \mu)$. Jones and Hobert (2001) established that the associated Markov chain is geometrically ergodic as long as $K \geq 5$. If $K > 10$, then the moment conditions ensuring the CLT and the regularity conditions for BM (with $\theta = 1/2$) hold.

Let $K = 11$, $\bar{y} = 1$, and $(K-1)s^2 = 14$ so that $E(\mu|y) = 1$ and $E(\lambda|y) = 2$; for the remainder of this chapter these settings will be used every time we consider this example. Consider estimating $E(\mu|y)$ and $E(\lambda|y)$ with $\bar{\mu}_n$ and $\bar{\lambda}_n$, respectively and using BM to calculate the MCSEs for each estimate. Specifically, we will use a 95% confidence level in (1.4) to construct an interval estimate. Let the initial value for the simulation be $(\lambda_1, \mu_1) = (1, 1)$. When we ran the Gibbs sampler for 1000 iterations we obtained $\bar{\lambda}_{1000} = 2.003$ with an MCSE of 0.055 and $\bar{\mu}_{1000} = 0.99$ with an MCSE of 0.016. Thus we would be comfortable reporting two significant figures

for the estimates of $E(\lambda|y)$ and $E(\mu|y)$, specifically 2.0 and 1.0, respectively. But when we started from $(\lambda_1, \mu_1) = (100, 100)$ and ran Gibbs for 1000 iterations we obtained $\bar{\lambda}_{1000} = 13.06$ with an MCSE of 11.01 and $\bar{\mu}_{1000} = 1.06$ with an MCSE of 0.071. Thus we would not be comfortable with *any* significant figures for the estimate of $E(\lambda|y)$ but we would trust one significant figure (i.e., 1) for $E(\mu|y)$. Unless the MCSE is calculated and reported a hypothetical reader would have no way to judge this independently.

Remarks

1. A common concern about MCSEs is that their use may require estimating $E_\pi g$ much too precisely relative to $\sqrt{\text{var}_\pi g}$. Of course, it would be a rare problem indeed where we would know $\sqrt{\text{var}_\pi g}$ and not $E_\pi g$. Thus we would need to estimate $\sqrt{\text{var}_\pi g}$ and calculate an MCSE (via the delta method) before we could trust the estimate of $\sqrt{\text{var}_\pi g}$ to inform us about the MCSE for $E_\pi g$.

2. We are not suggesting that all MCMC-based estimates should be reported in terms of significant figures; in fact we do not do this in the simulations that occur later. Instead, we are strongly suggesting that an estimate of the Monte Carlo standard error should be used to assess simulation error and reported. Without an attached MCSE a point estimate should not be trusted.

1.4 Stopping the Simulation

In this section we consider two formal approaches to terminating the simulation. The first is based on calculating an MCSE and is discussed in Section 1.4.1. The second is based on the method introduced in Gelman and Rubin (1992) and is one of many so-called convergence diagnostics (Cowles and Carlin, 1996). Our reason for choosing the Gelman-Rubin diagnostic (GRD) is that it appears to be far and away the most

popular method for stopping the simulation. GRD and MCSE are used to stop the simulation in a similar manner. After n iterations either the value of the GRD or MCSE is calculated and if it isn't sufficiently small then we continue the simulation until it is.

1.4.1 Fixed-Width Methods

Suppose we have an idea of how many significant figures we want in our estimate. Another way of saying this is that we want the half-width of the interval (1.4) to be less than some user-specified value, ϵ. Thus we might consider stopping the simulation when the MCSE of $\bar{g}_n$ is sufficiently small. This, of course, means that we may have to check whether this criterion is met many times. It is not obvious that such a procedure will be guaranteed to terminate the computation in a finite amount of time or whether the resulting intervals will enjoy the desired coverage probability and half-width. Also, we don't want to check too early in the simulation since we will run the risk of premature termination due to a poor estimate of the standard error.

Suppose we use BM to estimate the Monte Carlo standard error of $\bar{g}_n$, say $\hat{\sigma}_g/\sqrt{n}$, and use it to form a confidence interval for $E_\pi g$. If this interval is too large then the value of n is increased and simulation continues until the interval is sufficiently small. Formally, the criterion is given by

$$t_{a_n-1}\,\frac{\hat{\sigma}_g}{\sqrt{n}} + p(n) \leq \epsilon \tag{1.6}$$

where t_{a_n-1} is an appropriate quantile, $p(n) = \epsilon I(n < n^*)$ where, $n^* > 0$ is fixed, I is the usual indicator function on $\mathbb{Z}_+$ and $\epsilon > 0$ is the user-specified half-width. The role of p is to ensure that the simulation is not terminated prematurely due to a poor estimate of $\hat{\sigma}_g$. The conditions which guarantee $\hat{\sigma}_g^2$ is consistent also imply that this procedure will terminate in a finite amount of time with probability one

and that the resulting intervals asymptotically have the desired coverage (see Glynn and Whitt, 1992). However, the finite-sample properties of (1.4) have received less formal investigation but simulation results suggest that the resulting intervals have approximately the desired coverage and half-width in practice (see Flegal and Jones, 2008; Jones et al., 2006).

Remarks

1. The CLT and BM require a geometrically ergodic Markov chain. This can be difficult to check directly in any given application. On the other hand, considerable effort has been spent establishing (1.1) for a number of Markov chains; see the references given at the end of Section 1.2. In our view, this is not the obstacle that it was in the past.

2. The frequency with which (1.6) should be evaluated is an open question. Checking often, say every few iterations, may substantially increase the overall computational effort.

3. Consider $p(n) = \epsilon I(n < n^*)$. The choice of n^* is often made based on the user's experience with the problem at hand. However, for geometrically ergodic Markov chains there is some theory that can give guidance on this issue (see Jones and Hobert, 2001; Latuszynski, 2008; Rosenthal, 1995).

4. Stationarity of the Markov chain is not required for the CLT or the strong consistency of BM. One consequence is that burn-in is not required if we can find a reasonable starting value.

Toy Example

We consider implementation of fixed-width methods in the toy example introduced in Section 1.3.2. We performed 1000 independent replications of the following procedure.

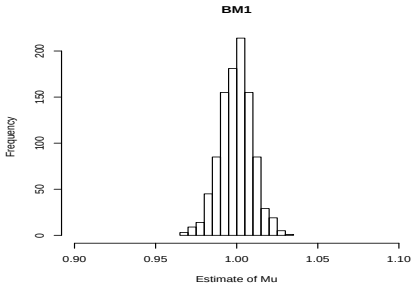
Method	Prop. at Min.	S.E.	Prop. ≤ 1000	S.E.	N	S.E.
BM1	0	0	0.011	0.0033	2191	19.9
BM2	0	0	0	0	5123	33.2
GRD1	0.576	0.016	0.987	0.0036	469	4.1
GRD2	0.587	0.016	0.993	0.0026	471	4.2
GRD3	0.062	0.0076	0.363	0.015	2300	83.5
GRD4	0.01	0.0031	0.083	0.0087	5365	150.5

Table 1.1: Summary of the proportion (and standard error) of the observed estimates which were based on the minimum number (400) of draws, less than or equal to 1000 draws, and the average total simulation effort for the toy example in Section 1.3.2.

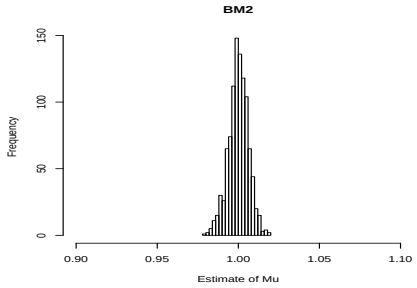
Each replication of the Gibbs sampler was started from $\bar{y}$. Using (1.6), a replication was terminated when the half-width of a 95% interval with $p(n) = \epsilon I(n < 400)$ was smaller than a prespecified cutoff, ϵ, for *both* parameters. If both standard errors were not less than the cutoff, then the current chain length was increased by 10% before checking again. We used two settings for the cutoff, $\epsilon = 0.06$ and $\epsilon = 0.04$. These settings will be denoted BM1 and BM2, respectively.

First, consider the estimates of $E(\mu|y)$. We can see from Figures 1.1a and 1.1b that the estimates of $E(\mu|y)$ are centered around the truth with both settings. Clearly, the cut-off of $\epsilon = 0.04$ is more stringent and yields estimates that are closer to the true value. It should come as no surprise that the cost of this added precision is increased computational effort; see Table 1.1. The corresponding plots for $\bar{\lambda}_n$ yield the same results and are therefore excluded.

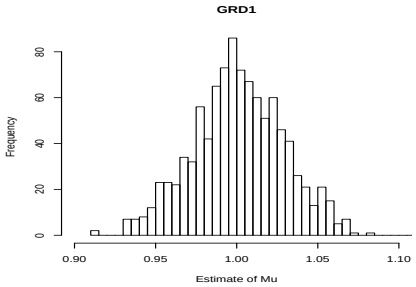
Consider BM2. In this case, 100% of the estimates, $\bar{\mu}_n$, of $E(\mu|y)$ and 96% of the estimates, $\bar{\lambda}_n$, of $E(\lambda|y)$ are within the specified $\epsilon = 0.04$ of the truth. In every replication the simulation was stopped when the criterion (1.6) for $E(\lambda|y)$ dropped below the cutoff. Similar results hold for the BM1 ($\epsilon = 0.06$) setting.



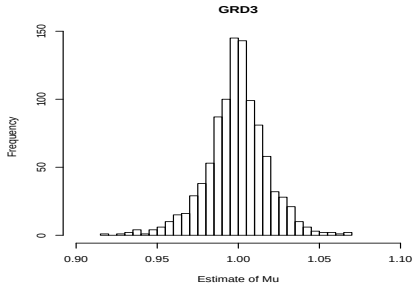
(a) BM1, with a cutoff of $\epsilon = 0.06$.



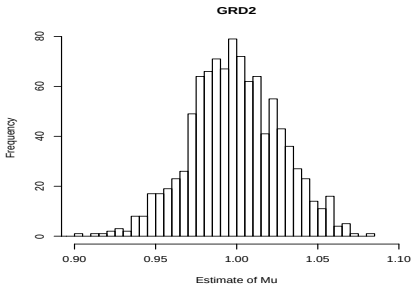
(b) BM2, with a cutoff of $\epsilon = 0.04$.



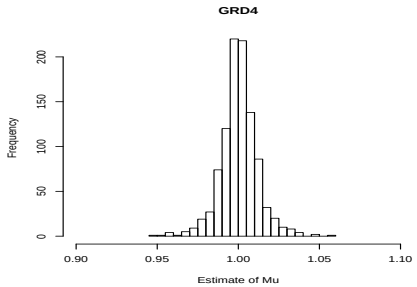
(c) GRD1, 2 chains and $\delta = 1.1$.



(d) GRD3, 2 chains and $\delta = 1.005$.



(e) GRD2, 4 chains and $\delta = 1.1$.



(f) GRD4, 4 chains and $\delta = 1.005$.

Figure 1.1: Histograms from 1000 replications estimating $E(\mu|y)$ for the toy example of Section 1.3.2 with BM and GRD. Simulation sample sizes are given in Table 1.1.

1.4.2 The Gelman-Rubin Diagnostic

The Gelman-Rubin diagnostic (GRD) introduced in Gelman and Rubin (1992) and refined by Brooks and Gelman (1998) is a popular method for assessing the output of MCMC algorithms. It is important to note that this method is also based on a Markov chain CLT (Gelman and Rubin, 1992, p.463) and hence does not apply more generally than approaches based on calculating an MCSE.

GRD is based on the simulation of m independent parallel Markov chains having invariant distribution π, each of length $2l$. Thus the total simulation effort is $2lm$. Gelman and Rubin (1992) suggest that the first l simulations should be discarded and inference based on the last l simulations; for the jth chain these are denoted $\{X_{1j}, X_{2j}, X_{3j}, \ldots, X_{lj}\}$ with $j = 1, 2, \ldots, m$. Recall that we are interested in estimating $E_\pi g$ and define $Y_{ij} = g(X_{ij})$,

$$B = \frac{l}{m-1}\sum_{j=1}^{m}(\bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot})^2 \qquad \text{and} \qquad W = \frac{1}{m}\sum_{j=1}^{m} s_j^2$$

where $\bar{Y}_{\cdot j} = l^{-1}\sum_{i=1}^{l} Y_{ij}$, $\bar{Y}_{\cdot\cdot} = m^{-1}\sum_{j=1}^{m}\bar{Y}_{\cdot j}$ and $s_j^2 = (l-1)^{-1}\sum_{i=1}^{l}(Y_{ij} - \bar{Y}_{\cdot j})^2$. Note that $\bar{Y}_{\cdot\cdot}$ is the resulting point estimate of $E_\pi g$. Let

$$\hat{V} = \frac{l-1}{l}W + \frac{(m+1)B}{ml}, \qquad d \approx \frac{2\hat{V}}{\hat{\text{var}}(\hat{V})},$$

and define the corrected *potential scale reduction factor*

$$\hat{R} = \sqrt{\frac{d+3}{d+1}\frac{\hat{V}}{W}}.$$

As noted by Gelman et al. (2004), $\hat{V}$ and W are essentially two different estimators of $\text{var}_\pi g$; *not* σ_g^2 from the Markov chain CLT. That is, neither $\hat{V}$ nor W address the sampling variability of $\bar{g}_n$ and hence neither does $\hat{R}$.

In our examples we used the R package `coda` which reports an upper bound on $\hat{R}$ (see Section 1.7.2). Specifically, a 97.5% upper bound for $\hat{R}$ is given by

$$\hat{R}_{0.975} = \sqrt{\frac{d+3}{d+1}\left[\frac{l-1}{l} + F_{0.975,m-1,w}\left(\frac{m+1}{ml}\frac{B}{W}\right)\right]},$$

where $F_{0.975,m-1,w}$ is the 97.5% percentile of an F_w^{m-1} distribution, $w = 2W^2/\hat{\sigma}_W^2$ and

$$\hat{\sigma}_W^2 = \frac{1}{m-1}\sum_{j=1}^{m}\left(s_j^2 - W\right)^2 .$$

In order to stop the simulation the user provides a cutoff, $\delta > 0$, and simulation continues until

$$\hat{R}_{0.975} + p(n) \leq \delta . \tag{1.7}$$

As with fixed-width methods, the role of $p(n)$ is to ensure that we do not stop the simulation prematurely due to a poor estimate, $\hat{R}_{0.975}$. By requiring a minimum total simulation effort of $n^* = 2lm$ we are effectively setting $p(n) = \delta I(n < n^*)$ where n indexes the total simulation effort.

Remarks

1. A rule of thumb suggested by Gelman et al. (2004) is to set $\delta = 1.1$. These authors also suggest that a value of δ closer to 1 will be desirable in a "final analysis in a critical problem" but give no further guidance. Since neither $\hat{R}$ nor $\hat{R}_{0.975}$ directly estimates the Monte Carlo error in $\bar{g}_n$ it is unclear to us that $\hat{R} \approx 1$ implies $\bar{g}_n$ is a good estimate of $E_\pi g$.

2. How large should m be? There seem to be few guidelines in the literature except that $m \geq 2$ since otherwise we cannot calculate B. Clearly, if m is too large then each chain will be too short to achieve any reasonable expectation of

convergence within a given computational effort.

3. The initial values, X_{j1}, of the m parallel chains should be drawn from an "over-dispersed" distribution. Gelman and Rubin (1992) suggest estimating the modes of π and then using a mixture distribution whose components are centered at these modes. Constructing this distribution could be difficult and is often not done in practice (Gelman et al., 2004, p. 593).

4. To our knowledge there has been no discussion in the literature about optimal choices of $p(n)$ or n^*. In particular, we know of no guidance about how long each of the parallel chains should be simulated before the first time we check that $\hat{R}_{0.975} < \delta$ or how often one should check after that. However, the same theoretical results that could give guidance in item 3 of Section 4.1.1 would apply here as well.

5. GRD was originally introduced simply as a method for determining an appropriate amount of burn-in. However, using diagnostics in this manner may introduce additional bias into the results, see Cowles et al. (1999).

Toy Example

We consider implementation of GRD in the toy example introduced in Section 1.3.2. The first issue we face is choosing the starting values for each of the m parallel chains. Notice that

$$\pi(\mu, \lambda|y) = g_1(\mu|\lambda)g_2(\lambda)$$

where $g_1(\mu|\lambda)$ is a $N(\bar{y}, \lambda/K)$ density and $g_2(\lambda)$ is an $IG((K-2)/2, (K-1)s^2/2)$ density. Thus we can sequentially sample the exact distribution by first drawing from $g_2(\lambda)$, and then conditionally, draw from $g_1(\mu|\lambda)$. We will use this to obtain starting values for each of the m parallel chains. Thus each of the m parallel Markov chains

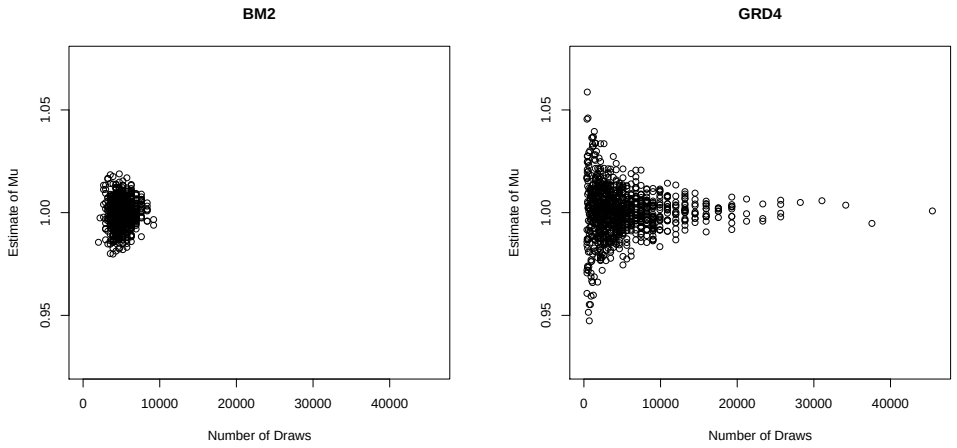
Method	Chains	Stopping Rule	MSE for $E(\mu\|y)$	S.E.	MSE for $E(\lambda\|y)$	S.E.
BM1	1	0.06	9.82e-05	4.7e-06	1.03e-03	4.5e-05
BM2	1	0.04	3.73e-05	1.8e-06	3.93e-04	1.8e-05
GRD1	2	1.1	7.99e-04	3.6e-05	8.7e-03	4e-04
GRD2	4	1.1	7.79e-04	3.7e-05	8.21e-03	3.6e-04
GRD3	2	1.005	3.49e-04	2.1e-05	3.68e-03	2e-04
GRD4	4	1.005	1.34e-04	9.2e-06	1.65e-03	1.2e-04

Table 1.2: Summary table for all settings and estimated mean-squared-error for estimating $E(\mu|y)$ and $E(\lambda|y)$ for the toy example of Section 1.3.2. Standard errors (S.E.) shown for each estimate.

will be stationary and hence GRD should be at a slight advantage compared to BM started from $\bar{y}$.

Our goal is to investigate the finite-sample properties of the GRD by considering the estimates of $E(\mu|y)$ and $E(\lambda|y)$ as in Section 1.4.1. To this end, we took multiple chains starting from different draws from the sequential sampler. The multiple chains were run until the total simulation effort was $n^* = 400$ draws; this is the same minimum simulation effort we required of BM in the previous section. If $\hat{R}_{0.975} < \delta$ for both, the simulation was stopped. Otherwise, 10% of the current chain length was added to each chain before $\hat{R}_{0.975}$ was recalculated. This continued until $\hat{R}_{0.975}$ was below δ for both. This simulation procedure was repeated independently 1000 times with each replication using the same initial values. We considered 4 settings using the combinations of $m \in \{2, 4\}$ and $\delta \in \{1.005, 1.1\}$. These settings will be denoted GRD1, GRD2, GRD3 and GRD4; see Table 1.2 for the different settings along with summary statistics that will be considered later.

Upon completion of each replication, the values of $\bar{\mu}_n$ and $\bar{\lambda}_n$ were recorded. Figures 1c–1f show histograms of $\bar{\mu}_n$ for each setting. We can see that all the settings center around the true value of 1, and setting $\delta = 1.005$ provides better estimates. Increasing the number of chains seems to have little impact on the quality of esti-



(a) $\bar{\mu}_n$ vs. n for BM2 (b) $\bar{\mu}_n$ vs. n for GRD4

Figure 1.2: Estimating $E(\mu|y)$ for the toy example of Section 1.3.2. Estimates of $E(\mu|y)$ versus number of draws for the BM2 and GRD4 settings.

mation, particularly when $\delta = 1.1$. Histograms of $\bar{\lambda}_n$ for each setting show similar trends.

In the settings we investigated, GRD often terminated the simulations much sooner than BM. Table 1.1 shows the percentage of the 1000 replications which were stopped at their minimum ($n^* = 400$) and the percentage with less than 1000 total draws. The data clearly shows that premature stopping was common with GRD but uncommon with BM. This is especially the case for GRD1 and GRD2 where over half the replications used the minimum simulation effort.

Also, the simulation effort for GRD was more variable than that of BM. In particular, the average simulation effort was comparable for BM1 and GRD3 and also BM2 and GRD4 but the standard errors are larger for GRD. Next, Figure 1.2 shows a plot of the estimates, $\bar{\mu}_n$, versus the total number of draws in the chains for BM2 and GRD4. The graphs clearly show that the total number of draws was more variable using GRD4 than using BM2. From a practical standpoint, this implies that when using GRD we are likely to run a simulation either too long or too short. Of course,

if we run the simulation too long, we will be likely to get a better estimate. But if the simulation is too short, the estimate can be poor.

Let's compare GRD and BM in terms of the quality of estimation. Table 1.2 gives the estimated mean-squared error (MSE) for each setting based on 1000 independent replications described above. The estimates for GRD were obtained using the methods described earlier in this subsection while the results for BM were obtained from the simulations performed for Section 1.4.1. It is clear that BM results in superior estimation. In particular, note that using the setting BM1 results in better estimates of $E(\mu|y)$ and $E(\lambda|y)$ than using setting GRD4 while using approximately half the average simulation effort (2191 (s.e. = 19.9) versus 5365 (150.5)); see Table 1.1.

Consider GRD4 and BM2. Note that these two settings have comparable average simulation effort. The MSE for $\bar{\mu}_n$ using GRD was 0.000134 (9.2e-6) and for BM we observed an MSE of 0.0000373 (1.8e-6). Now consider $\bar{\lambda}_n$. The MSE based on using GRD was 0.00165 (1.2e-4) and for BM we observed an MSE of 0.000393 (1.8e-5). Certainly, the more variable simulation effort of GRD contributes to this difference but so does the default use of burn-in

Recall that we employed a sequential sampler to draw from the target distribution implying that the Markov chain is stationary and hence burn-in is unnecessary. To understand the effect of using burn-in we calculated the estimates of $E(\mu|y)$ using the entire simulation; that is, we did not discard the first l draws of each of the m parallel chains. This yields an estimated MSE of 0.0000709 (4.8e-6) for GRD4. Thus, the estimates using GRD4 still have an estimated MSE 1.9 times larger than that obtained using BM2. The standard errors of the MSE estimates show that this difference is still significant, indicating BM, in terms of MSE, is still a superior method for estimating $E(\mu|y)$. Similarly, for estimating $E(\lambda|y)$ the MSE using GRD4 without discarding the first half of each chain is 2.1 higher than that of BM2.

Toy examples are useful for illustration, however it is sometimes difficult to know

just how much credence the resulting claims should be given. For this reason, we turn our attention to a setting that is "realistic" in the sense that it is similar to the type of setting encountered in practice. Specifically, we do not know the true values of the posterior expectations and implementing a reasonable MCMC strategy is not easy. Moreover, we do not know the convergence rate of the associated Markov chain.

1.5 Hierarchical Model for Geostatistics

The following example is directly from Flegal et al. (2008) which considers a data set on wheat crop flowering dates in the state of North Dakota (Haran et al., 2008). This data consists of experts' model-based estimates for the dates when wheat crops flower at 365 different locations across the state. Let D be the set of N sites and the estimate for the flowering date at site s be $Z(s)$ for $s \in D$. Let $X(s)$ be the latitude for $s \in D$. The flowering dates are generally expected to be later in the year as $X(s)$ increases so we assume that the expected value for $Z(s)$ increases linearly with $X(s)$. The flowering dates are also assumed to be spatially dependent, suggesting the following hierarchical model:

$$Z(s) \mid \beta, \xi(s) = X(s)\beta + \xi(s) \quad \text{for } s \in D\,,$$

$$\xi \mid \tau^2, \sigma^2, \phi \sim N(0, \Sigma(\tau^2, \sigma^2, \phi)),$$

where $\xi = (\xi(s_1), \ldots, \xi(s_N))^T$ with $\Sigma(\tau^2, \sigma^2, \phi) = \tau^2 I + \sigma^2 H(\phi)$ and $\{H(\phi)\}_{ij} = \exp((-\|s_i - s_j\|)/\phi)$, the exponential correlation function. We complete the specification of the model with priors on τ^2, σ^2, ϕ, and β,

$$\tau^2 \sim \text{IG}(2, 30), \quad \sigma^2 \sim \text{IG}(0.1, 30),$$
$$\phi \sim \text{Log-Unif}(0.6, 6), \quad \pi(\beta) \propto 1\,.$$

Setting $Z = (Z(s_1), \ldots, Z(s_N))$, inference is based on the posterior distribution $\pi(\tau^2, \sigma^2, \phi, \beta \mid Z)$. Note that MCMC may be required since the integrals required for inference are analytically intractable. Also, samples from this posterior distribution can then be used for prediction at any location $s \in D$.

Consider estimating the posterior expectation of τ^2, σ^2, ϕ, and β. Unlike the toy example considered earlier these expectations are not analytically available. Sampling from the posterior is accomplished via a Metropolis-Hastings sampler with a joint update for the τ^2, ϕ, β via a three-dimensional independent Normal proposal centered at the current state with a variance of 0.3 for each component and a univariate Gibbs update for σ^2.

To obtain a high quality approximation to the desired posterior expectations we used a single long run of 500,000 iterations of the sampler and obtained 23.23 (.0426), 25.82 (.0200), 2.17 (.0069), and 4.09 (4.3e-5) as estimates of the posterior expectations of τ^2, σ^2, ϕ, and β, respectively. These are assumed to be the truth. We also recorded the 10th, 30th, 70th and 90th percentiles of this long run for each parameter.

Our goal is to compare the finite-sample properties of GRD and BM in terms of quality of estimation and overall simulation effort. Consider implementation of GRD. We will produce 100 independent replications using the following procedure. For each replication we used $m = 4$ parallel chains from four different starting values corresponding to the 10th, 30th, 70th and 90th percentiles recorded above. A minimum total simulation effort of 1000 (250 per chain) was required. Also, no burn-in was employed. This is consistent with our finding in the toy example that estimation improved without using burn-in. Each replication continued until $\hat{R}_{0.975} \leq 1.1$ for all of the parameter estimates. Estimates of the posterior expectations were obtained by averaging draws across all 4 parallel chains.

Now consider the implementation of BM. For the purpose of easy comparison with GRD, we ran a total of 400 independent replications of our MCMC sampler, where the

Method	GRD		BM	
Parameter	MSE	S.E.	MSE	S.E.
$E(\tau^2\|z)$	0.201	0.0408	0.0269	0.00185
$E(\sigma^2\|z)$	0.0699	0.0179	0.00561	0.00039
$E(\phi\|z)$	0.00429	0.00061	0.000875	5.76e-05
$E(\beta\|z)$	1.7e-07	3.09e-08	3.04e-08	1.89e-09

Table 1.3: Summary of estimated mean-squared error obtained using BM and GRD for the model of Section 1.5. Standard errors (S.E.) shown for each estimate.

10th, 30th, 70th and 90th percentiles of the parameter samples from the long run were used as starting values for 100 replications each. Each replication was simulated for a minimum of 1000 iterations so $p(n) = \epsilon I(n < 1000)$. Thus the minimum simulation effort is the same as that for GRD. Using (1.6), a single replication (chain) was terminated when each of the half-widths of a 95% interval was smaller than 0.5, 0.5, 0.05 and 0.05 for the estimates of the posterior expectations of τ^2, σ^2, ϕ, and β, respectively. These thresholds correspond to reasonable desired accuracies for the parameters. If the half-width was not less than the cutoff, then 10 iterations were added to the chain before checking again.

The results from our simulation study are summarized in Table 1.3. Clearly, the MSE for estimates using GRD are significantly higher than the MSE for estimates obtained using BM. However, BM required a greater average simulation effort 31,568.9 (177.73) than did GRD 8,082 (525.7). To study whether the BM stopping rule delivered confidence intervals at the desired 95% levels, we also estimated the coverage probabilities for the intervals for the posterior expectations of τ^2, σ^2, ϕ, and β, which were 0.948 (0.0112), 0.945 (0.0114), 0.912 (0.0141), and 0.953 (0.0106) respectively. The coverage for all parameters is fairly close to the desired 95%.

Finally, we note that this simulation study was conducted on a Linux cluster using `R` (Ihaka and Gentleman, 1996), an MCMC package for spatial modeling, `spBayes` (Finley et al., 2007), and the parallel random number generator package `rlecuyer`

(L'Ecuyer et al., 2002).

1.6 Discussion

The point of this chapter is that those examining the results of MCMC computations are much better off when reliable techniques are used to estimate MCSEs and then *the MCSEs are reported*. An MCSE provides two desirable properties: (1) It gives useful information about the quality of the subsequent estimation and inference; and (2) it provides a theoretically justified, yet easily implemented, approach for determining appropriate stopping rules for their MCMC runs. On the other hand, a claim that a test indicated the sampler "converged" is simply nowhere near enough information to objectively judge the quality of the subsequent estimation and inference. Discarding a set of initial draws does not necessarily improve the situation.

A key requirement for reporting valid Monte Carlo standard errors is that the sampler mixes well. Finding a good sampler is likely to be the most challenging part of the recipe we describe. We have given no guidance on this other than one should look within the class of geometrically ergodic Markov chains if at all possible. This is an important matter in any MCMC setting; that is, a Markov chain that converges quickly is key to obtaining effective simulation results in finite time. Thus there is still a great deal of room for creativity and research in improving samplers but there are already many useful methods that can be implemented for difficult problems. For example, one of our favorite techniques is simulated tempering (Geyer and Thompson, 1995; Marinari and Parisi, 1992) but many others are possible.

1.7 Proofs and Calculations

1.7.1 Toy Example

This section contains calculations to verify the necessary conditions for a Markov chain CLT and fixed-width methods (see Chapter 2).

Normal Moments

In this section, we calculate the first eight moments for a normal distribution for use in Sections 1.7.1 and 1.7.1. Let $X \sim N(\mu, \sigma^2)$, then it's easy to calculate directly or by the moment generating function the following moments:

$$\begin{aligned}
\mathrm{E}X^1 &= \mu \,, \\
\mathrm{E}X^2 &= \mu^2 + \sigma^2 \,, \\
\mathrm{E}X^3 &= \mu^3 + 3\mu\sigma^2 \,, \\
\mathrm{E}X^4 &= \mu^4 + 6\mu^2\sigma^2 + 3\sigma^4 \,.
\end{aligned}$$

Then we can appeal to *Stein's Lemma*.

Lemma 1. *(Stein's Lemma) Let $X \sim N(\mu, \sigma^2)$, and let g be a differentiable function satisfying $E|g'(x)| < \infty$. Then*

$$E\left[g(X)(X-\mu)\right] = \sigma^2 E g'(x) \,.$$

Applying Stein's Lemma, we can calculate

$$\begin{aligned}
\mathrm{E}X^5 &= \mathrm{E}X^4(X-\mu+\mu) \\
&= \mathrm{E}X^4(X-\mu) + \mu\mathrm{E}X^4 \\
&= 4\sigma^2\mathrm{E}X^3 + \mu\mathrm{E}X^4 \\
&= 4\sigma^2\left(\mu^3+3\mu\sigma^2\right) + \mu\left(\mu^4+6\mu^2\sigma^2+3\sigma^4\right) \\
&= \mu^5+10\mu^3\sigma^2+15\mu\sigma^4 \ .
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathrm{E}X^6 &= \mathrm{E}X^5(X-\mu+\mu) \\
&= \mathrm{E}X^5(X-\mu) + \mu\mathrm{E}X^5 \\
&= 5\sigma^2\mathrm{E}X^4 + \mu\mathrm{E}X^5 \\
&= 5\sigma^2\left(\mu^4+6\mu^2\sigma^2+3\sigma^4\right) + \mu\left(\mu^5+10\mu^3\sigma^2+15\mu\sigma^4\right) \\
&= \mu^6+15\mu^4\sigma^2+45\mu^2\sigma^4+15\sigma^6 \ .
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathrm{E}X^7 &= \mathrm{E}X^6(X-\mu+\mu) \\
&= \mathrm{E}X^6(X-\mu) + \mu\mathrm{E}X^6 \\
&= 6\sigma^2\mathrm{E}X^5 + \mu\mathrm{E}X^6 \\
&= 6\sigma^2\left(\mu^5+10\mu^3\sigma^2+15\mu\sigma^4\right) + \mu\left(\mu^6+15\mu^4\sigma^2+45\mu^2\sigma^4+15\sigma^6\right) \\
&= \mu^7+21\mu^5\sigma^2+105\mu^3\sigma^4+105\mu\sigma^6 \ .
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathrm{E}X^8 &= \mathrm{E}X^7(X-\mu+\mu) \\
&= \mathrm{E}X^7(X-\mu) + \mu\mathrm{E}X^7 \\
&= 7\sigma^2\mathrm{E}X^6 + \mu\mathrm{E}X^7 \\
&= 7\sigma^2\left(\mu^6 + 15\mu^4\sigma^2 + 45\mu^2\sigma^4 + 15\sigma^6\right) + \mu\left(\mu^7 + 21\mu^5\sigma^2 + 105\mu^3\sigma^4 + 105\mu\sigma^6\right) \\
&= \mu^8 + 28\mu^6\sigma^2 + 210\mu^4\sigma^4 + 420\mu^2\sigma^6 + 105\sigma^8 \ .
\end{aligned}$$

Sequential Sampling

For the toy example, the posterior density is proportional to (1.5). That is

$$\pi(\mu,\lambda|y) = \frac{\lambda^{-\frac{m+1}{2}}\exp\left\{-\frac{1}{2\lambda}\sum_{j=1}^m(y_j-\mu)^2\right\}}{c}, \tag{1.8}$$

where

$$\begin{aligned}
c &= \int_{\mathbb{R}^+}\int_{\mathbb{R}} \lambda^{-\frac{m+1}{2}} e^{-\frac{1}{2\lambda}\sum_{j=1}^m(y_j-\mu)^2} d\mu d\lambda \\
&= \int_{\mathbb{R}^+} \lambda^{-\frac{m+1}{2}} \sqrt{\frac{2\pi\lambda}{m}} e^{-\frac{1}{2\lambda}\sum_{j=1}^m(y_j-\bar{y}_m)^2} \left[\int_{\mathbb{R}} \sqrt{\frac{m}{2\pi\lambda}} e^{-\frac{m}{2\lambda}(\bar{y}_m-\mu)^2} d\mu\right] d\lambda \\
&= \int_{\mathbb{R}^+} \lambda^{-\frac{m+1}{2}} \sqrt{\frac{2\pi\lambda}{m}} e^{-\frac{1}{2\lambda}\sum_{j=1}^m(y_j-\bar{y}_m)^2} d\lambda \\
&= \sqrt{\frac{2\pi}{m}}\frac{\Gamma(\frac{m-2}{2})}{(s^2/2)^{\frac{m-2}{2}}} \left[\int_{\mathbb{R}^+} \frac{(s^2/2)^{\frac{m-2}{2}}}{\Gamma(\frac{m-2}{2})} \lambda^{-(\frac{m-2}{2}+1)} e^{-\frac{s^2}{2\lambda}} d\lambda\right] \\
&= \sqrt{\frac{2\pi}{m}}\frac{\Gamma(\frac{m-2}{2})}{(s^2/2)^{\frac{m-2}{2}}}.
\end{aligned}$$

Next, we can plug this into (1.8) and group terms,

$$\begin{aligned}\pi(\mu,\lambda|\mathbf{y}) &= \left[\sqrt{\frac{m}{2\pi}}\frac{(s^2/2)^{\frac{m-2}{2}}}{\Gamma(\frac{m-2}{2})}\right]\lambda^{-\frac{m+1}{2}}e^{-\frac{1}{2\lambda}\sum_{j=1}^{m}(y_j-\mu)^2}\\ &= \left[\frac{(s^2/2)^{\frac{m-2}{2}}}{\Gamma(\frac{m-2}{2})}\lambda^{-(\frac{m-2}{2}+1)}e^{-\frac{s^2}{2\lambda}}\right]\left[\sqrt{\frac{m}{2\pi\lambda}}e^{-\frac{m}{2\lambda}(\bar{y}_m-\mu)^2}d\mu\right]\\ &= g_1(\lambda)g_2(\mu|\lambda),\end{aligned}$$

where $g_1(\lambda)$ is an $IG(\frac{m-2}{2}, s^2/2)$ and $g_2(\mu|\lambda)$ is a $N(\bar{y}_m, \lambda/m)$. Using this representation, we can sequentially sample the exact distribution. First, we can take a draw from $g_1(\lambda)$, and then conditionally, draw from $g_2(\mu|\lambda)$.

Calculating $\mathbf{E}(\mu|y)$ and $\mathbf{E}(\lambda|y)$

Using similar techniques as Section 1.7.1, we can calculate the mean of $\mathrm{E}_\pi(\mu|y)$;

$$\begin{aligned}\mathrm{E}_\pi(\mu|y) &= \int_{\mathbb{R}^+} g_1(\lambda)\left[\int_{\mathbb{R}}\mu g_2(\mu|\lambda)d\mu\right]d\lambda\\ &= \int_{\mathbb{R}^+} g_1(\lambda)\bar{y}_m d\lambda\\ &= \bar{y}_m\ .\end{aligned}$$

Similarly, we can calculate the mean of $\mathrm{E}_\pi(\lambda|y)$;

$$\begin{aligned}\mathrm{E}_\pi(\lambda|y) &= \int_{\mathbb{R}^+}\lambda g_1(\lambda)\left[\int_{\mathbb{R}} g_2(\mu|\lambda)d\mu\right]d\lambda\\ &= \int_{\mathbb{R}^+}\lambda g_1(\lambda)d\lambda\\ &= \frac{\frac{s^2}{2}}{\frac{m-2}{2}-1} = \frac{s^2}{m-4}\ ,\end{aligned}$$

if $m > 4$.

Calculating M

Using similar techniques as Section 1.7.1, we can verify the M is $\bar{y}_m$;

$$\begin{aligned}
\Pr_\pi(\mu < \bar{y}_m) &= \int_{\mathbb{R}^+} g_1(\lambda) \left[\int_{-\infty}^{\bar{y}_m} g_2(\mu|\lambda) d\mu \right] d\lambda \\
&= \int_{\mathbb{R}^+} g_1(\lambda) \frac{1}{2} d\lambda \\
&= \frac{1}{2} ,
\end{aligned}$$

similarly, $\Pr_\pi(\mu > \bar{y}_m) = 1/2$.

Verifying the Central Limit Theorem

Consider estimation of $\mathrm{E}_\pi(\mu|y)$ with $\bar{\mu}_n$. To apply the CLT, Theorem 2, we need to show that $\mathrm{E}_\pi(\mu^{2+\epsilon}|y) < \infty$. Here, we will show that $\mathrm{E}_\pi(\mu^3|y) < \infty$.

$$\begin{aligned}
\mathrm{E}_\pi(\mu^3|y) &= \int_{\mathbb{R}^+} g_1(\lambda) \left[\int_{\mathbb{R}} \mu^3 g_2(\mu|\lambda) d\mu \right] d\lambda \\
&= \int_{\mathbb{R}^+} g_1(\lambda) \left[\bar{y}_m^3 + 3\bar{y}_m \frac{\lambda}{m} \right] d\lambda \\
&= \bar{y}_m^3 + \frac{3\bar{y}_m}{m} \int_{\mathbb{R}^+} \lambda g_1(\lambda) d\lambda \\
&= \bar{y}_m^3 + \frac{3\bar{y}_m}{m} \frac{s^2}{m-4},
\end{aligned}$$

which is clearly finite for $m > 4$ and any other parameter values we are considering. This, combined with Jones and Hobert (2001) showing this sampler is geometrically ergodic when $m > 4$, implies we can apply Theorem 2.

Next, we can consider the case of trying to estimate $\mathrm{E}_\pi(\lambda|y)$ with $\bar{\lambda}_n$. Here, we

will look at $\mathrm{E}_\pi(\lambda^3|y)$,

$$\begin{aligned}
\mathrm{E}_\pi(\lambda^3|y) &= \int_{\mathbb{R}^+} \lambda^3 g_1(\lambda) \left[\int_{\mathbb{R}} g_2(\mu|\lambda) d\mu \right] d\lambda \\
&= \int_{\mathbb{R}^+} \lambda^3 g_1(\lambda) d\lambda \\
&= \int_{\mathbb{R}^+} \lambda^3 \left[\frac{(s^2/2)^{\frac{m-2}{2}}}{\Gamma(\frac{m-2}{2})} \lambda^{-(\frac{m-2}{2}+1)} e^{-\frac{s^2}{2\lambda}} \right] d\lambda \\
&= \frac{(s^2/2)^{\frac{m-2}{2}}}{\Gamma(\frac{m-2}{2})} \int_{\mathbb{R}^+} \lambda^{-(\frac{m-8}{2}+1)} e^{-\frac{s^2}{2\lambda}} d\lambda \\
&= \frac{\Gamma(\frac{m-8}{2})}{\Gamma(\frac{m-2}{2})} \left(\frac{s^2}{2} \right)^3
\end{aligned}$$

if $m > 8$. Again, we can appeal to Theorem 2 to estimate $\mathrm{E}_\pi(\lambda|y)$.

Verifying Proposition 2 for BM

Consider estimation of $\mathrm{E}_\pi(\mu|y)$ with $\bar{\mu}_n$ and calculating the MCSE with BM. First, we will show that $\mathrm{E}_\pi(\mu^8|y) < \infty$.

$$\begin{aligned}
\mathrm{E}_\pi(|\mu^8| \mid y) &= \mathrm{E}_\pi(\mu^8|y) \\
&= \int_{\mathbb{R}^+} g_1(\lambda) \left[\int_{\mathbb{R}} \mu^8 g_2(\mu|\lambda) d\mu \right] d\lambda \\
&= \int_{\mathbb{R}^+} g_1(\lambda) \left[\bar{y}_n^8 + 28\bar{y}_n^6 \frac{\lambda}{m} + 210\bar{y}_n^4 \frac{\lambda^2}{m^2} + 420\bar{y}_n^2 \frac{\lambda^3}{m^3} + 105\frac{\lambda^4}{m^4} \right] d\lambda \\
&\leq \text{Constant} + \text{Constant} \int_{\mathbb{R}^+} g_1(\lambda)\lambda^4 d\lambda < \infty
\end{aligned}$$

if $m > 10$. With this moment condition, we can apply Proposition 2 with a batch size of $b_n = \lfloor n^\nu \rfloor$ for any ν such that $1/4 < \nu < 1$.

Similarly, consider estimating $\mathrm{E}_\pi(\lambda|y)$ with $\bar{\lambda}_n$. If $m = 11$ as in our example, then

$$\begin{aligned}\mathrm{E}_\pi(|\lambda^{2+\delta}| \quad | \quad y) &= \mathrm{E}_\pi(\lambda^{2+\delta}|y) \\ &= \int_{\mathbb{R}^+} \lambda^{2+\delta} g_1(\lambda) d\lambda \\ &< \infty\end{aligned}$$

for $\delta < 5/2$. Then with this moment condition, we can apply Proposition 2 with a batch size of $b_n = \lfloor n^\nu \rfloor$ for any ν such that $4/9 < \nu < 1$.

1.7.2 More on the Gelman-Rubin Diagnostic

While Brooks and Gelman (1998) propose the use of $\hat{R}$ as a tool to evaluate the convergence of a Markov chain, they give no practical manner in which to calculate $\hat{\mathrm{var}}(\hat{V})$. Examination of the `coda` code reveals $\hat{\mathrm{var}}(\hat{V})$ is calculated in the following manner:

$$\begin{aligned}\hat{\mathrm{var}}(\hat{V}) &= \hat{\mathrm{var}}\left(\frac{l-1}{l}W + \frac{(m+1)B}{ml}\right) \\ &= \frac{(l-1)^2}{l^2}\hat{\sigma}_W^2 + \frac{(m+1)^2}{m^2 l^2}\hat{\sigma}_B^2 + 2\frac{l-1}{l}\frac{(m+1)}{ml}\hat{\mathrm{cov}}(B, W)\,,\end{aligned}$$

where

$$\hat{\sigma}_B^2 = \frac{2 * B^2}{m-1}\,,$$

and

$$\begin{aligned}
\hat{\text{cov}}(B,W) &= \hat{\text{cov}}\left(\frac{l}{m-1}\sum_{j=1}^{m}(\bar{Y}_{\cdot j}-\bar{Y}_{\cdot\cdot})^2\ ,\ \frac{1}{m}\sum_{j=1}^{m}s_j^2\right)\\
&= \frac{l}{(m-1)m}\hat{\text{cov}}\left(\sum_{j=1}^{m}\left[\bar{Y}_{\cdot j}^2+\bar{Y}_{\cdot\cdot}^2-2\bar{Y}_{\cdot j}\bar{Y}_{\cdot\cdot}\right]\ ,\ \sum_{j=1}^{m}s_j^2\right)\\
&= \frac{l}{(m-1)m}\left[\hat{\text{cov}}\left(\sum_{j=1}^{m}\bar{Y}_{\cdot j}^2\ ,\ \sum_{j=1}^{m}s_j^2\right)+\hat{\text{cov}}\left(m\bar{Y}_{\cdot\cdot}^2\ ,\ \sum_{j=1}^{m}s_j^2\right)\right.\\
&\left.-2\hat{\text{cov}}\left(\sum_{j=1}^{m}\bar{Y}_{\cdot j}\bar{Y}_{\cdot\cdot}\ ,\ \sum_{j=1}^{m}s_j^2\right)\right]\\
&= \frac{l}{(m-1)m}\left[\hat{\text{cov}}\left(\sum_{j=1}^{m}\bar{Y}_{\cdot j}^2\ ,\ \sum_{j=1}^{m}s_j^2\right)+0-2\bar{Y}_{\cdot\cdot}\hat{\text{cov}}\left(\sum_{j=1}^{m}\bar{Y}_{\cdot j}\ ,\ \sum_{j=1}^{m}s_j^2\right)\right]\ .
\end{aligned}$$

Then define $\bar{Y}_{\cdot} = (\bar{Y}_{\cdot 1},\ldots,\bar{Y}_{\cdot m})^T$, $\bar{Y}_{\cdot}^2 = (\bar{Y}_{\cdot 1}^2,\ldots,\bar{Y}_{\cdot m}^2)^T$ and $s^2 = (s_1^2,\ldots,s_m^2)^T$. Then `coda` estimates the following quantities as

$$\hat{\text{cov}}\left(\sum_{j=1}^{m}\bar{Y}_{\cdot j}^2\ ,\ \sum_{j=1}^{m}s_j^2\right) = m^2\text{cor}\left(\bar{Y}_{\cdot}^2\ ,\ s^2\right)$$

and

$$\hat{\text{cov}}\left(\sum_{j=1}^{m}\bar{Y}_{\cdot j}\ ,\ \sum_{j=1}^{m}s_j^2\right) = m^2\text{cor}\left(\bar{Y}_{\cdot}\ ,\ s^2\right)\ .$$

Of course, `coda` uses $m-1$ instead of m^2, but I think this is just an error in the code. With my correction, this yields the following estimate

$$\begin{aligned}
\hat{\text{var}}(\hat{V}) = &\frac{(l-1)^2}{l^2}\hat{\sigma}_W^2+\frac{(m+1)^2}{m^2l^2}\hat{\sigma}_B^2\\
&+2\frac{(l-1)(m+1)}{l(m-1)}\left[\text{cor}\left(\bar{Y}_{\cdot}^2\ ,\ s^2\right)-2\bar{Y}_{\cdot\cdot}\text{cor}\left(\bar{Y}_{\cdot}\ ,\ s^2\right)\right]\ .
\end{aligned}$$

An obvious question to ask would be "Why calculate $\hat{\text{var}}(\hat{V})$ in this manner?" Particularly, the above calculation assumes that $\bar{Y}_{\cdot\cdot}$ is a fixed quantity (which it is

clearly not). However, we are not interested in improving this ad-hoc diagnostic considering the poor results from implementation of the GRD in our examples.

Chapter 2

Markov Chain Monte Carlo

MCMC has become a standard technique in the toolbox of applied statisticians. Simply put, MCMC is a method for using a computer to generate data in order to estimate fixed, unknown quantities of a target distribution. That is, a common use of MCMC is to produce a point estimate of some characteristic of a given target distribution.

As stated above, consider the specific case where we are interested in finding $E_\pi g := \int_{\mathsf{X}} g(x)\,\pi(dx)$ where π is a probability distribution with support X and g is a real-valued, π-integrable function on X. (Chapter 4 considers a more general case.) In modern applications we often have to resort to MCMC methods to approximate $E_\pi g$. To this end, this chapter discusses the relevant Markov chain theory with particular attention paid to the conditions and definitions needed to establish a Markov chain central limit theorem. In addition, these conditions are verified in the context of two examples.

2.1 Markov Chains

This section provides a brief discussion of Markov chain theory; for more details see Jones and Hobert (2001); Meyn and Tweedie (1993); Tierney (1994).

Let $X = \{X_1, X_2, X_3, \dots\}$ be a discrete-time Markov chain on a general state space X and let $\mathcal{B}$ denote the associated Borel σ-algebra. Then let $P(x, dy)$ denote the associated Markov transition kernel; that is, for $x \in \mathsf{X}$ and $A \in \mathcal{B}$,

$$P(x, A) = \Pr(X_{i+1} \in A | X_i = x).$$

For $n \in \mathbb{N} := \{1, 2, 3, \dots\}$, let $P^n(x, dy)$ denote the n-step Markov transition kernel; that is, for $x \in \mathsf{X}$, $A \in \mathcal{B}$, and $i \in \mathbb{N}$,

$$P^n(x, A) = \Pr(X_{n+i} \in A | X_i = x).$$

For ease of exposition, we will often assume that the probability measure $P(x, \cdot)$ has a conditional density, $k(\cdot|x)$, with respect to Lebesgue measure so that,

$$P(x, A) = \int_A k(u|x) du \ .$$

We will call k a Markov transition density. Further, if there exists a density π such that

$$\pi(x) = \int_{\mathsf{X}} k(x|y)\pi(y) dy \ , \tag{2.1}$$

then π is called the stationary or invariant density for the Markov chain X. Consider the idea of stationarity in (2.1). If we begin by drawing y from π and apply the Markov transition kernel, $P(x, dy)$, resulting in the transition $x \to y$. Then the joint density of (y, x) has the same formula as seen in (2.1), implying the marginal density of x is also π. Consequently, if we can draw $X_1 \sim \pi$, then X is a sequence of dependent observations from π, or the chain is stationary.

Practically, it is usually impossible to draw from π (hence the use of MCMC in the first place), therefore we might consider conditions under which the chain "converges"

to π. Define

$$\|P^n(x,\cdot) - \pi(\cdot)\| := \sup_{A\in\mathcal{B}} |P^n(x,A) - \pi(A)| \ ,$$

the **total variation** distance between the probability measures $P^n(x,\cdot)$ and $\pi(\cdot)$. Under regularity assumptions, something can be said about the total variation distance. First, the following are some important definitions.

Definition 1. A Markov chain transition kernel P is **π-irreducible** if for any $x \in \mathsf{X}$ and for any set A with $\pi(A) > 0$, there exists an n such that $P^n(x,A) > 0$.

In other words, starting from any point in the state space, there exists an n such that there is positive probability we can reach any set having positive π-probability. If X is a Markov chain with a π-irreducible transition kernel, we will say X is a π-irreducible Markov chain.

Definition 2. A π-irreducible Markov transition kernel P is **periodic** if there exists an integer $d \geq 2$ and a collection of disjoint sets $A_1, \ldots, A_d \in \mathcal{B}$ such that for each $x \in A_j$, $P(x, A_{j+1}) = 1$ for $j = 1, \ldots, d-1$, and for each $x \in A_d$, $P(x, A_1) = 1$. Otherwise, P is said to be **aperiodic**.

That is, P is periodic if we can partition the state space in such as way as to introduce cyclic behavior. If no such partition exists, then P is aperiodic. Similarly, if X is a Markov chain with periodic (aperiodic) P, then we will say X is periodic (aperiodic).

Definition 3. If X is a π-irreducible Markov chain where π is the stationary distribution, then X is **recurrent** if for every set A with $\pi(A) > 0$,

$$\Pr(X_n \in A \text{ i.o.} | X_1 = x) > 0 \text{ for all } x \ ,$$
$$\Pr(X_n \in A \text{ i.o.} | X_1 = x) = 1 \text{ for } \pi\text{-almost all } x \ .$$

The chain is **Harris recurrent** if $\Pr(X_n \in A \text{ i.o.} | X_1 = x) = 1$ for all x.

If π is a probability distribution, then X is **positive recurrent** (or **positive Harris recurrent**).

When a chain is π-irreducible, aperiodic, and positive Harris recurrent, we will call it **Harris ergodic**.

Proposition 1. *Suppose P is π-irreducible and π is an invariant distribution of P. Then P is positive recurrent and π is the unique invariant distribution of P. If P is also aperiodic, then, for π-almost all x,*

$$\|P^n(x,\cdot) - \pi(\cdot)\| \to 0 \text{ as } n \to \infty .$$

If P is positive Harris recurrent, then the convergence occurs for all x.

Athreya et al. (1996) provide a proof of Proposition 1. The proposition shows the limit of the total variation norm is zero for Harris ergodic chains, but says nothing about the rate of convergence. We are particularly interested in bounding the rate of convergence of the total variation norm because of its connection to Markov chain CLTs and consistent estimators of the associated asymptotic variance.

The previous chapter defined one rate of convergence which we will use in the following sections. The formal definition is as follows.

Definition 4. Let X be a Harris ergodic Markov chain with invariant distribution π. The chain is **geometrically ergodic** if there exists a constant $0 < t < 1$ and a function $M : \mathsf{X} \mapsto \mathbb{R}^+$ such that for any $x \in \mathsf{X}$,

$$\|P^n(x,\cdot) - \pi(\cdot)\| \leq M(x)t^n \tag{2.2}$$

for $n \in \mathbb{N}$. If there exists a bounded M satisfying (2.2), then X is **uniformly ergodic** (and if X has a finite number of elements, then M is clearly bounded).

2.1.1 Establishing Geometric Ergodicity

In this section, two approaches for establishing geometric ergodicity are presented. Assume X is a Harris ergodic Markov chain with invariant distribution π. We will say a **drift condition** holds if for some function $W : \mathsf{X} \mapsto [1, \infty)$, some $0 < \gamma < 1$ and some $L < \infty$,

$$E\left[W(X_{i+1})|X_i = x\right] \leq \gamma W(x) + I_{(x \in S)}L \quad \text{for all } x \in \mathsf{X} \tag{2.3}$$

where $S = \{x \in \mathsf{X} : W(x) \leq d\}$ and

$$d = \frac{L}{1-\gamma} - 1.$$

Next, a **minorization condition** holds if for some probability measure Q on $\mathcal{B}$, some set C with $\pi(C) > 0$, and some $\epsilon > 0$

$$P(x, A) \geq \epsilon Q(A) \quad \text{for all } x \in C \quad \text{and for all } A \in \mathcal{B}\,. \tag{2.4}$$

If an associated minorization condition holds with $C = S$, then it is well known that the associated minorization condition with (2.3) imply X is geometrically ergodic (Meyn and Tweedie, 1993).

AR(1) Example

Consider the first order autoregressive process as follows

$$X_i = \rho X_{i-1} + \epsilon_i \quad \text{for } i = 1, 2, \ldots\,,$$

where ϵ_i is an i.i.d. N(0,τ^2) for $i = 1, 2, \ldots$. It is easy to show the distribution of $X_{i+1}|X_i = x$ is given by N($\rho x, \tau^2$) resulting in a normal conditional density, say $k(\cdot|x)$.

Let $W(x) = x^2 + 1$, then

$$\begin{aligned} E\left[W(X_{i+1})|X_i = x\right] &= \rho^2 x^2 + \tau^2 + 1 \\ &= \rho^2\left(x^2+1\right) + \tau^2 + \left(1-\rho^2\right) \\ &= \frac{1+\rho^2}{2}W(x) - \frac{1-\rho^2}{2}W(x) + \tau^2 + \left(1-\rho^2\right) \\ &= \gamma W(x) - (1-\gamma)\,W(x) + L\ , \end{aligned}$$

where $\gamma := (1+\rho^2)/2$ and $L := [\tau^2 + (1-\rho^2)]$. Notice, for all $x \in \mathbb{R}$

$$E\left[W(X_{i+1})|X_i = x\right] \leq \gamma W(x) + L \tag{2.5}$$

and if $|\rho| < 1$ and $W(x) > L/(1-\gamma)$

$$E\left[W(X_{i+1})|X_i = x\right] \leq \gamma W(x)\ . \tag{2.6}$$

Combining (2.5) and (2.6) yields

$$E\left[W(X_{i+1})|X_i = x\right] \leq \gamma W(x) + I_{(x\in S)}L \quad \text{for all } x \in \mathsf{X}$$

where $S = \{x \in \mathsf{X} : W(x) \leq d_{RT}\}$ and $d_{RT} = L/(1-\gamma)$. Further, since

$$\frac{L}{1-\gamma} > \frac{L}{1-\gamma} - 1$$

this constitutes a drift condition of the form (2.3).

Next, we will show the associated minorization condition when $W(x) \leq d^2$. Let $\mathrm{N}(\mu, \tau^2; x)$ denote the value of the $\mathrm{N}(\mu, \tau^2)$ density at the point $x \in \mathbb{R}$. If $\tau^2 > 0$ and

$d \in \mathbb{R}$ are fixed, then define $h(x)$

$$h(x) := \inf_{-d \leq \mu \leq d} \mathrm{N}\left(\mu, \tau^2; x\right) = \begin{cases} \mathrm{N}\left(d, \tau^2; x\right) & \text{if } x < 0 \,, \\ \mathrm{N}\left(-d, \tau^2; x\right) & \text{if } x \geq 0 \,. \end{cases}$$

Then for all $x \in [-d, d]$

$$k(x|x') \geq \epsilon q(x)$$

where $\epsilon = \int_{\mathbb{R}} h(x)dx$ and

$$q(x) = \frac{h(x)}{\int_{\mathbb{R}} h(x)dx} \,.$$

Thus, the associated minorization condition from (2.4) holds for all $x \in \mathsf{X}$ with $W(x) \leq d^2$ and for all $A \in \mathcal{B}$.

Finally, for $|\rho| < 1$ the AR(1) model is geometrically ergodic.

Rosenthal-type Drift Condition

The drift condition in (2.3) is sometimes referred to as a Roberts-and-Tweedie-type drift condition (Roberts and Tweedie, 1999, 2001). However, there is an alternative equivalent drift condition sometimes referred to as a Rosenthal-type drift condition (Rosenthal, 1995). Again, we will assume X is a Harris ergodic Markov chain with invariant distribution π. A Rosenthal-type **drift condition** holds if for some function $V : \mathsf{X} \mapsto \mathbb{R}^+$, some $0 < \lambda < 1$, and some constant $b < \infty$

$$E\left[V(X_{i+1})|X_i = x\right] \leq \lambda V(x) + b \quad \text{for all } x \in \mathsf{X} \,. \tag{2.7}$$

Notice, in (2.7) we are taking the expectation with respect to the Markov transition kernel. The function V is sometimes called an *energy function* from the fact when the drift condition holds, the chain tends to "drift" towards states of lower energy in terms of expectation.

Often, the drift condition in (2.7) can be easier to establish.

Connection Between Drift Functions

Clearly, the drift condition in (2.3) implies the drift condition in (2.7). Jones and Hobert (2004) show that (2.7) implies (2.3) in general.

Lemma 2. *(Jones and Hobert, 2004, Lemma 3.1) Let X be a Harris ergodic Markov chain with invariant distribution π. Suppose there exists $V : \mathsf{X} \mapsto \mathbb{R}^+$, $0 < \lambda < 1$, and $b < \infty$ such that*

$$E\left[V(X_{i+1})|X_i = x\right] \leq \lambda V(x) + b \quad \textit{for all } x \in \mathsf{X}\,. \tag{2.8}$$

Set $W(x) = 1 + V(x)$. Then, for any $a > 0$,

$$E\left[W(X_{i+1})|X_i = x\right] \leq \gamma W(x) + I_{(x\in S)}L \quad \textit{for all } x \in \mathsf{X}\,, \tag{2.9}$$

where $\gamma = (a + \lambda)/(a + 1)$, $L = b + (1 - \lambda)$ and

$$S = \left\{x \in \mathsf{X} : W(x) \leq \frac{(a+1)L}{a(1-\gamma)}\right\}\,.$$

Then since

$$\frac{(a+1)L}{a(1-\gamma)} \geq \frac{L}{(1-\gamma)} - 1\,,$$

(2.9) constitutes a drift condition of the form of (2.3). Hence, if we can establish (2.7) and the associated minorization condition

$$P(x, A) \geq \epsilon Q(A) \quad \text{for all } x \in \mathsf{X} \quad \text{with } V(x) \leq d \quad \text{and for all } A \in \mathcal{B}$$

then the Markov chain is geometrically ergodic.

AR(1) Example

Recall the AR(1) model and let $V(x) = x^2$, then

$$\begin{aligned} E\left[V(X_{i+1})|X_i = x\right] &= \rho^2 x^2 + \tau^2 \\ &= \rho^2 V(x) + \tau^2 \ , \end{aligned}$$

for all $x \in \mathsf{X}$. Suppose $|\rho| < 1$, then the drift condition in (2.7) holds where $\lambda \in [\rho^2, 1)$ and $b \in [\tau^2, \infty)$.

An associated minorization condition can be established as before, and hence the chain is geometrically ergodic if $|\rho| < 1$.

2.2 Markov Chain Monte Carlo

Suppose that $X = \{X_1, X_2, X_3, \dots\}$ is a Harris ergodic Markov chain with state space X and invariant distribution π (for definitions see Section 2.1). We will maintain these assumptions throughout this thesis. Typically, estimating $E_\pi g$ is natural by appealing to the Ergodic Theorem.

Theorem 1. *Let X be a Harris recurrent Markov chain on X with invariant distribution π and $g : \mathsf{X} \to \mathbb{R}$ be a Borel function. If $E_\pi|g| < \infty$ then, as $n \to \infty$*

$$\bar{g}_n := \frac{1}{n}\sum_{i=1}^{n} g(X_i) \to E_\pi g \quad \textit{almost surely,} \tag{2.10}$$

for any initial distribution.

This follows directly from Theorems 17.0.1 and 17.1.6 in Meyn and Tweedie (1993). Applying Theorem 1 to estimate $E_\pi g$ is easily accomplished by using $\bar{g}_n$.

MCMC methods entail constructing a Markov chain X satisfying the regularity conditions described above and then simulating X for a finite number of steps, say

n, and using $\bar{g}_n$ to estimate $E_\pi g$. The popularity of MCMC methods result from the ease with which such an X can be simulated (Chen et al., 2000; Robert and Casella, 1999; Liu, 2001).

An obvious question is when should we stop the simulation? That is, how large should n be? Or, when is $\bar{g}_n$ a good estimate of $E_\pi g$? An obvious method of addressing the quality of the estimate is to calculate the associated Monte Carlo standard error of $\bar{g}_n$. This requires a Central Limit Theorem (CLT), and some stronger regularity conditions. Specifically, we require the chain to be **geometrically ergodic** or **uniformly ergodic** depending on the moment condition. We will also need to remember a kernel P satisfies the *detailed balance equation* with respect to π if

$$\pi(dx)P(x,dy) = \pi(dy)P(y,dx) \quad \text{for all } x, y \in \mathsf{X} \,. \tag{2.11}$$

The following states three different sets of regularity conditions (without proof) for a Markov chain CLT. Other conditions and discussion can be found in Jones (2004) and Roberts and Rosenthal (2004).

Theorem 2. *Let X be a Harris ergodic Markov chain on X with invariant distribution π, and let $g : \mathsf{X} \to \mathbb{R}$ be a Borel function. Assume one of the following conditions:*

1. *(Doukhan et al., 1994) X is geometrically ergodic and $E_\pi\left[g^2(X)(log^+|g(X)|)\right] < \infty$;*

2. *(Roberts and Rosenthal, 1997) X is geometrically ergodic, satisfies* (2.11)*, and $E_\pi g^2(X) < \infty$; or*

3. *(Ibragimov and Linnik, 1971) X is uniformly ergodic and $E_\pi g^2(X) < \infty$.*

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

$$\sqrt{n}(\bar{g}_n - E_\pi g) \xrightarrow{d} N(0, \sigma_g^2) \,, \tag{2.12}$$

where $\sigma_g^2 := var_\pi\{g(X_1)\} + 2\sum_{i=2}^{\infty} cov_\pi\{g(X_1), g(X_i)\}$*; the subscript* π *means that the expectations are calculated assuming* $X_1 \sim \pi$*.*

Application of Theorem 2 poses two distinct requirements. First, we must verify the necessary conditions. Specifically, we must ensure the appropriate moment conditions. We must also construct a Harris ergodic Markov chain, X, with the appropriate invariant distribution π satisfying geometric (or uniform) ergodicity. Once we have a suitable X, we need to estimate σ_g^2. This requires specialized techniques that take into account the dependence in the observed Markov chain. This will be addressed in Chapter 3.

2.3 Examples

Section 2.1.1 shows the AR(1) model is geometrically ergodic. In future sections, MCMC will be implemented to estimate $E_\pi X$. In this case it is easy to show that $E_\pi X = 0$ and $E_\pi |X|^d < \infty$ for all $d < \infty$, hence Theorem 2 can be applied. This section considers a more realistic example.

2.3.1 Hierarchical Linear Mixed Models

Consider the usual frequentist general linear mixed model

$$Y = X\beta + Zu + \varepsilon \,,$$

where Y is an $n \times 1$ vector of observations, X is a known $n \times p$ matrix, Z is a known $n \times q$ matrix, β is a $p \times 1$ vector of parameters, u is a $q \times 1$ vector of random variables, and ε is an $n \times 1$ vector of residual errors. We assume that X is of full column rank so that $X^T X$ is invertible. A Bayesian version of this model may be expressed as the

following conditionally independent hierarchical model

$$Y|\beta, u, R, D \sim \ \mathrm{N}_n(X\beta + Zu, R^{-1})$$

$$\beta|u, R, D \sim \ \mathrm{N}_p(\beta_0, B^{-1})$$

$$u|D, R \sim \ \mathrm{N}_q(0, D^{-1})$$

with as yet unspecified priors $f(R)$ and $f(D)$. Here β_0 and B^{-1} are assumed to be known. The posterior density of (β, u, R, D) given the data, y, is characterized by

$$\pi(\beta, u, R, D|y) \propto f(y|\beta, u, R, D)f(\beta|u, R, D)f(u|D, R)f(R)f(D)\ . \tag{2.13}$$

We assume that the priors on R and D are such that the resulting posterior (2.13) is proper. Even if proper conjugate priors are chosen, the integrals required for inference through this posterior can not be evaluated in closed form. Thus, exploring the posterior in order to make inferences might require MCMC.

Block Gibbs Sampler

Consider a block Gibbs sampler with components R, D and $\xi = (u^T, \beta^T)^T$. The full conditional densities for R and D are given by

$$\begin{aligned}
\pi(R|\xi, D, y) &= C_R^{-1}(\xi)|R|^{1/2}\exp\{-0.5(y - X\beta - Zu)^T R(y - X\beta - Zu)\}f(R) \\
\pi(D|\xi, R, y) &= C_D^{-1}(\xi)|D|^{1/2}\exp\{-0.5u^T Du\}f(D)
\end{aligned}$$

where $|\cdot|$ is a determinant,

$$C_R(\xi) = \int |R|^{1/2}\exp\{-0.5(y - X\beta - Zu)^T R(y - X\beta - Zu)\}f(R)\,dR\ ,$$

and

$$C_D(\xi) = \int |D|^{1/2} \exp\{-0.5u^T Du\} f(D)\, dD \; .$$

The density $\pi(\xi|R, D, y)$ is $(p+q)$–variate Normal with mean ξ_0 and covariance matrix Σ^{-1} where

$$\Sigma = \begin{pmatrix} Z^T RZ + D & Z^T RX \\ X^T RZ & X^T RX + B \end{pmatrix} \qquad \text{and} \qquad \Sigma\xi_0 = \begin{pmatrix} Z^T Ry \\ X^T Ry + B\beta_0 \end{pmatrix} \; . \tag{2.14}$$

Consider the block Gibbs sampler corresponding to the following updating scheme:

$$(D', R', \xi') \;\to\; (D, R', \xi') \;\to\; (D, R, \xi') \;\to\; (D, R, \xi) \; .$$

Conditional on ξ, D and R are independent and hence the order in which they are updated is irrelevant. That is, we are effectively dealing with a two-variable Gibbs sampler. Suppressing dependence on the data, the transition density is given by

$$k(D, R, \xi|D', R', \xi') = \pi(D|\xi')\, \pi(R|\xi')\, \pi(\xi|R, D) \; .$$

A Special Case

In this section, we identify a specific example of the model in Section 2.3.1 and establish drift and minorization conditions for the block Gibbs sampler. Johnson and Jones (2008) consider a much broader class of models. Suppose that $p = 1$ so that $X = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$ and that $q = n$ with $Z = I_n$. Fix $\beta_0 = 0$ and $B^{-1} = 1$. Assume that $R^{-1} = \lambda_R^{-1} I_n$ and $D^{-1} = \lambda_D^{-1} I_n$ where λ_R^{-1} and λ_D^{-1} are scalar variance components whose reciprocals are assigned the following conjugate priors

$$\lambda_R \sim \text{Gamma}(r_1, r_2) \qquad \text{and} \qquad \lambda_D \sim \text{Gamma}(d_1, d_2) \; .$$

Set $\xi = (u^T, \beta)^T$ and $\lambda = (\lambda_D, \lambda_R)^T$.

Recall that the block Gibbs sampler from the previous section uses the sampling scheme: $(\lambda', \xi') \to (\lambda, \xi') \to (\lambda, \xi)$. The full conditionals for the precision parameters are given by

$$\lambda_R|\xi, y \sim \text{Gamma}\left(r_1 + \frac{n}{2}\,,\, r_2 + \frac{1}{2}(y - X\beta - u)^T(y - X\beta - u)\right)\,,$$
$$\lambda_D|\xi, y \sim \text{Gamma}\left(d_1 + \frac{n}{2}\,,\, d_2 + \frac{1}{2}u^T u\right)\,.$$

Let P be the transition kernel of this Gibbs sampler. Now $\xi|\lambda_R, \lambda_D, y \sim \text{N}_{n+1}(\xi_0, \Sigma^{-1})$ where

$$\Sigma = \begin{pmatrix} (\lambda_R + \lambda_D)I_n & \lambda_R X \\ \lambda_R X^T & 1 + \lambda_R X^T X \end{pmatrix} \quad \text{and} \quad \Sigma\xi_0 = \lambda_R \begin{pmatrix} y \\ X^T y \end{pmatrix}\,. \tag{2.15}$$

We establish the drift condition for this sampler in Section 2.4.3 and the associated minorization condition in Section 2.4.3. From these, we can conclude the chain is geometrically ergodic (see Section 2.1.1 for details).

2.4 Proofs and Calculations

2.4.1 Proof of Lemma 2

This proof is shown in Jones and Hobert (2004). Clearly, (2.8) implies that

$$E\left[W(X_{i+1})|X_i = x\right] \leq \lambda W(x) + b + (1 - \lambda) = \lambda W(x) + L \quad \text{for all } x \in \mathsf{X}\,.$$

Set $\Delta W(x) = E\left[W(X_{i+1})|X_i = x\right] - W(x)$ and $\beta = (1 - \lambda)/(a + 1)$. Then

$$E\left[W(X_{i+1})|X_i = x\right] \leq \left[1 - (a+1)\beta\right] W(x) + L$$

or, equivalently,

$$\Delta W(x) \leq -\beta W(x) - a\beta W(x) + L \quad \text{for all } x \in \mathsf{X} .$$

If $x \notin C$, then

$$W(x) > \frac{(a+1)L}{a(1-\gamma)} > \frac{(a+1)L}{a(1-\lambda)} = \frac{L}{a\beta} .$$

Now write $W(x) = \frac{L}{a\beta} + s(x)$, where $s(x) > 0$. Then

$$\begin{aligned} \Delta W(x) &\leq -\beta W(x) - a\beta \left[\frac{L}{a\beta} + s(x) \right] + L \\ &= -\beta W(x) - a\beta s(x) \\ &\leq -\beta W(x) . \end{aligned}$$

If, on the other hand, $x \in C$, then

$$\begin{aligned} \Delta W(x) &\leq -\beta W(x) - a\beta W(x) + L \\ &\leq -\beta W(x) + L . \end{aligned}$$

Now putting these together gives

$$\begin{aligned} E\left[W(X_{i+1}) | X_i = x\right] &\leq (1-\beta) W(x) + I_{(x \in S)} L \\ &= \gamma W(x) + I_{(x \in S)} L . \end{aligned}$$

2.4.2 Mixing Conditions

This section will provide connections between different mixing conditions for future use. Specifically, the goal of this section is to show that geometrically ergodic Markov chains are exponentially fast alpha-mixing. We begin by deriving the **coupling in-**

equality.

Let $X = \{X_1, X_2, X_3, \dots\}$ and $Y = \{Y_1, Y_2, Y_3, \dots\}$ be two Markov chains with common transition kernel P satisfying the minorization condition in (2.4). Suppose X_1 is an arbitrary starting point in X and $Y_1 \sim \pi$, hence Y is stationary.

Then for each $x \in C$, define the residual kernel as

$$R(x, dy) = \frac{P(x, dy) - \epsilon Q(dy)}{1 - \epsilon} \quad \text{for } \epsilon < 1 \ ,$$

and $R(x, \cdot) := 0$ for $\epsilon = 1$. It is easy to verify R is a transition kernel, and that

$$P(x, dy) = \epsilon Q(dy) + (1 - \epsilon) R(x, dy) \ .$$

Using this representation, we can consider updating X and Y in the following manner.

Let $X_n = x$ and $Y_n = y$.

1. If $x = y$, generate $Z \sim P(x, \cdot)$, and set $X_{n+1} = Y_{n+1} = Z$.

2. If $x \neq y$, $x \in C$, and $y \in C$, generate $\delta \sim \text{Bernoulli}(\epsilon)$ then

   (a) if $\delta = 0$, generate $X_{n+1} \sim R(x, \cdot)$ and $Y_{n+1} \sim R(y, \cdot)$ independently;

   (b) if $\delta = 1$, generate $Z \sim Q(\cdot)$ and set $X_{n+1} = Y_{n+1} = Z$.

3. If $x \neq y$ and $x \notin C$ or $y \notin C$ generate $X_{n+1} \sim P(x, \cdot)$ and $Y_{n+1} \sim P(y, \cdot)$ independently.

It is clear this method retains the original transition kernel P for both chains. It is also clear that if $X_n = Y_n$, then $X_{n+k} = Y_{n+k}$ for all $k \in \mathbb{N}$. In other words, once the chains have **coupled**, all future draws of the two chains remain equal. Define the **coupling time** as $T = \inf\{t : X_t = Y_t\}$, the random time when the two chains couple.

The following is the classical derivation of the coupling inequality

$$\begin{aligned}
|P^n(x_0, A) - \pi(A)| &= |\Pr(X_n \in A) - \Pr(Y_n \in A)| \\
&= |\Pr(X_n \in A, X_n = Y_n) + \Pr(X_n \in A, X_n \neq Y_n) - \\
&\qquad \Pr(Y_n \in A, X_n = Y_n) - \Pr(Y_n \in A, X_n \neq Y_n)| \\
&= |\Pr(X_n \in A, X_n \neq Y_n) - \Pr(Y_n \in A, X_n \neq Y_n)| \\
&\leq \max\{\Pr(X_n \in A, X_n \neq Y_n), \Pr(Y_n \in A, X_n \neq Y_n)\} \\
&\leq \Pr(X_n \neq Y_n) \leq \Pr(T > n) \; .
\end{aligned}$$

Resulting in the **coupling inequality**

$$\|P^n(x, \cdot) - \pi(\cdot)\| \leq \Pr(T > n) \; . \tag{2.16}$$

Under our assumptions (Jones, 2004), the coupling time is almost surely finite and $\Pr(T > n) \to 0$ as $n \to \infty$.

We will use the coupling inequality to show that Harris ergodic Markov chains satisfying (2.4) are alpha mixing. First, let $\mathcal{F}_k^m = \sigma(X_k, \ldots, X_m)$.

Definition 5. The sequence X is said to be **alpha-mixing** (or **strongly mixing**) if $\alpha(n) \to 0$ as $n \to \infty$ where

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

Then let A and B be Borel sets so that by (2.16)

$$|P^n(x, A) - \pi(A)| \leq \Pr_x(T > n)$$

and

$$\begin{aligned}\int_B \Pr_x(T > n)\pi(dx) &\geq \int_B |P^n(x, A) - \pi(A)|\pi(dx) \\ &\geq |\int_B [P^n(x, A) - \pi(A)]\,\pi(dx)| \\ &= |\Pr(X_n \in A \text{ and } X_1 \in B) - \pi(A)\pi(B)|\,.\end{aligned}$$

Then $\alpha(n) \leq \mathrm{E}_\pi\left[\Pr_x(T > n)\right]$ and a dominated convergence argument shows that

$$\mathrm{E}_\pi\left[\Pr_x(T > n)\right] \to 0 \quad \text{as} \quad n \to \infty\,,$$

and hence $\alpha(n) \to 0$ as $n \to \infty$.

Similarly, consider a geometrically ergodic Markov chain satisfying (2.2). Using the same argument, let A and B be Borel sets so

$$|P^n(x, A) - \pi(A)| \leq M(x)t^n$$

and

$$\begin{aligned}\int_B M(x)t^n\pi(dx) &\geq \int_B |P^n(x, A) - \pi(A)|\pi(dx) \\ &\geq |\int_B [P^n(x, A) - \pi(A)]\,\pi(dx)| \\ &= |\Pr(X_n \in A \text{ and } X_1 \in B) - \pi(A)\pi(B)|\,.\end{aligned}$$

Then $\alpha(n) \leq t^n \mathrm{E}_\pi M(x)$ and if $\mathrm{E}_\pi M(x) < \infty$, $\alpha(n) \to 0$ as $n \to \infty$.

Following the argument from Jones (2004), we want to show $\mathrm{E}_\pi M(x) < \infty$. First, notice that from either drift condition (2.7) or (2.3), we can take $M(x) \propto W(x)$ in (2.2). Then we can appeal to Theorem 14.3.7 in Meyn and Tweedie (1993) that shows if (2.3) holds then $\mathrm{E}_\pi W(x) < \infty$. Since a geometrically ergodic chain is equivalent

to (2.3) (Meyn and Tweedie, 1993, Chapter 16) we can conclude that geometrically ergodic Markov chains satisfy (2.2) with $\mathrm{E}_\pi M(x) < \infty$. In other words, geometrically ergodic Markov chains are exponentially fast alpha-mixing.

Remark 1. Results for a Markov chain CLT for quantiles require $\sum_{i=1}^{\infty} \alpha(i) < \infty$ (Chapter 4). If X is a geometrically ergodic Markov chain, then

$$\begin{aligned}\sum_{i=1}^{\infty} \alpha(i) &\leq \sum_{i=1}^{\infty} t^i \left[\mathrm{E}_\pi M(x)\right] \\ &= \left[\mathrm{E}_\pi M(x)\right] \frac{1}{1-t} < \infty \,,\end{aligned}$$

because $0 < t < 1$.

2.4.3 Block Gibbs Sampler

To simulate from this multivariate normal distribution we require the Cholesky decomposition of Σ in (2.15), $\Sigma = LL^T$, where L is

$$L = \begin{pmatrix} l_1 & 0 \\ l_2 & l_3 \end{pmatrix} \,.$$

Solving for L we obtain

$$l_1 = aI_n, \quad l_2 = bX^T, \quad \text{and} \quad l_3 = c$$

where $a = \sqrt{\lambda_R + \lambda_D}\,$, $b = \lambda_R/a$ and $c = \sqrt{1 + (\lambda_R\lambda_D/a^2)X^TX}$. It is easy to see that

$$L^{-1} = \begin{pmatrix} a^{-1}I_n & 0 \\ -b(ac)^{-1}X^T & c^{-1} \end{pmatrix} \,,$$

and hence

$$\Sigma^{-1} = \begin{pmatrix} a^{-2}I_n + (b/ac)^2 XX^T & -b/ac^2 X \\ -b/ac^2 X^T & c^{-2} \end{pmatrix}$$
$$= \begin{pmatrix} \frac{1}{\lambda_R+\lambda_D}\left[I_n + \frac{\lambda_R^2}{\lambda_R+\lambda_D+\lambda_R\lambda_D X^TX} XX^T\right] & \frac{-\lambda_R}{\lambda_R+\lambda_D+\lambda_R\lambda_D X^TX} X \\ \frac{-\lambda_R}{\lambda_R+\lambda_D+\lambda_R\lambda_D X^TX} X^T & \frac{\lambda_R+\lambda_D}{\lambda_R+\lambda_D+\lambda_R\lambda_D X^TX} \end{pmatrix}.$$

Also,

$$\xi_0 = \begin{pmatrix} \frac{\lambda_R}{\lambda_R+\lambda_D} Y - \frac{\lambda_R^2\lambda_D}{(\lambda_R+\lambda_D)(\lambda_R+\lambda_D+\lambda_R\lambda_D X^TX)} XX^TY \\ \frac{\lambda_R\lambda_D}{\lambda_R+\lambda_D+\lambda_R\lambda_D X^TX} X^TY \end{pmatrix}.$$

It is now easy to obtain the following expectations that we will require later.

$$\begin{aligned}
\mathrm{E}(\lambda_D|\xi) &= \frac{2d_1+n}{2d_2+u^Tu} \\
\mathrm{E}(\lambda_D^{-1}|\xi) &= \frac{2d_2+u^Tu}{2d_1+n-2} \\
\mathrm{E}(\lambda_R|\xi) &= \frac{2r_1+n}{2r_2+(Y-X\beta-u)^T(Y-X\beta-u)} \\
\mathrm{E}(\lambda_R^{-1}|\xi) &= \frac{2r_2+(Y-X\beta-u)^T(Y-X\beta-u)}{2r_1+n-2} \\
\mathrm{E}(\beta|\lambda) &= \frac{\lambda_R\lambda_D}{\lambda_R+\lambda_D+\lambda_R\lambda_D X^TX} X^TY \\
\mathrm{E}(u_i|\lambda) &= \frac{\lambda_R}{\lambda_R+\lambda_D}\left[y_i - x_i\mathrm{E}(\beta|\lambda)\right] \\
\mathrm{Var}(\beta|\lambda) &= \frac{\lambda_R+\lambda_D}{\lambda_R+\lambda_D+\lambda_R\lambda_D X^TX} \\
\mathrm{Var}(u_i|\lambda) &= \frac{1}{\lambda_R+\lambda_D}\left[1 + \frac{\lambda_R^2 x_i^2}{\lambda_R+\lambda_D+\lambda_R\lambda_D X^TX}\right] \\
\mathrm{Cov}(\beta, u_i|\lambda) &= \frac{-\lambda_R x_i}{\lambda_R+\lambda_D+\lambda_R\lambda_D X^TX} \\
\mathrm{Cov}(u_i, u_j|\lambda) &= \frac{\lambda_R^2 x_i x_j}{(\lambda_R+\lambda_D)(\lambda_R+\lambda_D+\lambda_R\lambda_D X^TX)}, \quad i \neq j .
\end{aligned}$$

Now

$$\begin{aligned} y_i - x_i \mathrm{E}(\beta|\lambda) &= y_i - x_i \frac{\lambda_R \lambda_D}{\lambda_R + \lambda_D + \lambda_R \lambda_D X^T X} X^T Y \\ &= \frac{(\lambda_R + \lambda_D) y_i + \lambda_R \lambda_D X^T X (y_i - x_i X^T Y / X^T X)}{\lambda_R + \lambda_D + \lambda_R \lambda_D X^T X} , \end{aligned}$$

and hence $y_i - x_i \mathrm{E}(\beta|\lambda)$ is a convex combination of y_i and $y_i - x_i X^T Y / X^T X$. Let Δ denote the convex hull of the set

$$\left\{ \left(y_i,\ y_i - x_i X^T Y / X^T X \right) \quad \text{for } i = 1, \ldots, n \right\} .$$

Then

$$y_i - x_i \mathrm{E}(\beta|\lambda) - \mathrm{E}(u_i|\lambda) = \frac{\lambda_D}{\lambda_R + \lambda_D} \left[y_i - x_i \mathrm{E}(\beta|\lambda) \right]$$

and hence

$$\left[y_i - x_i \mathrm{E}(\beta|\lambda) - \mathrm{E}(u_i|\lambda) \right]^2 \leq \Delta^2 .$$

Drift for the Block Gibbs Sampler

In this section, we develop a drift condition as in (2.7) for the block Gibbs sampler. Set

$$V_1(\xi) = (Y - X\beta - u)^T (Y - X\beta - u)$$

and $V_2(u) = u^T u$. These will be used in establishing both the drift and minorization conditions.

Theorem 3. *Suppose $r_1 > 1$ and $d_1 > 1$ and fix*

$$\gamma \in \left(\max \left\{ \frac{n}{2r_1 + n - 2}, \frac{n}{2d_1 + n - 2} \right\}, 1 \right) .$$

Define $V(\xi) = \phi_1 V_1(\xi) + \phi_2 V_2(\xi)$. *Then for any positive* ϕ_1 *and* ϕ_2

$$PV(\xi') \leq \gamma V(\xi') + b$$

where $PV(\xi') = E[V(\xi)|\lambda', \xi']$ *and*

$$\begin{aligned} b =& n\phi_1\Delta^2 + \phi_1 X^T X + 2\phi_1 \frac{r_2 n}{2r_1 + n - 2} \\ &+ n\phi_2\Delta^2 + \phi_2 X^T X + 2\phi_2 \frac{d_2(n+1)}{2d_1 + n - 2} \ . \end{aligned}$$

Proof. Notice

$$\mathrm{E}(V(\xi)|\lambda', \xi') = \phi_1 \mathrm{E}(V_1(\xi)|\xi') + \phi_2 \mathrm{E}(V_2(\xi)|\xi') \ .$$

The required expectations will be calculated separately via the following rule:

$$\mathrm{E}[V_i(\xi)|\lambda', \xi'] = \mathrm{E}[V_i(\xi)|\xi'] = \mathrm{E}[\mathrm{E}(V_i(\xi)|\lambda)|\xi'] \ ,$$

for $i = 1, 2$.

First consider the inner expectation in $\mathrm{E}[\mathrm{E}(V_1(\xi)|\lambda)|\xi']$.

$$\begin{aligned} \mathrm{E}(V_1(\xi)|\lambda) &= \sum_{i=1}^{n} \mathrm{E}[(y_i - x_i\beta - u_i)^2|\lambda] \\ &= \sum_{i=1}^{n} \mathrm{Var}[(y_i - x_i\beta - u_i)|\lambda] + [\mathrm{E}(y_i - x_i\beta - u_i)|\lambda]^2 \\ &= \sum_{i=1}^{n} x_i^2 \mathrm{Var}(\beta|\lambda) + \mathrm{Var}(u_i|\lambda) + 2x_i \mathrm{Cov}(\beta, u_i|\lambda) \\ &\quad + [y_i - x_i \mathrm{E}(\beta|\lambda) - \mathrm{E}(u_i|\lambda)]^2 \end{aligned}$$

and hence

$$\begin{aligned}
\mathrm{E}(V_1(\xi)|\lambda) &\leq \sum_{i=1}^{n} x_i^2 \mathrm{Var}(\beta|\lambda) + \mathrm{Var}(u_i|\lambda) + 2x_i \mathrm{Cov}(\beta, u_i|\lambda) + n\Delta^2 \\
&= \frac{n}{\lambda_R + \lambda_D} + \frac{(\lambda_R + \lambda_D)^2 + \lambda_R^2 - 2\lambda_R(\lambda_R + \lambda_D)}{(\lambda_R + \lambda_D)(\lambda_R + \lambda_D + \lambda_R\lambda_D X^T X)} X^T X + n\Delta^2 \\
&= \frac{n}{\lambda_R + \lambda_D} + \frac{\lambda_D^2}{(\lambda_R + \lambda_D)(\lambda_R + \lambda_D + \lambda_R\lambda_D X^T X)} X^T X + n\Delta^2 \\
&\leq \frac{n}{\lambda_R} + \left(\frac{\lambda_D}{\lambda_R + \lambda_D + \lambda_R\lambda_D X^T X} \right) X^T X + n\Delta^2 \\
&\leq \frac{n}{\lambda_R} + \frac{\lambda_D}{\lambda_D} X^T X + n\Delta^2 \\
&= \frac{n}{\lambda_R} + X^T X + n\Delta^2 \, .
\end{aligned}$$

Now for the outer expectation.

$$\begin{aligned}
\mathrm{E}(V_1(\xi)|\xi') &\leq n\mathrm{E}(\lambda_R^{-1}|\xi') + X^T X + n\Delta^2 \\
&= n\frac{2r_2 + (Y - X\beta' - u')^T (Y - X\beta' - u')}{2r_1 + n - 2} + X^T X + n\Delta^2 \\
&= \frac{n}{2r_1 + n - 2} V_1(\xi') + \frac{2r_2 n}{2r_1 + n - 2} + X^T X + n\Delta^2 \, .
\end{aligned}$$

Now consider calculating $\mathrm{E}[\mathrm{E}(V_2(\xi)|\lambda)|\xi']$. We will again consider the inner expectation first.

$$\begin{aligned}
\mathrm{E}(V_2(u)|\lambda) &= \sum_{i=1}^{n} \mathrm{Var}(u_i|\lambda) + [\mathrm{E}(u_i|\lambda)]^2 \\
&= \frac{n}{\lambda_R + \lambda_D} + \frac{\lambda_R^2}{(\lambda_R + \lambda_D)(\lambda_R + \lambda_D + \lambda_R\lambda_D X^T X)} X^T X \\
&\quad + \left(\frac{\lambda_R}{\lambda_R + \lambda_D} \right)^2 n\Delta^2 \\
&\leq \frac{n}{\lambda_D} + \left(\frac{\lambda_R}{\lambda_R + \lambda_D} \right) \left(\frac{\lambda_R}{\lambda_R + \lambda_D + \lambda_R\lambda_D X^T X} \right) X^T X + n\Delta^2 \\
&\leq \frac{n}{\lambda_D} + X^T X + n\Delta^2 \, .
\end{aligned}$$

Hence

$$\begin{aligned}\mathrm{E}(V_2(u)|\xi') &\leq n\mathrm{E}(\lambda_D^{-1}|\xi') + X^T X + n\Delta^2 \\ &= \frac{n}{2d_1 + n - 2} V_2(u') + \frac{2d_2 n}{2d_1 + n - 2} + X^T X + n\Delta^2 \ .\end{aligned}$$

Combining these two calculations yields

$$\begin{aligned}PV(\xi') &\leq \phi_1 \left[\frac{n}{2r_1 + n - 2} V_1(\xi')\right] + \phi_2 \left[\frac{n}{2d_1 + n - 2} V_2(u')\right] + b \\ &\leq \gamma V(\xi') + b \ .\end{aligned}$$

□

Minorization for the Block Gibbs Sampler

In this section, we develop a minorization condition of the form (2.4) for the block Gibbs sampler by closely following the argument of Jones and Hobert (2004). We wish to establish the minorization condition on the set

$$S_B = \{\xi : V(\xi) \leq d_R\} = \{\xi : \phi_1 V_1(\xi) + \phi_2 V_2(\xi) \leq d_R\} \ ,$$

for any $d_R > 0$. First note that S_B is contained in $C_B := C_{B_1} \cap C_{B_2}$, where

$$C_{B_1} = \{\xi : V_1(\xi) \leq d_R/\phi_1\} \quad \text{and} \quad C_{B_2} = \{\xi : V_2(\xi) \leq d_R/\phi_2\} \ .$$

Hence, it suffices to establish a minorization condition that holds on C_B. This will require the following lemma.

Lemma 3. *(Jones and Hobert, 2004, Lemma 4.1) Let* $\mathrm{Gamma}(\alpha, \beta; x)$ *denote the value of the* $\mathrm{Gamma}(\alpha, \beta)$ *density at the point* $x > 0$*. If* $\alpha > 1$*,* $b > 0$*, and* $c > 0$ *are*

fixed, then, as a function of x

$$\inf_{0<\beta<c} \text{Gamma}(\alpha, b+\beta/2; x) = \begin{cases} \text{Gamma}(\alpha, b; x) & \text{if } x < x^* \,, \\ \text{Gamma}(\alpha, b+c/2; x) & \text{if } x > x^* \,, \end{cases}$$

where

$$x^* = \frac{2\alpha}{c} \log\left(1 + \frac{c}{2b}\right).$$

We now present the minorization condition for the block Gibbs sampler.

Theorem 4. *Let* $q(\lambda, \xi)$ *be a density on* $\mathbb{R}_+^2 \times \mathbb{R}^{n+1}$ *such that*

$$q(\lambda, \xi) = \left[\frac{h_1(\lambda_R)}{\int_{\mathbb{R}_+} h_1(\lambda_R) d\lambda_R}\right] \left[\frac{h_2(\lambda_D)}{\int_{\mathbb{R}_+} h_2(\lambda_D) d\lambda_D}\right] \pi(\xi|\lambda, y),$$

where

$$h_1(\lambda_R) = \begin{cases} \text{Gamma}\left(r_1 + \frac{n}{2}, r_2; \lambda_R\right) & \text{if } \lambda_R < \lambda_R^* \\ \text{Gamma}\left(r_1 + \frac{n}{2}, r_2 + \frac{d_R}{2\phi_1}; \lambda_R\right) & \text{if } \lambda_R \geq \lambda_R^* \end{cases}$$

for

$$\lambda_R^* = \frac{\phi_1(2r_1 + n)}{d_R} \log\left(1 + \frac{d_R}{2\phi_1 r_2}\right)$$

and

$$h_2(\lambda_D) = \begin{cases} \text{Gamma}\left(d_1 + \frac{n}{2}, d_2; \lambda_D\right) & \text{if } \lambda_D < \lambda_D^* \\ \text{Gamma}\left(d_1 + \frac{n}{2}, d_2 + \frac{d_R}{2\phi_2}; \lambda_D\right) & \text{if } \lambda_D \geq \lambda_D^* \end{cases}$$

for

$$\lambda_D^* = \frac{\phi_2(2d_1 + n)}{d_R} \log\left(1 + \frac{d_R}{2\phi_2 d_2}\right).$$

Then the following minorization condition is satisfied for the Markov transition density of the block Gibbs sampler:

$$k(\lambda, \xi|\lambda', \xi') \geq \varepsilon q(\lambda, \xi) \qquad \text{for all } \xi' \in C_B$$

where $\varepsilon = \left[\int_{\mathbb{R}_+} h_1(\lambda_R)d\lambda_R\right]\left[\int_{\mathbb{R}_+} h_2(\lambda_D)d\lambda_D\right]$.

Proof. Assume $\xi' \in C_B$ and recall that

$$\pi(\lambda_R|\xi') = \text{Gamma}(r_1 + n/2, r_2 + V_1(\xi')/2; \lambda_R) ,$$
$$\pi(\lambda_D|\xi') = \text{Gamma}(d_1 + n/2, d_2 + V_2(\xi')/2; \lambda_D) .$$

Therefore,

$$\begin{aligned}
k(\lambda, \xi|\lambda', \xi') &= \pi(\lambda_R|\xi', y)\pi(\lambda_D|\xi', y)\pi(\xi|\lambda, y) \\
&\geq \pi(\xi|\lambda, y) \inf_{\xi' \in C_{B_1} \cap C_{B_2}} [\pi(\lambda_R|\xi', y)\pi(\lambda_D|\xi', y)] \\
&\geq \pi(\xi|\lambda, y) \left[\inf_{\xi' \in C_{B_1} \cap C_{B_2}} \pi(\lambda_R|\xi', y)\right] \left[\inf_{\xi' \in C_{B_1} \cap C_{B_2}} \pi(\lambda_D|\xi', y)\right] \\
&\geq \pi(\xi|\lambda, y) \left[\inf_{\xi' \in C_{B_1}} \pi(\lambda_R|\xi', y)\right] \left[\inf_{\xi' \in C_{B_2}} \pi(\lambda_D|\xi', y)\right]
\end{aligned}$$

where

$$\inf_{\xi' \in C_{B_1}} \pi(\lambda_R|\xi', y) = \inf_{\xi': v_1(\xi') \leq d_R/\phi_1} \pi(\lambda_R|\xi', y) = h_1(\lambda_R)$$

and

$$\inf_{\xi' \in C_{B_2}} \pi(\lambda_D|\xi', y) = \inf_{\xi': v_2(\xi') \leq d_R/\phi_2} \pi(\lambda_D|\xi', y) = h_2(\lambda_D).$$

This gives us $k(\lambda, \xi|\lambda', \xi') \geq q(\lambda, \xi)$ for

$$q(\lambda, \xi) \propto \pi(\xi|\lambda)h_1(\lambda_R)h_2(\lambda_D).$$

□

Chapter 3

Monte Carlo Error

As we have seen, MCMC is a common statistical method where the goal is estimating characteristics of a target distribution. An important, and often overlooked, secondary goal is estimation of the associated asymptotic variance in the Markov chain central limit theorem. Specifically, the variance is required in evaluating the Monte Carlo standard error (MCSE) which is useful in measuring the accuracy of the resulting estimate. We introduce several techniques for estimating the variance: batch means, overlapping batch means, regeneration and spectral variance estimation. In addition, we establish conditions under which these methods produce strongly consistent estimators. For batch means and overlapping batch means, we establish conditions which ensure consistency in the mean-square sense. Using mean-square consistency, we calculate "optimal" batch sizes that minimize the asymptotic mean-square error. Finally, we examine the finite sample properties in the context of some examples and provide recommendations for practitioners.

Some of this chapter is contained in Flegal and Jones (2008). The results here are expanded to contain the relevant supporting material from the literature.

3.1 Introduction

Suppose the goal is to calculate $E_\pi g := \int_{\mathsf{X}} g(x)\,\pi(dx)$ where g is a real-valued, π-integrable function on X and π is a probability distribution with support X. Let $X = \{X_1, X_2, X_3, \dots\}$ be a Harris ergodic Markov chain on a general space X having invariant distribution π. The Ergodic Theorem from Chapter 2 shows with probability one,

$$\bar{g}_n := n^{-1} \sum_{i=1}^{n} g(X_i) \to E_\pi g \quad \text{as} \;\; n \to \infty \,. \tag{3.1}$$

Clearly, the resulting Monte Carlo approximation will not be exact. In other words, we would rarely expect our estimate, $\bar{g}_n$, to equal the true quantity of interest, $E_\pi g$. The estimate is bound to be off by some amount, $\bar{g}_n - E_\pi g$, previously defined as the **Monte Carlo error**. Unless $E_\pi g$ is known, we will never know the true Monte Carlo error. However, we can assess this error by estimating the variance from the asymptotic distribution of $\bar{g}_n$.

More specifically, suppose a Markov chain central limit theorem (CLT) exists for g, for any initial distribution,

$$\sqrt{n}(\bar{g}_n - E_\pi g) \xrightarrow{d} \mathrm{N}(0, \sigma_g^2) \quad \text{as} \;\; n \to \infty \tag{3.2}$$

where $\sigma_g^2 := \mathrm{var}_\pi\{g(X_1)\} + 2\sum_{i=2}^{\infty} \mathrm{cov}_\pi\{g(X_1), g(X_i)\}$. The issue we study here is how to consistently estimate σ_g^2.

3.1.1 Stopping the Simulation

Estimating σ_g^2 is an important issue in MCMC output analysis since the estimate can be used to decide when to terminate the simulation or assess the reliability of the current point estimate $\bar{g}_n$; see Flegal et al. (2008), Hobert et al. (2002) and Jones et al. (2006).

When choosing a stopping criteria for MCMC simulations, a practitioner is forced to choose between one long run or several independent smaller runs. Using one long run requires specialized techniques which incorporate the dependent nature of the underlying Markov chain. In contrast, the use of multiple Markov chains (i.e. the Gelman-Rubin Diagnostic from Chapter 1) uses independent chains but each chain may be too short to overcome the initial bias. Alexopoulos and Goldsman (2004) investigate this issue for general processes and conclude that one long run is preferable for nonstationary processes. With this in mind, we will focus on these techniques, but further investigation may be warrented for geometrically (or uniformly) ergodic Markov chains.

A number of other stopping criteria have been proposed based on convergence diagnostics (Cowles and Carlin, 1996). However, convergence diagnostics have done little other than confuse practitioners; see Flegal et al. (2008) and Jones et al. (2006). In general, diagnostics simply ignore the appropriate problem of accurately estimating σ_g^2, and hence we ignore them in this discussion.

Instead, we will focus on two main approaches to stopping an MCMC simulation, fixed-length methods and fixed-width methods.

Fixed-Length Approach

The simplest approach to stopping an MCMC simulation is to run a single chain for a fixed number of iterations then use $\bar{g}_n$ to estimate $E_\pi g$. Practitioners using this approach should report the resulting MCSE to allow the reader to infer the accuracy of the estimate. Hence, estimating σ_g^2 is required. Suppose we can construct an estimator, say $\hat{\sigma}_n^2$, such that $\hat{\sigma}_n^2$ is *mean-square consistent*, or as $n \to \infty$

$$\text{MSE}(\hat{\sigma}_n^2) := E_\pi(\hat{\sigma}^2 - \sigma_g^2)^2 \to 0 \,. \tag{3.3}$$

Damerdji (1995) establishes (3.3) for general stationary processes. Section 3.2.3 specializes this work to MCMC and addresses errors from Damerdji (1995).

Given potential estimators of σ_g^2, we can use (3.3) to choose nuisance parameters to minimize the asymptotic mean-square error (MSE). Song and Schmeiser (1995) propose one such approach to batch size selection for general stationary processes; Section 3.2.3 considers this in the context of MCMC.

Fixed-Width Methods

Suppose we have an idea of the level of accuracy we want in our estimate, $\bar{g}_n$. For example, suppose we want to report estimates with three significant figures. One way of achieving this goal is through evaluation of the Monte Carlo error.

Assessing the Monte Carlo error is usually accomplished by appealing to a Markov chain CLT (Theorem 2). Then one can calculate an estimate of the MCSE of $\bar{g}_n$, say $\hat{\sigma}_n/\sqrt{n}$ and form a confidence interval for $E_\pi g$. If this interval is too large, then the value of n is increased and simulation continues until the interval is sufficiently small. The half-width of the interval is given by

$$t^* \frac{\hat{\sigma}_n}{\sqrt{n}} < \epsilon \tag{3.4}$$

where t^* is an appropriate quantile and $\epsilon > 0$ is the user-specified half-width.

Not knowing ahead of time the number of iterations necessary to ensure the desired precision, the method requires a sequential approach where the chain will be run for a random number of iterations. The theoretical justification of this method requires a *strongly consistent* estimator of σ_g^2. Specifically, an estimator is strongly consistent if with probability one,

$$\hat{\sigma}_n^2 \to \sigma_g^2 \quad \text{as} \quad n \to \infty \, . \tag{3.5}$$

We can use $\hat{\sigma}_n^2$ to form an asymptotically valid confidence interval for $E_\pi g$ or use it to

implement the fixed-width methods for stopping the simulation introduced in Glynn and Whitt (1992) for general processes and in MCMC by Jones et al. (2006).

Remark 2. In general, both (3.3) and (3.5) imply convergence in probability, but mean-square consistency and strong consistency do not in general imply one another.

3.2 Variance Estimation

Estimating σ_g^2 in (3.2) requires specialized techniques that take into account the dependence in the observed Markov chain. Many methods have been proposed to estimate σ_g^2, including non-overlapping batch means (BM), overlapping batch means (OBM), spectral variance methods (SV), and regenerative simulation (RS); see Fishman (1996) for an overview. These methods all have advantages and disadvantages. For example, RS will produce an estimator satisfying (3.5) under the conditions guaranteeing (3.2), however, implementing RS typically requires additional theoretical work. Some common ways of implementing BM and OBM lead to estimators that are demonstrably not consistent (Glynn and Iglehart, 1990; Glynn and Whitt, 1991) thus some authors encourage caution in their use; see Roberts (1996). On the other hand, both BM and OBM are simple to implement but OBM has the reputation for being more efficient that BM. SV methods are also easy to implement, however they have received limited attention in MCMC settings.

In our theoretical work, we focus on conditions which guarantee (3.5) for SV and OBM estimators. These estimators are closely connected since it is well known (Anderson, 1984; Meketon and Schmeiser, 1984) that OBM is, aside from the end-effect terms, equal to a spectral variance estimator with the modified Bartlett lag window. This has also been addressed by Damerdji (1994) but our regularity conditions are weaker. Further, in BM and OBM, batch size selection is still an open research problem. Song and Schmeiser (1995) propose an approach that minimizes the mean-

squared error in (3.3) with a certain rate of convergence. We will show the estimators of σ_g^2 using BM and OBM are mean-square consistent. Under additional conditions, we show the "optimal" batch size in terms of MSE is proportional to $n^{1/3}$. However, we will show that it is impractical to use this result in finite sample settings.

3.2.1 Notation and Assumptions

Throughout this thesis, the symbol "O" will represent the usual big-O notation and the symbol "o" will represent the usual little-o notation. Suppose $f(n)$ and $g(n)$ are functions defined on $\mathbb{R}$. Formally, if $f(n) = O(g(n))$, then there exists an n_0 and $M > 0$ such that $|f(n)| \leq M|g(n)|$ for all $n > n_0$. Further, if $f(n) = o(g(n))$, then for any $M > 0$ there exists an n_0 such that $|f(n)| < M|g(n)|$ for all $n > n_0$.

Either geometric or uniform ergodicity along with a moment condition on g will ensure a Markov chain CLT in (3.2) (Jones, 2004; Roberts and Rosenthal, 2004). However, throughout this discussion we will require a different asymptotic property, specifically a **strong invariance principle** which is now described. Let $B = \{B(t), t \geq 0\}$ denote a standard Brownian motion. A strong invariance principle holds if there exists a nonnegative increasing function $\psi(n)$ on the positive integers, a constant $0 < \sigma_g < \infty$, and a sufficiently rich probability space such that

$$\left| \sum_{i=1}^{n} g(X_i) - nE_\pi g - \sigma_g B(n) \right| = O(\psi(n)) \quad w.p.1 \text{ as } n \to \infty \tag{3.6}$$

where the w.p.1 in (3.6) means for almost all sample paths $\omega \in \Omega$. Alternatively, (3.6) can be expressed as there exists n_0 and a finite random variable C such that for almost all sample paths $\omega \in \Omega$ of the process,

$$\left| \sum_{i=1}^{n} Y_i - nE_\pi g - \sigma_g B(n) \right| < C(\omega)\psi(n) \tag{3.7}$$

for all $n > n_0$. A strong invariance principle is enough to guarantee both (3.1) and (3.2) among other properties; see Philipp and Stout (1975) and Damerdji (1991). Of course, directly establishing a strong invariance principle may be difficult in practice. We will rely on the following result first established by Jones et al. (2006) and later extended by Bednorz and Latuszyński (2007).

Lemma 4. *Let $g : \mathsf{X} \mapsto \mathbb{R}$ be a Borel function and let X be a Harris ergodic Markov chain with invariant distribution π.*

1. *If X is uniformly ergodic and $E_\pi|g|^{2+\delta'} < \infty$ for some $\delta' > 0$, then (3.6) holds with $\psi(n) = n^{1/2-\alpha'}$ where $\alpha' \leq \delta'/(24 + 12\delta')$.*

2. *If X is geometrically ergodic and $E_\pi|g|^{2+\delta+\epsilon} < \infty$ for some $\delta > 0$ and some $\epsilon > 0$, then (3.6) holds with $\psi(n) = n^\alpha \log n$ where $\alpha = 1/(2 + \delta)$.*

3.2.2 Spectral Density Estimation

In this section, we investigate conditions which guarantee strong consistency, i.e. (3.5), of spectral variance estimators. For a detailed review of spectral analysis in a time–series context see Anderson (1994) and Priestley (1981).

Define the process $Y = \{Y_i = g(X_i) - E_\pi g \}$ for $i = 1, 2, 3, \ldots$ and $\bar{Y}_j(k) := k^{-1} \sum_{i=1}^{k} Y_{j+i}$ for $j = 0, \ldots, n - b_n$ and $k = 1, \ldots, b_n$. Further define $\bar{Y}_n := Y_1(n) = n^{-1} \sum_{i=1}^{n} Y_i$. First note that

$$\sigma_g^2 = \sum_{s=-\infty}^{\infty} \gamma(s)$$

where $\gamma(s) := E_\pi [Y_t Y_{t+s}] = E_\pi [(g(X_t) - E_\pi g)(g(X_{t+s}) - E_\pi g)]$. This representation is similar to the spectral density function, or the Fourier transform of the covariance sequence,

$$f(\phi) = \frac{1}{2\pi} \sum_{s=-\infty}^{\infty} \gamma(s) \cos(\phi s)$$

where the parameter ϕ is the frequency.

Remark 3. General time-series settings may require additional work to establish the existence of $f(\phi)$. Given a Markov chain CLT this quantity is certain to exist and be finite.

First consider estimating $\gamma(s)$ using the sample analog of the covariance of lag $s \geq 0$,

$$\gamma_n(s) = n^{-1} \sum_{t=1}^{n-s} (Y_t - \bar{Y}_n)(Y_{t+s} - \bar{Y}_n) \; . \tag{3.8}$$

If $E_\pi g^2 < \infty$, then we can appeal to the Ergodic Theorem to show that $\gamma_n(s) \to \gamma(s)$ almost surely for each fixed s as $n \to \infty$. One could then use the sample analog in (3.8) to estimate $f(\phi)$, though this turns out to be a poor estimator (see Anderson, 1994; Bratley et al., 1987). Instead, we will introduce the use of a weight function, $w_n(\cdot)$, commonly called the *lag window* in the literature. We will restrict our attention to lag windows fulfilling the following requirements.

Assumption 1. The lag window $w_n(\cdot)$ is an even function defined on $\mathbb{Z}$ such that

$$\begin{aligned} |w_n(s)| &\leq 1 \quad \text{for all } n \text{ and } s, \\ w_n(0) &= 1 \quad \text{for all } n, \\ w_n(s) &= 0 \quad \text{for all } |s| \geq b_n \; , \end{aligned}$$

where b_n is the *truncation point*.

We will also require the following assumption on the truncation point.

Assumption 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 monotonically nondecreasing.

Note $2\pi f(0) = \sum_{-\infty}^{\infty} \gamma(s) = \sigma_g^2$ which can be estimated with

$$2\pi f_n(0) = \sum_{s=-(b_n-1)}^{b_n-1} w_n(s)\gamma_n(s) \; . \tag{3.9}$$

The main result of this section follows.

Theorem 5. *Let X be a geometrically ergodic Markov chain with invariant distribution π and $g : \mathsf{X} \to \mathbb{R}$ be a Borel function with $E_\pi|g|^{2+\delta+\epsilon} < \infty$ for some $\delta > 0$ and $\epsilon > 0$. Suppose Assumptions 1 and 2 hold and define $\Delta_1 w_n(k) = w_n(k-1) - w_n(k)$ and $\Delta_2 w_n(k) = w_n(k-1) - 2w_n(k) + w_n(k+1)$. Further suppose (a) $b_n n^{-1} \sum_{k=1}^{b_n} k\,|\Delta_1 w_n(k)| \to 0$ as $n \to \infty$; (b) there exists a constant $c \geq 1$ such that $\sum_n (b_n/n)^c < \infty$; (c) $b_n^{-1} \log n$ stays bounded as $n \to \infty$; (d) $b_n n^{-1} \log n \to 0$ as $n \to \infty$; (e)*

$$b_n n^{2\alpha} (\log n)^3 \left(\sum_{k=1}^{b_n} |\Delta_2 w_n(k)| \right)^2 \to 0 \text{ as } n \to \infty, \text{ and}$$

$$n^{2\alpha} (\log n)^2 \sum_{k=1}^{b_n} |\Delta_2 w_n(k)| \to 0 \text{ as } n \to \infty$$

where $\alpha = 1/(2+\delta)$; and (f) $b_n^{-1} n^{2\alpha} \log n \to 0$ as $n \to \infty$. Then

$$2\pi f_n(0) \to \sigma_g^2 \text{ as } n \to \infty \quad w.p.1.$$

Proof. Consider the following quantity,

$$\hat{\sigma}^2(n) = n^{-1} \sum_{j=0}^{n-b_n} \sum_{k=1}^{b_n} \alpha_n(k) \left(\bar{Y}_j(k) - \bar{Y}_n \right)^2 \tag{3.10}$$

where $\alpha_n(k)$ is a sequence of weights. Proposition 3 in Section 3.4.2 shows there exists a sequence of weights $\alpha_n(k) := k^2 \Delta_2 w_n(k)$ and a sequence d_n due to some end effects,

such that

$$\hat{\sigma}^2(n) = 2\pi f_n(0) - d_n \ .$$

Proposition 7 in Appendix A.1.1 shows $\tilde{\sigma}_*^2 \to 1$ as $n \to \infty$ where $\tilde{\sigma}_*^2$ is the Brownian motion equivalent of (3.10). Lemma 4 with Lemma 8 show $\hat{\sigma}^2(n) - \sigma_g^2 \tilde{\sigma}_*^2 \to 0$ as $n \to \infty$. Finally, Lemma 9 with Lemma 10 shows that $d_n \to 0$ as $n \to \infty$. Thus, combining all of these yields the desired result. □

Remark 4. It is convenient in applications to take $b_n = \lfloor n^\nu \rfloor$ for some $0 < \nu < 1$, and hence (b), (c), and (d) are satisfied.

Lag Windows

Anderson (1994) gives an extensive collection of lag windows satisfying Assumption 1. In this section, we discuss some of the more popular windows.

The Parzen lag windows, $w_n(k) = 1 - |k|^q / b_n^q$ for $|k| < b_n$ and 0 otherwise where $q > 0$. For this type of window, (a) of Theorem 5 reduces to $b_n^2 n^{-1} \to 0$ as $n \to \infty$ and (e) requires $b_n^{1-2q} n^{2\alpha} (\log n)^3 \to 0$ and $b_n^{-q} n^{2\alpha} (\log n)^2 \to 0$ as $n \to \infty$. The modified Bartlett lag window is a special case of a Parzen window where $q = 1$, hence the result applies.

The Tukey-Hanning window, $w_n(k) = [1 + \cos(\pi k / b_n)] / 2$ for $|k| < b_n$ and 0 otherwise. Then (a) reduces to $b_n^2 n^{-1} \to 0$ as $n \to \infty$ and (e) requires $b_n^{-1} n^{2\alpha} (\log n)^3 \to 0$ as $n \to \infty$.

In general, (e) requires $\lim_{b_n \to \infty} w_n(b_n - 1) = 0$. Unfortunately, the truncated periodogram ($w_n(k) = 1$ for $|k| < b_n$ and 0 otherwise), the general Blackman-Tukey window ($w_n(k) = 1 - 2a + 2a\cos(\pi k / b_n)$ for $|k| < b_n$ and 0 otherwise where a is a positive constant), and the scale-parameter class of lag window functions ($w_n(k) = 1 - \delta |k|^q / b_n^q$ for $|k| < b_n$ and 0 otherwise where q and δ are positive constants) do not generally satisfy this requirement. Damerdji (1991) incorrectly applies his result to

the latter two windows listed above.

Using a similar proof as Theorem 6 with results from Damerdji (1991, 1994), one can show the spectral variance estimator is strongly consistent for uniformly ergodic chains. In this case we need to replace condition (e) with (e')

$$b_n n^{1-2\alpha'}(\log n)\left(\sum_{k=1}^{b_n}|\Delta_2 w_n(k)|\right)^2 \to 0 \text{ as } n\to\infty, \text{ and}$$
$$n^{1-2\alpha'}\sum_{k=1}^{b_n}|\Delta_2 w_n(k)| \to 0 \text{ as } n\to\infty$$

where $0 < \alpha' \leq \delta'/(24+12\delta')$ and $\delta' = \delta + \epsilon$. It turns out this result is not very practical when $b_n = \lfloor n^\nu \rfloor$. Consider using the modified Bartlett or Tukey-Hanning lag windows. Either of these require $b_n^2 n^{-1} \to 0$ as $n\to\infty$ and (e') requires $b_n^{-1} n^{1-2\alpha'}(\log n) \to 0$ as $n\to\infty$, but there is no ν value that will satisfy both of these requirements. Parzen windows with $q > 3/2$ will satisfy these conditions.

3.2.3 Batch Means

In non-overlapping batch means the output is broken into blocks of equal size. Suppose the algorithm is run for a total of $n = a_n b_n$ iterations and for $k = 0,\ldots,a_n-1$ define $\bar{Y}_k := b_n^{-1}\sum_{i=1}^{b_n} g(X_{kb_n+i})$. The batch means estimate of σ_g^2 is

$$\hat{\sigma}^2_{BM} = \frac{b_n}{a_n-1}\sum_{k=0}^{a_n-1}(\bar{Y}_k - \bar{g}_n)^2 \,. \tag{3.11}$$

It is well known that generally (3.11) is not a consistent estimator of σ_g^2 (Glynn and Iglehart, 1990; Glynn and Whitt, 1991). On the other hand, Jones et al. (2006) show that if the batch size and number of batches are allowed to increase as the overall length of the simulation does (e.g., by setting $a_n = b_n = \lfloor n^{1/2} \rfloor$) then $\hat{\sigma}^2_{BM} \to \sigma_g^2$ with

probability one as $n \to \infty$.

Proposition 2. *(Jones et al., 2006, Proposition 3) Let X be a geometrically ergodic Markov chain with invariant distribution π and $g : \mathsf{X} \to \mathbb{R}$ be a Borel function with $E_\pi|g|^{2+\delta+\epsilon} < \infty$ for some $\delta > 0$ and $\epsilon > 0$. Suppose Assumption 2 holds (hence $a_n \to \infty$ as $n \to \infty$) and (a) there exists a constant $c \geq 1$ such that $\sum_n (b_n/n)^c < \infty$; and (b) $b_n^{-1} n^{2\alpha} (\log n)^3 \to 0$ as $n \to \infty$ where $\alpha = 1/(2+\delta)$, then as $n \to \infty$, $\hat{\sigma}^2_{BM} \to \sigma_g^2$ w.p.1.*

Proof. Lemma 22 in Appendix A.1.3 shows $\tilde{\sigma}^2_{BM} \to 1$ where $\tilde{\sigma}^2_{BM}$ is the Brownian motion equivalent to (3.11). From Lemma 4 and Lemma 11 we can conclude $\hat{\sigma}^2_{BM} - \sigma_g^2 \tilde{\sigma}^2_{BM} \to 0$ as $n \to \infty$, hence the desired result. □

Notice that there is no assumption of stationarity, implying that burn-in is not required to implement BM. Also, Jones et al. (2006) found that the finite sample properties that can be less desirable than expected, thus we consider OBM.

OBM is just a generalization of BM but it is also well known that the OBM estimator is equal, except for some end-effect terms, to the SV estimator arising from the modified Bartlett lag window–a relationship we will exploit later. Note that there are $n - b_n + 1$ batches of length b_n indexed by k running from 0 to $n - b_n$. OBM averages across all batches. Its estimate of σ_g^2 is

$$\hat{\sigma}^2_{OBM} = \frac{n b_n}{(n - b_n)(n - b_n + 1)} \sum_{j=0}^{n-b_n} (\bar{Y}_j(b_n) - \bar{Y}_n)^2 \,, \tag{3.12}$$

where b_n is the batch length previously defined. The next result establishes strong consistency of the OBM estimator.

Theorem 6. *Let X be a geometrically ergodic Markov chain with invariant distribution π and $g : \mathsf{X} \to \mathbb{R}$ be a Borel function with $E_\pi|g|^{2+\delta+\epsilon} < \infty$ for some $\delta > 0$ and*

$\epsilon > 0$. *Suppose Assumption 2 holds and (a)* $b_n^2 n^{-3/2} \to 0$ *as* $n \to \infty$*; (b) there exists a constant* $c \geq 1$ *such that* $\sum_n (b_n/n)^c < \infty$*; and (c)* $b_n^{-1} n^{2\alpha} (\log n)^3 \to 0$ *as* $n \to \infty$ *where* $\alpha = 1/(2+\delta)$*, then as* $n \to \infty$*,* $\hat{\sigma}^2_{OBM} \to \sigma_g^2$ *w.p.1.*

Proof. The proof is similar to the proof of Theorem 5 with the modified Bartlett lag window. Specifically, if $w_n(k) = 1 - |k|/b_n$ for $|k| < b_n$ and 0 otherwise, then $\alpha_n(b_n) = b_n$ and $\alpha_n(k) = 0$ for all $k = 1, 2, \ldots, b_n - 1$. Furthermore, (3.10) reduces to

$$\hat{\sigma}^2(n) = b_n n^{-1} \sum_{j=0}^{n-b_n} \left(\bar{Y}_j(b_n) - \bar{Y}_n \right)^2 \tag{3.13}$$

which is asymptotically equivalent to (3.12).

It suffices to show that $\hat{\sigma}^2(n) \to \sigma_g^2$ w.p.1. To this end Lemma 4 with Lemma 8 show $\hat{\sigma}^2(n) - \sigma_g^2 \tilde{\sigma}^2(n) \to 0$ as $n \to \infty$ where $\tilde{\sigma}^2(n)$ is the Brownian motion equivalent of (3.13). (With the modified Bartlett window $\sum_{k=1}^{b_n} |\Delta_2 w_n(k)| = b_n^{-1}$, hence (3.26) and (3.27) from Lemma 8 are satisfied if $b_n^{-1} n^{2\alpha} (\log n)^3 \to 0$ as $n \to \infty$.) Proposition 8 in Appendix A.1.2 shows $\tilde{\sigma}^2(n) \to 1$ completing the proof. □

Remark 5. An alternative proof of Theorem 6 follows from Theorem 5 with the modified Bartlett lag window. However, this requires the use of Lemma 10 making condition (a) of Theorem 5 necessary. Specifically, the modified Bartlett window results in $\sum_{k=1}^{b_n} k \, |\Delta_1 w_n(k)| = (b_n + 1)/2$ and condition (a) requires $b_n^2 n^{-1} \to 0$ as $n \to \infty$.

Corollary 1. *Let* X *be a geometrically ergodic Markov chain with invariant distribution* π *and* $g : \mathsf{X} \to \mathbb{R}$ *be a Borel function with* $E_\pi |g|^{2+\delta+\epsilon} < \infty$ *for some* $\delta > 0$ *and* $\epsilon > 0$*. If* $b_n = \lfloor n^\nu \rfloor$ *where* $3/4 > \nu > (1 + \delta/2)^{-1}$*, then* $\hat{\sigma}^2_{OBM} \to \sigma_g^2$ *w.p.1.*

Remark 6. Damerdji (1994) and Jones et al. (2006) show $\hat{\sigma}^2_{BM}$ is strongly consistent for uniformly and geometrically ergodic chains respectively under nearly the same

conditions as required for Theorem 6. Specifically, Proposition 2 shows if $b_n = \lfloor n^\nu \rfloor$, then strong consistency requires $1 > \nu > (1 + \delta/2)^{-1} > 0$ for $\delta > 0$ resulting in less stringent regularity conditions than those required in Theorem 6.

Mean-Square Consistency

We turn our attention to showing that BM and OBM estimators are consistent in the mean-square sense, i.e. (3.3). Let $\hat{\sigma}^2$ be an estimator of σ_g^2 and assume that $X_1 \sim \pi$ so that X is stationary in this subsection. Recall that strong consistency and mean-square consistency do not imply each other, hence we can not appeal to the results in the previous section.

Establishing (3.3) for $\hat{\sigma}^2_{BM}$ is a well studied problem in the operations research literature, see Chien et al. (1997) for an overview. For geometrically ergodic Markov chains where Assumption 2 holds and $E_\pi g^4 < \infty$, Song and Schmeiser (1995) show

$$b_n \text{Bias}\left[\hat{\sigma}^2\right] = \Gamma + o\left(1\right) \ , \tag{3.14}$$

where $\hat{\sigma}^2$ is the estimator from BM or OBM and $\Gamma := -2\sum_{s=1}^{\infty} s\gamma(s)$. Chien et al. (1997) show that Γ is well defined for geometrically ergodic chains if $E_\pi g^2 < \infty$. Under the additional condition that $E_\pi g^{12} < \infty$, Chien et al. (1997) further show

$$\frac{n}{b_n}\text{Var}(\hat{\sigma}^2_{BM}) = 2\sigma_g^4 + o\left(1\right) \ . \tag{3.15}$$

Combining (3.14) and (3.15) imply (3.3) for $\hat{\sigma}^2_{BM}$.

We will establish conditions for (3.3) where $\hat{\sigma}^2$ is $\hat{\sigma}^2_{BM}$ or $\hat{\sigma}^2_{OBM}$ under a less stringent moment condition on g.

Theorem 7. *Let X be a geometrically ergodic Markov chain with invariant distribution π and $g : \mathsf{X} \to \mathbb{R}$ be a Borel function with $E_\pi|g|^{2+\delta+\epsilon} < \infty$ for some $\delta \geq 2$ and*

$\epsilon > 0$. *Suppose Assumption 2 holds and* $E_\pi C^4 < \infty$ *where* C *is defined in* (3.7). *If* $b_n^{-1} n^{2\alpha} (\log n)^3 \to 0$ *as* $n \to \infty$ *where* $\alpha = 1/(2+\delta)$, *then MSE*$(\hat{\sigma}^2) \to 0$ *as* $n \to \infty$ *for BM and OBM.*

Remark 7. The proof of Theorem 7 with results in Damerdji (1995) shows that the conclusions also hold for uniformly ergodic chains. The resulting condition on b_n requires $b_n^{-1} n^{1-2\alpha'} (\log n) \to 0$ as $n \to \infty$ where $\alpha' \leq \delta'/(24 + 12\delta')$ and $\delta' = \delta + \epsilon$.

Optimal Batch Sizes in Terms of MSE

In this section, we will use the previous results to calculate "optimal" batch sizes. Chien et al. (1997) and Song and Schmeiser (1995) study the case of BM. Combining (3.14) and (3.15) yields

$$\mathrm{MSE}(\hat{\sigma}^2_{BM}) = \frac{\Gamma^2}{b_n^2} + \frac{2 b_n \sigma_g^4}{n} + o\left(\frac{1}{b_n^2}\right) + o\left(\frac{b_n}{n}\right) .$$

It is easy to use the above expression to see that $\mathrm{MSE}(\hat{\sigma}^2_{BM})$ will be minimized asymptotically by selecting the "optimal" batch size of

$$\hat{b}^* := \left(\frac{\Gamma^2 n}{\sigma_g^4}\right)^{1/3} .$$

Notice that this optimal batch size is dependent on Γ^2/σ_g^4 which is an unknown parameter relating to the process. However, this result implies that the optimal batch size should increase proportional to $n^{1/3}$.

The main result of this section follows.

Theorem 8. *Let* X *be a geometrically ergodic Markov chain with invariant distribution* π *and* $g : \mathsf{X} \to \mathbb{R}$ *be a Borel function with* $E_\pi |g|^{2+\delta+\epsilon} < \infty$ *for some* $\delta \geq 2$ *and* $\epsilon > 0$. *Suppose Assumption 2 holds and* $E_\pi C^4 < \infty$ *where* C *is defined in* (3.7). *If*

$b_n^{-1} n^{1/2+\alpha} (\log n)^{3/2} \to 0$ *as* $n \to \infty$ *where* $\alpha = 1/(2+\delta)$*, then*

$$\frac{n}{b_n} Var(\hat{\sigma}^2) = c\sigma_g^4 + o(1) \ . \tag{3.16}$$

where $c = 2$ *for BM and* $c = 4/3$ *for OBM.*

Remark 8. If $b_n = \lfloor n^\nu \rfloor$, the result here for geometrically ergodic chains allows $\nu \in (1/2 + \alpha, 1)$.

Remark 9. The proof of Theorem 8 coupled with results from Damerdji (1995) can be applied to uniformly ergodic Markov chains. Specifically, the condition on b_n would be changed to $b_n^{-1} n^{1-\alpha'} (\log n)^{1/2} \to 0$ as $n \to \infty$ where $\alpha' \leq \delta'/(24 + 12\delta')$ and $\delta' = \delta + \epsilon$. However, if $b_n = \lfloor n^\nu \rfloor$, the best we can do is $\nu \in (11/12, 1)$, and hence the result is not very useful.

Combining (3.14) and (3.16) yields

$$\text{MSE}(\hat{\sigma}^2_{OBM}) = \frac{\Gamma^2}{b_n^2} + \frac{c b_n \sigma_g^4}{n} + o\left(\frac{1}{b_n^2}\right) + o\left(\frac{b_n}{n}\right) \ ,$$

which will be minimized asymptotically by selecting the "optimal" batch size of

$$\hat{b}^* := \left(\frac{2\Gamma^2 n}{c\sigma_g^4}\right)^{1/3} .$$

Remark 10. Selecting $b_n = dn^{1/3}$ where d is a constant will not satisfy the necessary conditions of Theorem 8. Specifically, if $b_n = \lfloor n^\nu \rfloor$ then Theorem 8 requires $\nu \in (1/2+\alpha, 1)$. Since the function is increasing around $\hat{b}^*$, we could select $\nu = 1/2+\alpha+\epsilon$ where $\epsilon > 0$.

OBM versus BM

Comparing the results for OBM and BM, we can see that the regularity conditions necessary to ensure strong consistency and mean-square consistency for OBM are more stringent. For strong consistency, OBM requires $b_n^2 n^{-3/2} \to 0$ as $n \to \infty$ while there is no such requirement to implement BM. In the case of a typical sampling plan where $b_n = \lfloor n^\nu \rfloor$ for some $0 < \nu < 1$, OBM results in a smaller range of choices for ν. In fact, we require $\nu < 3/4$ to implement OBM resulting in a higher necessary moment condition on g. For mean-square consistency with OBM, we must appeal to Theorems 7 and 8, and hence a moment condition on C which is not required in Chien et al. (1997) and Song and Schmeiser (1995).

Under this additional restriction, why should one use OBM? Originally argued by Meketon and Schmeiser (1984), $\hat{\sigma}^2_{OBM}$ has a lower asymptotic variance compared to $\hat{\sigma}^2_{BM}$. Looking closely at (3.16) yields

$$\frac{\text{Var}(\hat{\sigma}^2_{OBM})}{\text{Var}(\hat{\sigma}^2_{BM})} \to \frac{2}{3}$$

as $n \to \infty$. Welch (1987) argues that most of this benefit can be achieved by a modest amount of overlapping. For example, using a batch of size 64 and splitting the batch into 4 sub-batches, then it is only necessary to consider the overlapping batches (of length 64) starting at $X_1, X_{17}, X_{33}, X_{49}, X_{65}, \ldots$ This modification is meant to reduce computational time required. This could also reduce necessary memory if one used a sampling plan such that b_n was restricted to values such that $b_n = 2^k$ for $k = 0, 1, \ldots$

3.2.4 Regeneration

Regeneration techniques are another method to ensure strongly consistent estimators of σ_g^2 (Hobert et al., 2002). Suppose the Markov chain transition kernel P satisfies the minorization condition in (2.4) on some set C with $\pi(C) > 0$. As we saw in

Section 2.4.2, then P can be rewritten as a mixture of two transition kernels

$$P(x, dy) = s(x)Q(dy) + (1 - s(x))\, R(x, dy)\ , \tag{3.17}$$

where $s(x) : \mathsf{X} \to [0, 1]$. Then for each $x \in C$ the residual kernel is

$$R(x, dy) = \frac{P(x, dy) - s(x)Q(dy)}{1 - s(x)} \quad \text{for } s(x) < 1\ ,$$

and $R(x, \cdot) := 0$ for $s(x) = 1$.

Instead of using P exclusively to simulate the next step in the Markov chain, the mixture density in (3.17) can be incorporated as follows. Suppose the current state is $X_i = x$.

1. If $x \in C$, generate $\delta_i \sim \text{Bernoulli}\,(s(x))$ then

   (a) if $\delta_i = 0$, generate $X_{i+1} \sim R(x, \cdot)$;

   (b) if $\delta_i = 1$, generate $X_{i+1} \sim Q(\cdot)$.

2. If $x \notin C$ generate $X_{i+1} \sim P(x, \cdot)$ and $\delta_i = 0$.

Notice that if $\delta_i = 1$, the subsequent draw from $Q(\cdot)$ is independent of the current state, hence the chain **regenerates**. Suppose we can start the chain with a draw from $Q(\cdot)$, then every time $\delta_i = 1$ the next step $i + 1$ is a **regeneration time** since X_{i+1} is drawn from $Q(\cdot)$ starting the process over. These **tours** of the Markov chain are i.i.d., meaning standard techniques can be used to establish an alternative CLT, and hence a simple method to estimate the asymptotic variance.

Furthermore, we can avoid drawing from $Q(\cdot)$ entirely by changing the order slightly. Suppose x is the current state, then we can simply generate $X_{i+1} \sim P(x, \cdot)$ in the usual manner. Then if $x \in C$, we can generate $\delta_i | X_i, X_{i+1}$. Nummelin (1984,

p. 62) notes that

$$Pr\left(\delta_i | X_i, X_{i+1}\right) = \frac{s(X_i)q(X_{i+1})}{k(X_{i+1}|X_i)} ,$$

where $q(\cdot)$ and $k(\cdot|x)$ are the densities corresponding to $Q(\cdot)$ and P.

Using the structure above, we now explain how to calculate a consistent estimator of the asymptotic variance. Suppose $X_1 \sim Q(\cdot)$ and that the Markov chain is run for R regenerations, or tours. In other words, the simulation is run until the Rth time that $\delta_i = 1$. Further suppose $0 = \tau_0 < \tau_1 < \tau_2 < \cdots < \tau_R$ are the random regeneration times ($\tau_i := \min\{i > \tau i - 1 : \delta_{i-1} = 1\}$) and $N_1, N_2, \ldots, N_R$ are the random lengths of the tours ($N_i := \tau_i - \tau_{i-1}$). If we define

$$S_i = \sum_{j=\tau_{i-1}}^{\tau_i - 1} g(X_j) ,$$

then the (N_i, S_i) are i.i.d. and the resulting strongly consistent estimator of $E_\pi g$ is

$$\bar{g}_{\tau_R} := \frac{\bar{S}}{\bar{N}} = \frac{1}{\tau_R} \sum_{j=0}^{\tau_R - 1} g(X_j)$$

where $\bar{S} = R^{-1} \sum_{i=1}^{R} S_i$ and $\bar{N} = R^{-1} \sum_{i=1}^{R} N_i$. If the underlying Markov chain is geometrically ergodic and $E_\pi |g|^{2+\epsilon} < \infty$ for some $\epsilon > 0$, then

$$R^{1/2} \left(\bar{g}_{\tau_R} - E_\pi g\right) \xrightarrow{d} \mathrm{N}(0, \sigma_R^2) \tag{3.18}$$

as $R \to \infty$. Using this alternative CLT, Hobert et al. (2002) show that there exists an easily computed strongly consistent estimator of σ_R^2 defined as

$$\hat{\sigma}_R^2 = \frac{\sum_{i=1}^{R} \left(S_i - \bar{g}_{\tau_R} N_i\right)^2}{R\bar{N}^2} .$$

Remark 11. Notice that the CLT from (3.2) is different from the CLT in (3.18).

Specifically, Hobert et al. (2002) show that $\sigma_R^2 = \sigma_g^2 E_\pi s$.

3.3 Examples

In this section, we investigate the finite sample properties in two examples to compare BM, OBM, RS, and SV. First, we examine the AR(1) model and assess the "optimal" batch size selection. Next, we examine a more realistic Bayesian probit regression model and compare our methods to RS. The finite sample properties for the competing methods will be evaluated based on the length and coverage probabilities of confidence intervals.

3.3.1 AR(1) Model

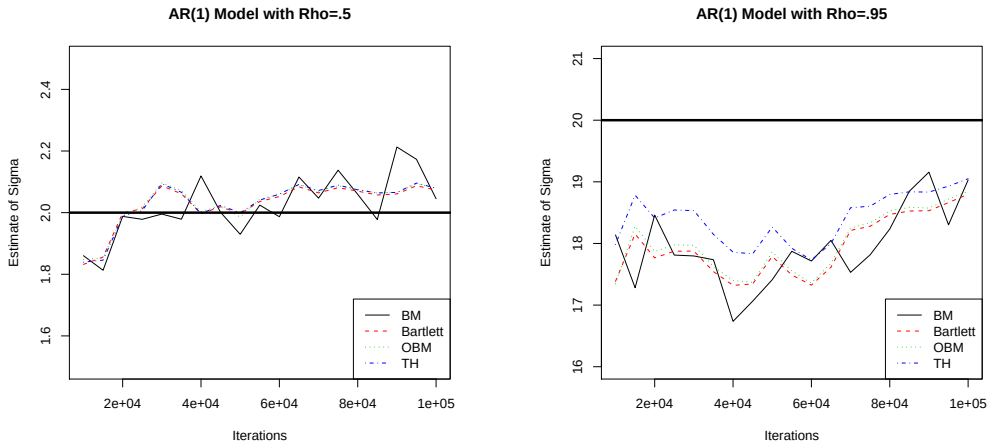
Recall the first order autoregressive process introduced in Chapter 2,

$$X_i = \rho X_{i-1} + \epsilon_i \quad \text{for } i = 1, 2, \dots ,$$

where ϵ_i is an i.i.d. N(0,τ^2) for $i = 1, 2, \dots$. We have previously shown that this chain is geometrically ergodic if $|\rho| < 1$. This is a well studied problem where it is easy to show that $\pi \sim N\left(0, \tau^2/(1-\rho^2)\right)$, and hence $E_\pi X = 0$ and $E_\pi |X|^d < \infty$ for all $d < \infty$. Thus we can appeal to a CLT, strong consistency, and mean-square consistency results. In addition, we can show $\text{cov}_\pi\{X_0, X_i\} = \tau^2\rho^i/(1-\rho)^2$ and $\sigma_g^2 = \tau^2/(1-\rho)^2$. The usefulness of this example is that we can easily control the correlation between iterations to make estimation arbitrarily hard.

Consider estimating $E_\pi X$ with $\bar{x}_n$ and calculating a confidence interval for the resulting estimate. Using the CLT in (3.2), we can calculate the interval

$$\bar{x}_n \pm t^* \frac{\hat{\sigma}}{\sqrt{n}} , \tag{3.19}$$



(a) $\rho = 0.5$ (b) $\rho = 0.95$

Figure 3.1: Plot of $\hat{\sigma}$ versus the number of iterations in the chain using BM, OBM, TH, and Brt for the AR(1) model with $\nu = 1/2$. The solid black lines represent the actual values of σ_g.

where t^* is the appropriate critical value and $\hat{\sigma}$ is an estimate for σ_g^2. We will use multiple replications and the known value of $E_\pi X$ to evaluate BM, OBM, and SV by examining the coverage probabilities of intervals in (3.19). For SV estimators, we will use the Tukey-Hanning window (TH) and Bartlett window (Brt). For BM the degrees of freedom for t^* are $a_n - 1$ and for OBM, TH, and Brt the degrees of freedom for t^* are $n - b_n$.

We consider $\tau^2 = 1$ for the white noise in the AR(1) model and start from $X_1 = 0$. We considered two different autocorrelations, $\rho = \{0.5, 0.95\}$.

Consider the case with $\rho = 0.5$ and $b_n = \lfloor n^\nu \rfloor$ where $\nu = 1/2$. Here the autocorrelation is relatively low resulting in the relative ease in which we can estimate $\sigma_g = 2$. Figure 3.1a shows a plot of $\hat{\sigma}$ versus the number of iterations in the chain using BM, OBM, TH, and Brt. This plot is based on one realization of a chain of length 1e5, but multiple replications result in similar conclusions. We can see that for small numbers of iterations, the estimates for all the methods are biased down.

Method	$b_n =$	Number of Iterations 1e3	5e3	1e4	5e4	1e5
BM	$\lfloor n^{1/3} \rfloor$	0.9315	0.939	0.937	0.942	0.943
Brt		0.93	0.9395	0.936	0.9415	0.944
OBM		0.9305	0.9395	0.936	0.942	0.944
TH		0.936	0.9465	0.9395	0.9465	0.947
BM	$\lfloor n^{1/2} \rfloor$	0.9415	0.948	0.939	0.947	0.949
Brt		0.933	0.946	0.935	0.947	0.9475
OBM		0.9385	0.947	0.9355	0.9475	0.9475
TH		0.9365	0.9465	0.9365	0.948	0.948
BM	$\lfloor n^{2/3} \rfloor$	0.9475	0.9445	0.9385	0.95	0.9465
Brt		0.9105	0.9265	0.9275	0.9445	0.9425
OBM		0.9245	0.935	0.932	0.947	0.944
TH		0.9115	0.927	0.927	0.9435	0.9425

Table 3.1: Table of coverage probabilities for 2000 replications using the AR(1) example with $\rho = 0.5$. All calculations were based on the nominal level of 0.95. The standard errors for these numbers are easily calculated as $\sqrt{\hat{p}(1-\hat{p})/2000}$ which results in a largest standard error of 6.4e-3.

However, after a sufficient number of iterations ($\sim 20,000$) all of the methods seem to provide good estimates for σ_g. Figure 3.1b shows the case where $\rho = 0.95$ and $\nu = 1/2$ resulting in $\sigma_g = 20$. Estimating σ_g here is much more difficult because of the high autocorrelation. We can see that the estimates using all of our methods are biased down even after 100,000 iterations (and much longer in some runs) though TH seems to perform better than the other three.

Next, we wish to compare different batch sizes and variance estimation techniques using finite samples. Using a single chain, σ_g was estimated at five different points (at 1e3, 5e3, 1e4, 5e4, and 1e5 iterations) using each BM, OBM, Brt, and TH with three different sampling plans, $b_n = \lfloor n^\nu \rfloor$ where $\nu = \{1/3, 1/2, 2/3\}$. Here, the calculations for one replication were done using same chain eliminating the effect of different random numbers. Using each setting, a confidence interval for $E_\pi X$ was calculated. To asses the performance of each method, this procedure was repeated

Method	$b_n =$	Number of Iterations 1e3	5e3	1e4	5e4	1e5
BM	$\lfloor n^{1/3} \rfloor$	0.115 (1.8e-4)	0.0534 (4.9e-5)	0.038 (2.8e-5)	0.0172 (7.4e-6)	0.0122 (4.1e-6)
Brt		0.114 (1.5e-4)	0.0531 (4.3e-5)	0.0379 (2.4e-5)	0.0172 (6.2e-6)	0.0122 (3.4e-6)
OBM		0.114 (1.6e-4)	0.0532 (4.3e-5)	0.0379 (2.4e-5)	0.0172 (6.2e-6)	0.0122 (3.4e-6)
TH		0.117 (1.6e-4)	0.0544 (4.4e-5)	0.0387 (2.5e-5)	0.0175 (6.5e-6)	0.0124 (3.6e-6)
BM	$\lfloor n^{1/2} \rfloor$	0.125 (3.6e-4)	0.0556 (1e-4)	0.0394 (6.4e-5)	0.0176 (1.9e-5)	0.0124 (1.1e-5)
Brt		0.119 (2.8e-4)	0.0544 (8.2e-5)	0.0387 (5e-5)	0.0174 (1.5e-5)	0.0124 (9.1e-6)
OBM		0.121 (2.9e-4)	0.0548 (8.3e-5)	0.0389 (5e-5)	0.0175 (1.5e-5)	0.0124 (9.1e-6)
TH		0.121 (3e-4)	0.0549 (8.7e-5)	0.039 (5.2e-5)	0.0175 (1.6e-5)	0.0124 (9.7e-6)
BM	$\lfloor n^{2/3} \rfloor$	0.139 (7.3e-4)	0.0591 (2.3e-4)	0.0412 (1.5e-4)	0.018 (4.8e-5)	0.0127 (3e-5)
Brt		0.116 (4.7e-4)	0.0533 (1.6e-4)	0.0379 (1.1e-4)	0.0173 (3.7e-5)	0.0123 (2.3e-5)
OBM		0.121 (5.2e-4)	0.0548 (1.7e-4)	0.0388 (1.1e-4)	0.0175 (3.8e-5)	0.0124 (2.4e-5)
TH		0.116 (5e-4)	0.0534 (1.7e-4)	0.0379 (1.1e-4)	0.0173 (3.9e-5)	0.0123 (2.5e-5)

Table 3.2: Table of mean confidence interval half-widths with standard errors for 2000 replications using the AR(1) example with $\rho = 0.5$.

Method	$b_n =$	Number of Iterations 1e3	5e3	1e4	5e4	1e5
BM	$\lfloor n^{1/3} \rfloor$	0.614	0.738	0.766	0.842	0.872
Brt		0.606	0.736	0.764	0.841	0.871
OBM		0.61	0.736	0.764	0.842	0.872
TH		0.61	0.74	0.77	0.854	0.886
BM	$\lfloor n^{1/2} \rfloor$	0.838	0.903	0.9155	0.94	0.9425
Brt		0.807	0.893	0.911	0.9365	0.9385
OBM		0.821	0.895	0.913	0.937	0.9395
TH		0.822	0.9055	0.9235	0.943	0.945
BM	$\lfloor n^{2/3} \rfloor$	0.927	0.9385	0.933	0.948	0.9465
Brt		0.872	0.916	0.9185	0.944	0.942
OBM		0.89	0.925	0.924	0.9455	0.943
TH		0.885	0.92	0.924	0.9435	0.9425

Table 3.3: Table of coverage probabilities for 2000 replications using the AR(1) example with $\rho = 0.95$. All calculations were based on the nominal level of 0.95. The standard errors for these numbers are easily calculated as $\sqrt{\hat{p}(1-\hat{p})/2000}$ which results in a largest standard error of 0.011.

Method	$b_n =$	Number of Iterations 1e3	5e3	1e4	5e4	1e5
BM	$\lfloor n^{1/3} \rfloor$	0.544 (1.4e-3)	0.319 (3.9e-4)	0.244 (2.2e-4)	0.129 (6.3e-5)	0.0973 (3.5e-5)
Brt		0.536 (1.3e-3)	0.317 (3.9e-4)	0.243 (2.2e-4)	0.129 (6.2e-5)	0.0973 (3.5e-5)
OBM		0.539 (1.4e-3)	0.318 (3.9e-4)	0.244 (2.2e-4)	0.129 (6.2e-5)	0.0973 (3.5e-5)
TH		0.54 (1.3e-3)	0.322 (3.9e-4)	0.247 (2.2e-4)	0.132 (6.1e-5)	0.0999 (3.4e-5)
BM	$\lfloor n^{1/2} \rfloor$	0.883 (2.9e-3)	0.478 (8.9e-4)	0.355 (5.7e-4)	0.168 (1.8e-4)	0.121 (1.1e-4)
Brt		0.835 (2.6e-3)	0.467 (8.3e-4)	0.349 (5.1e-4)	0.167 (1.6e-4)	0.12 (9.3e-5)
OBM		0.854 (2.7e-3)	0.471 (8.4e-4)	0.351 (5.2e-4)	0.167 (1.6e-4)	0.12 (9.3e-5)
TH		0.855 (2.6e-3)	0.482 (8.3e-4)	0.361 (5.1e-4)	0.172 (1.6e-4)	0.123 (9.6e-5)
BM	$\lfloor n^{2/3} \rfloor$	1.24 (6.7e-3)	0.57 (2.2e-3)	0.403 (1.4e-3)	0.179 (4.8e-4)	0.127 (3e-4)
Brt		1.01 (4.8e-3)	0.514 (1.7e-3)	0.371 (1.1e-3)	0.171 (3.7e-4)	0.122 (2.3e-4)
OBM		1.07 (5.3e-3)	0.529 (1.8e-3)	0.38 (1.1e-3)	0.174 (3.8e-4)	0.123 (2.4e-4)
TH		1.05 (4.9e-3)	0.527 (1.7e-3)	0.377 (1.1e-3)	0.172 (3.9e-4)	0.123 (2.5e-4)

Table 3.4: Table of mean confidence interval half-widths with standard errors for 2000 replications using the AR(1) example with $\rho = 0.95$.

for 2000 independent replications recording the resulting coverage probabilities and confidence interval lengths.

The results from this simulation with $\rho = 0.5$ are in Table 3.1. In the calculations with $\nu = 1/3$ and $\nu = 1/2$, all of the calculated coverage probabilities are within 2 standard errors of the nominal .95 level when at least 5e3 iterations are used. We can also see that for all the settings, the coverage probabilities improve as the number of iterations increase. The choice of $\nu = 2/3$ seems to slightly underestimate the coverage probabilities for small numbers of iterations. Examining the mean confidence interval length in Table 3.2, it is no surprise that the lengths decrease for all of the settings as the number of iterations increase. All of the methods produce similar interval lengths, however, they are slightly longer using BM as a result of the selection of t^*. This is clearly magnified when $\nu = 2/3$ and there is a smaller number of batches. In general, when $\rho = 0.5$ the problem is relatively easy, and all the methods and settings seem to perform well.

Consider the more difficult problem of $\rho = 0.95$. Table 3.3 shows the calculated coverage probabilities. We can see the coverage probabilities get closer to the nominal .95 level as the number of iterations increases. We can also see that as the ν increases

the confidence intervals become more accurate because the strong correlation in the model is better captured with larger batch sizes. In this case, the choice of $\nu = 1/3$ performs much worse than the other options analyzed. For $\nu = 2/3$, BM performs better in terms of coverage probabilities but results in longer confidence intervals (see Table 3.4). Again, this is a result of the selection of t^* for BM.

3.3.2 Bayesian Probit Regression

Suppose $Y_1, \ldots, Y_m$ are independent Bernoulli random variables with $Pr(Y_i = 1) = \Phi(x_i^T\beta)$ where x_i is a $p \times 1$ vector of known covariates associated with Y_i, β is a $p \times 1$ vector of unknown regression coefficients, and $\Phi(\cdot)$ denotes the standard normal distribution function. Then for $y_i \in \{0, 1\}$

$$\Pr(Y_1 = y_1, \ldots, Y_n = y_m|\beta) = \prod_{i=1}^{m} \Phi(x_i^T\beta)^{y_i} \left[1 - \Phi(x_i^T\beta)\right]^{1-y_i}.$$

Bayesian inference on β with a flat prior (p-dimensional Lebesgue measure) is common resulting in

$$\pi(\beta|y) \propto \prod_{i=1}^{m} \Phi(x_i^T\beta)^{y_i} \left[1 - \Phi(x_i^T\beta)\right]^{1-y_i},$$

which under regularity conditions is proper (Roy and Hobert, 2007).

We will sample from $\pi(\beta|y)$ using the PX-DA algorithm of Liu and Wu (1999). First, let $TN(\mu, \sigma^2, w)$ denote a normal distribution with mean μ and variance σ^2 that is truncated to be positive if $w = 1$ and negative if $w = 0$. The procedure requires:

1. Draw $z_1, \ldots, z_m$ independently with $z_i \sim TN(x_i^T\beta, 1, y_i)$.
2. Draw $g^2 \sim \Gamma\left(\frac{n}{2}, \frac{1}{2}\sum_{i=1}^{m}\left[z_i - x_i^T(X^TX)^{-1}X^Tz\right]^2\right)$ and set $z' = (gz_1, \ldots, gz_m)^T$.
3. Draw $\beta' \sim N\left((X^TX)^{-1}X^Tz', (X^TX)^{-1}\right)$.

	$\hat{\beta}_i$	$\hat{\sigma}_{\beta_i}$	MCSE
β_0	-3.0192	11.830	0.004
β_1	6.9136	22.654	0.008
β_2	3.9804	14.765	0.005

Table 3.5: Results from 9e6 iterations for the Bayesian probit regression using the Lupus data from van Dyk and Meng (2001). These values were treated as the "truth" for estimating confidence interval coverage probabilities.

We will use the general framework above to analyze the Lupus Data from van Dyk and Meng (2001). The goal of this example is to predict the occurrence of latent membranous lupus nephritis using x_{i1}, the difference between IgG3 and IgG4 (immunoglobulin G), and x_{i2}, IgA (immunoglobulin A). The response variable is y_i, an indicator of the disease (1 for present). We consider the a Bayesian analysis using a flat prior of the following model

$$\Pr(Y_i = 1) = \Phi\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}\right).$$

Hence, we are interested in estimating the regression parameters $\beta := (\beta_0, \beta_1, \beta_2)$.

Chen and Shao (2001) show conditions to ensure the appropriate moment conditions for estimating the posterior expectation of the regression parameters, β. Roy and Hobert (2007) verify these conditions and show conditions to ensure the chain is geometrically ergodic.

In the following sections, we will attempt to compare different methods based on their coverage probabilities. However, this requires the actual value of β which can never be known in a real example such as this. In an attempt to solve this problem, we will calculate a very precise estimate of β from one long run, and treat this as the "truth" when evaluating confidence intervals in future sections. Table 3.5 shows the calculated values from one long run of 9e6 iterations. MCSEs were calculated using BM with a batch size of $n^{1/2}$.

Fixed-Width Methods

Suppose we want to estimate the three parameters of interest to within ± 0.25. To this end, we will use the fixed-width methods outlined by Jones et al. (2006). In this example, the first confidence interval will be calculated at 1e4 iterations in the chain. If the maximum half-width was greater than 0.25, then 1000 iterations were added to the chain before checking again. Formally, a simulation was terminated when

$$t^* \frac{\hat{\sigma}}{\sqrt{n}} + 0.25 I(n < 10000) < 0.25$$

where t^* is the appropriate critical value and $\hat{\sigma}$ is an estimate for σ_g. Ensuring the resulting confidence intervals are asymptotically valid requires a strongly consistent estimator of σ_g^2. We have shown conditions that result in strongly consistent estimators using BM, OBM, Brt, and TH. The goal in this section, is to compare coverage probabilities and confidence interval lengths for each combination of BM, OBM, Brt, and TH used to estimate σ_g^2 with sampling plans, $b_n = \lfloor n^\nu \rfloor$ where $\nu = \{1/3, 1/2\}$.

The simulations were started from the maximum likelihood estimate of β given by $\hat{\beta} = (-1.778, 4.374, 2.482)$. To estimate the coverage probabilities and interval lengths in this example, we ran 1000 independent replications of the procedure outlined above. Table 3.6 shows the calculated coverage probabilities for β, mean number of iterations at termination, and mean confidence interval lengths for each method.

With $\nu = 1/3$, the results are terrible with any estimate σ_g^2. With the smaller truncation point (or batch size), the estimates are not capturing enough of the correlation in the chain. We can also see that all of the simulations stop very early, most with 10,000 or 11,000 iterations in the chain resulting poor estimates of β.

With $\nu = 1/2$, all of the methods result in coverage probabilities slightly lower than the nominal 0.95 level. It also appears that using TH results in slightly better coverage probabilities while this method requires slightly more simulation effort. The lengths

		β	BM	Brt	OBM	TH
$b_n = \lfloor n^{1/3} \rfloor$	Coverage Probability	β_0	0.742	0.733	0.730	0.742
		β_1	0.739	0.738	0.735	0.742
		β_2	0.739	0.727	0.726	0.741
	N	-	1.1e4 (32)	1.07e4 (29)	1.08e4 (29)	1.1e4 (32)
	C.I. Length	β_0	0.128	0.128	0.128	0.128
		β_1	0.243	0.242	0.242	0.243
		β_2	0.160	0.159	0.159	0.160
$b_n = \lfloor n^{1/2} \rfloor$	Coverage Probability	β_0	0.929	0.927	0.932	0.936
		β_1	0.921	0.921	0.923	0.936
		β_2	0.924	0.927	0.929	0.930
	N	-	2.68e4 (99)	2.66e4 (94)	2.68e4 (95)	2.86e4 (97)
	C.I. Length	β_0	0.129	0.130	0.130	0.130
		β_1	0.245	0.247	0.247	0.247
		β_2	0.160	0.162	0.162	0.162

Table 3.6: Summary of results for using fixed-width methods for the Lupus data Bayesian probit regression. Coverage probabilities using calculated half-width have MCSEs of 1.4e-2 when $b_n = \lfloor n^{1/3} \rfloor$ and between 7.7e-3 and 8.5e-3 when $b_n = \lfloor n^{1/2} \rfloor$. The table also shows the mean simulation effort at termination in terms of number of iterations. The mean confidence interval lengths reported all have MCSEs below 2e-4.

of the resulting half-widths showed virtually no difference between the methods.

Bonferonni Correction

Until this point, we have examined the performance of multiple confidence intervals individually. Alternatively, this section considers a Bonferonni correction to calculate simultaneous confidence intervals. As before, the simulation will be run until the maximum calculated half-width is below 0.25. However, instead of using nominal 95% confidence intervals to decide when to stop the simulation, we will use nominal 98 1/3% confidence intervals. The resulting simultaneous confidence intervals should have at least nominal 95% coverage based on the Bonferonni correction.

To examine the finite sample properties, we ran 1000 replications of the procedure outlined in the previous section changing only the critical value used to stop the

		β	BM	Brt	OBM	TH
$b_n = \lfloor n^{1/3} \rfloor$	Coverage Probability	β_0	0.852	0.847	0.846	0.853
		β_1	0.847	0.844	0.843	0.847
		β_2	0.856	0.841	0.843	0.860
	Simultaneous	β	0.819	0.806	0.807	0.825
	N	-	1.81e4 (59)	1.75e4 (59)	1.75e4 (59)	1.82e4 (60)
	C.I. Length	β_0	0.130	0.130	0.130	0.130
		β_1	0.247	0.247	0.247	0.247
		β_2	0.162	0.162	0.162	0.162
$b_n = \lfloor n^{1/2} \rfloor$	Coverage Probability	β_0	0.974	0.971	0.969	0.977
		β_1	0.975	0.973	0.972	0.978
		β_2	0.972	0.971	0.968	0.978
	Simultaneous	β	0.964	0.961	0.959	0.972
	N	-	4.08e4 (131)	4.08e4 (125)	4.11e4 (125)	4.37e4 (127)
	C.I. Length	β_0	0.129	0.130	0.130	0.130
		β_1	0.245	0.248	0.248	0.248
		β_2	0.161	0.162	0.162	0.162

Table 3.7: Summary of results for using fixed-width methods with a Bonferonni correction for the Lupus data Bayesian probit regression. Coverage probabilities using calculated half-width have MCSEs of between 1.1e-2 and 1.3e-2 when $b_n = \lfloor n^{1/3} \rfloor$ and between 4.6e-3 and 6.3e-3 when $b_n = \lfloor n^{1/2} \rfloor$. The table also shows the mean simulation effort at termination in terms of number of iterations. The mean confidence interval lengths reported all have MCSEs below 1e-4.

simulation. Table 3.7 shows the calculated coverage probabilities for β, mean number of iterations at termination, and mean confidence interval lengths for each method. The resulting coverage probabilities for β_0, β_1, and β_2 have a nominal level of 0.9833 while the resulting nominal simultaneous level is 0.95.

With $\nu = 1/3$, the results have improved but are still poor with any estimate σ_g^2. The simulations ran longer than the minimum values but are still not capturing enough of the correlation in the chain. With $\nu = 1/2$, all individual confidence intervals perform well with observed coverage probabilities close to the nominal 0.9833 level. The simultaneous intervals have observed coverage probabilities greater than the 0.95 nominal level meaning the estimates are correlated. Again, the resulting

	β	BM	Brt	OBM	TH	RS
Coverage Probability	β_0	0.946	0.944	0.944	0.948	0.940
	β_1	0.941	0.941	0.941	0.946	0.938
	β_2	0.952	0.950	0.950	0.955	0.937
C.I. Length	β_0	0.027	0.027	0.027	0.028	0.028
	β_1	0.052	0.052	0.052	0.053	0.053
	β_2	0.034	0.034	0.034	0.035	0.034

Table 3.8: Coverage probabilities and mean confidence interval lengths comparing BM, OBM, and SV using 7e5 Iterations to RS. MCSEs vary between 6.6e-3 and 7.7e-3 for the coverage probabilities and are less that 3e-4 for the mean interval lengths.

half-widths showed virtually no difference.

Comparison to Regeneration

In this section, we will use the same example to compare our methods to RS. Again, we will compere the methods by looking at the resulting coverage probabilities and confidence interval length. Roy and Hobert (2007) implement RS for this example which we use here under identical settings other than the number of regenerations. We implemented RS starting from the appropriate residual density and ran the simulation until there were 50 regenerations in the chain. This procedure was repeated 1000 times resulting in mean simulation effort of 7.12e5 (3200). For an appropriate comparison in terms of simulation effort, confidence intervals for β were calculated using BM, OBM, and SV from simulated chains with 7e5 iterations. Again, 1000 replications of this procedure were done to estimate the coverage probabilities. Table 3.8 shows the resulting coverage probabilities for β using the "truth" calculated previously and the mean confidence interval lengths. The results from all the methods are within two standard errors of the nominal 0.95 level meaning all of the methods provide quality estimates. In addition, for each β_i, the confidence interval lengths were very close for all methods.

3.3.3 Summary

In our examples, we consider different estimators of σ_g^2. In general, all of the methods considered resulted in similar performance for a given simulation setting. Recall, OBM and Brt are asymptotically equivalent and the simulation results show there is little difference between the two in finite samples. In our experience, Brt (and SV methods in general) tended to run slightly faster computationally than OBM. The TH estimator exhibits very similar behavior to OBM and Brt, though there seems to be a slight improvement in performance. Confirming the theoretical results from Section 3.2.3, the estimator from BM was more variable than OBM (Figure 3.1). This result was consistent in both examples in multiple realizations.

Using the Bayesian probit regression model we compared our methods to RS. The resulting simulation showed all the methods performed very well. The advantage of RS is that the actual chain does not need to be stored as the simulation progresses. However, RS requires a considerable theoretical cost that is likely to dissuade a practitioner. The resulting simulation is also dependent on the length of the regeneration tours which can become very long as the dimension increases (Johnson and Jones, 2008). In contrast, BM, OBM, and SV are relatively simple to implement though they can require saving the entire chain. Given the current price of computer memory, this is clearly not the obstacle it was in the past.

The second simulation goal was to investigate the finite sample behavior of different batch size selection. Theoretically, we showed that the batch size should increase at a rate proportional to $n^{1/3}$ where the proportionality constant is unknown. In our examples, using $b_n = \lfloor n^{1/3} \rfloor$ seemed to give very poor results because the batch size or truncation point was too small. In realistic examples with higher correlations, the larger batch size $b_n = \lfloor n^{1/2} \rfloor$ worked well agreeing with the previous work of Jones et al. (2006). Our investigation of $b_n = \lfloor n^{2/3} \rfloor$ worked well in high correlation settings, though for long chains more computational effort was necessary.

3.4 Proofs and Calculations

3.4.1 Results for Proof of Lemma 4

This proof is contained in Jones et al. (2006) and extended by Bednorz and Latuszyński (2007). The first part of the lemma is an immediate consequence of the Theorem 4.1 of Philipp and Stout (1975) and the fact that uniformly ergodic Markov chains enjoy exponentially fast uniform mixing (see Chapter 2). The second part follows from our Lemma 5 and Theorem 2.1 in Csáki and Csörgő (1995).

Lemma 5. *Let X be a Harris ergodic Markov chain on X with invariant distribution π. Assume that* (2.4) *holds and that X is geometrically ergodic. If $E_\pi|g|^{p+\delta} < \infty$ for some $p > 0$ and $\delta > 0$, then $E_Q N_1^p < \infty$ and $E_Q S_1^p < \infty$.*

Preliminary Results

In the proof of this Lemma 5, we will require two additional Lemmas.

Lemma 6. *(Hobert et al., 2002, Lemma 1) Let X be a Harris ergodic Markov chain and assume that* (2.4) *holds. Then for any function $h : \mathsf{X}^\infty \mapsto \mathbb{R}$*

$$E_\pi|h(X_1, X_2, \dots)| \geq cE_Q|h(X_1, X_2, \dots)|,$$

where $c = E_\pi s$.

Proof. For any measurable set A it follows from (2.4) that

$$\pi(A) = \int_{\mathsf{X}} \pi(dx)P(x, A) \geq Q(A) \int_{\mathsf{X}} \pi(dx)s(x) \tag{3.20}$$

and hence $\pi(\cdot) \geq cQ(\cdot)$. Next note that

$$E_\pi|h(X_1, X_2, \dots)| = E_\pi \left[E\left\{|h(X_1, X_2, \dots)| \mid X_1\right\}\right] .$$

The inner expectation is a nonnegative function of X_1 not depending on the starting distribution. Thus, we can use (3.20) and the Markov property to obtain

$$E_\pi|h(X_1, X_2, \dots)| \geq cE_Q\left[E\left\{|h(X_1, X_2, \dots)| \mid X_1\right\}\right] = cE_Q|h(X_1, X_2, \dots)|.$$

□

Lemma 7. *(Hobert et al., 2002, Lemma 2) Let X be a Harris ergodic Markov chain and assume that (2.4) holds. If X is geometrically ergodic, then there exists a $\beta > 1$ such that $E_\pi\beta^{\tau_1} < \infty$.*

Proof. First notice that $\tau_1 = \min\{i > 0 : (X_{i-1}, \delta_{i-1}) \in \mathsf{X} \times \{1\}\}$; or just the hitting time on the set $\mathsf{X} \times \{1\}$. Also note that X and X' converge to stationarity at the exact same rate, consequently, since X is geometrically ergodic, so is X'. Now let π' denote the invariant distribution of X' and note that a random vector (X, δ) with distribution π' satisfies $X \sim \pi(\cdot)$, and, conditional on X, $\delta|X \sim \text{Bernoulli}(s(X))$. Thus $\pi'(\mathsf{X} \times \{1\}) = E_\pi(s) > 0$, and, since X' is geometrically ergodic, Theorem 2.5 of Nummelin and Tuominen (1982) then implies that there exists a $\beta > 1$ such that

$$E_\pi\beta^{\tau_1} < \infty.$$

□

Corollary 2. *Assume the conditions in Lemma 7. For any $a > 0$*

$$\sum_{i=0}^{\infty}[Pr_\pi(\tau_1 \geq i+1)]^a \leq (E_\pi\beta^{\tau_1})^a \sum_{i=0}^{\infty}\beta^{-a(i+1)} < \infty\ .$$

Proof of Lemma 5

By Lemma 6, it is enough to verify that $E_\pi\tau_1^p < \infty$ and $E_\pi S_1^p < \infty$. Lemma 7 shows that $E_\pi\tau_1^p < \infty$ for any $p > 0$.

To show that $E_\pi S_1^p < \infty$, we will first note that

$$C := \left(\left(E_\pi |g(X_i)|^{p+\delta} \right)^{\frac{p}{p+\delta}} \right)^{1/p} < \infty \, . \tag{3.21}$$

For $p \geq 1$ we use the triangle inequality in L^p, Hölder's inequality, the inequality in (3.21) and finally Corollary 2.

$$\begin{aligned}
(E_\pi S_1^p)^{1/p} &\leq \left[E_\pi \left(\sum_{i=0}^{\tau_1 - 1} |g(X_i)| \right)^p \right]^{1/p} \\
&= \left[E_\pi \left(\sum_{i=0}^{\infty} \mathbf{I}(i \leq \tau_1 - 1) |g(X_i)| \right)^p \right]^{1/p} \\
&\leq \sum_{i=0}^{\infty} \left[E_\pi \mathbf{I}(i \leq \tau_1 - 1) |g(X_i)|^p \right]^{1/p} \\
&\leq \sum_{i=0}^{\infty} \left[(E_\pi \mathbf{I}(i \leq \tau_1 - 1))^{\frac{\delta}{p+\delta}} \left(E_\pi |g(X_i)|^{p+\delta} \right)^{\frac{p}{p+\delta}} \right]^{1/p} \\
&= C \sum_{i=0}^{\infty} \left(P_\pi(\tau_1 \geq i+1) \right)^{\frac{\delta}{p(p+\delta)}} < \infty \, .
\end{aligned}$$

For $0 < p < 1$ we can use the fact x^p is concave and then proceed similarly as above to obtain

$$\begin{aligned}
E_\pi S_1^p &\leq E_\pi \left(\sum_{i=0}^{\infty} \mathbf{I}(i \leq \tau_1 - 1) |g(X_i)| \right)^p \\
&\leq \sum_{i=0}^{\infty} E_\pi \mathbf{I}(i \leq \tau_1 - 1) |g(X_i)|^p \\
&= C^p \sum_{i=0}^{\infty} \left(P_\pi(\tau_1 \geq i+1) \right)^{\frac{\delta}{p+\delta}} < \infty \, .
\end{aligned}$$

3.4.2 Results for Proof of Theorem 5

Recall that $X = \{X_1, X_2, \dots\}$ is a Harris ergodic Markov chain. Define the process $Y = \{Y_i = g(X_i) - E_\pi g\}$ for $i = 1, 2, 3, \dots$ with $\bar{Y}_j(k) := k^{-1} \sum_{i=1}^{k} Y_{j+i}$ for $j = 0, \dots, n - b_n$ and $k = 1, \dots, b_n$ and $\bar{Y}_n := Y_1(n) = n^{-1} \sum_{i=1}^{n} Y_i$.

Proposition 3. *(Damerdji, 1991, Theorem 3.1) Under Assumption 1, there exist sequences $\alpha_n(k)$ and d_n such that $\hat{\sigma}^2(n) = 2\pi f_n(0) - d_n$ where*

$$\alpha_n(k) = k^2 \Delta_2 w_n(k) \text{ and}$$
$$d_n = n^{-1} \left(\left[\sum_{l=1}^{b_n} \Delta_1 w_n(l) \left(\sum_{i=1}^{l-1} Z_i^2 + \sum_{i=n-b_n+l+1}^{n} Z_i^2 \right) \right] + 2 \sum_{s=1}^{b_n-1} \left[\sum_{l=1}^{b_n-s} \Delta_1 w_n(s+l) \left(\sum_{i=1}^{l-1} Z_i Z_{s+i} + \sum_{i=n-b_n+l+1}^{n-s} Z_i Z_{s+i} \right) \right] \right)$$

where $Z_i = Y_i - \bar{Y}_n$ for all $i = 1, 2, \dots, n$ and any empty sums are defined to be zero.

Proof. Damerdji (1991) contains a general proof for Proposition 3 that generalizes this result to all frequencies. This proof contains the specific proof for $\phi = 0$ with some added details for clarification.

For clarity, we will define $w_n(k) := w_k$, $\Delta_1 w_n(k) := \Delta_1 w_k$, and $\Delta_2 w_n(k) := \Delta_2 w_k$. Then we can notice that

$$\Delta_1 w_l = \sum_{k=l}^{b_n} \Delta_2 w_k \ , \tag{3.22}$$

$$\sum_{l=s+1}^{b_n} \Delta_1 w_l = w_s \ , \text{ and} \tag{3.23}$$

$$\sum_{l=1}^{b_n} \Delta_1 w_l = 1 \ . \tag{3.24}$$

Recall that $Z_i = Y_i - \bar{Y}_n$; therefore

$$(\bar{Y}_j(k) - \bar{Y}_n)^2 = k^{-2}\left((Y_{j+1} + \cdots + X_{j+k}) - k\bar{Y}_n\right)^2 = k^{-2}\left(\sum_{l=1}^{k} Z_{j+l}\right)^2$$
$$= k^{-2}\left(\sum_{l=1}^{k} Z_{j+l}^2 + 2\sum_{s=1}^{k-1}\sum_{l=1}^{k-s} Z_{j+l}Z_{j+l+s}\right)$$

and from (3.10) with $\alpha_n(k) = k^2\Delta_2 w_n(k)$,

$$\begin{aligned}\hat{\sigma}^2(n) &= n^{-1}\sum_{j=0}^{n-b_n}\sum_{k=1}^{b_n} k^2\Delta_2 w_k(\bar{Y}_j(k) - \bar{Y}_n)^2 \\ &= n^{-1}\sum_{j=0}^{n-b_n}\sum_{k=1}^{b_n}\Delta_2 w_k\left[\sum_{l=1}^{k} Z_{j+l}^2\right] + \left[2n^{-1}\sum_{j=0}^{n-b_n}\sum_{k=1}^{b_n}\Delta_2 w_k\sum_{s=1}^{k-1}\sum_{l=1}^{k-s} Z_{j+l}Z_{j+l+s}\right] .\end{aligned} \tag{3.25}$$

By permuting sums in (3.25), we will arrive at the characterization. Denote the first term in (3.25) as A and the second as B. Then we have that

$$\begin{aligned}A &= n^{-1}\sum_{j=0}^{n-b_n}\sum_{k=1}^{b_n}\sum_{l=1}^{k}\Delta_2 w_k Z_{j+l}^2 \\ &= n^{-1}\sum_{j=0}^{n-b_n}\sum_{l=1}^{b_n}\sum_{k=l}^{b_n}\Delta_2 w_k Z_{j+l}^2 \\ &= n^{-1}\sum_{j=0}^{n-b_n}\sum_{l=1}^{b_n} Z_{j+l}^2\sum_{k=l}^{b_n}\Delta_2 w_k .\end{aligned}$$

Then at lag $s = 0$ from (3.8) and (3.22),

$$A = n^{-1} \sum_{j=0}^{n-b_n} \sum_{l=1}^{b_n} \Delta_1 w_l Z_{j+l}^2 = \sum_{l=1}^{b_n} \Delta_1 w_l n^{-1} \sum_{j=0}^{n-b_n} Z_{j+l}^2$$
$$= \sum_{l=1}^{b_n} \Delta_1 w_l \left(\gamma_n(0) - n^{-1} \left(Z_1^2 + \ldots Z_{l-1}^2 + Z_{n-b_n+1+l}^2 + \cdots + Z_n^2 \right) \right) .$$

For $s \geq 0$, let

$$a_n(s) = n^{-1} \sum_{l=1}^{b_n - s} \Delta_1 w_{s+l} \left(\sum_{i=1}^{l-1} Z_i Z_{s+i} + \sum_{i=n-b_n+l+1}^{n-s} Z_i Z_{s+i} \right) .$$

Then from (3.24),

$$A = \gamma_n(0) - a_n(0) .$$

Now we can turn our attention to B,

$$B = 2n^{-1} \sum_{j=0}^{n-b_n} \sum_{k=1}^{b_n} \sum_{s=1}^{k-1} \sum_{l=1}^{k-s} \Delta_2 w_k Z_{j+l} Z_{j+l+s} .$$

Again, we will begin by permuting the order of the summations

$$\begin{aligned}
B &= 2n^{-1} \sum_{j=0}^{n-b_n} \sum_{k=1}^{b_n} \sum_{s=1}^{k-1} \sum_{l=1}^{k-s} \Delta_2 w_k Z_{j+l} Z_{j+l+s} \\
&= 2n^{-1} \sum_{j=0}^{n-b_n} \sum_{s=1}^{b_n-1} \sum_{k=s}^{b_n} \sum_{l=1}^{k-s} \Delta_2 w_k Z_{j+l} Z_{j+l+s} \\
&= 2n^{-1} \sum_{s=1}^{b_n-1} \sum_{j=0}^{n-b_n} \sum_{k=s}^{b_n} \sum_{l=1}^{k-s} \Delta_2 w_k Z_{j+l} Z_{j+l+s} \\
&= 2n^{-1} \sum_{s=1}^{b_n-1} \sum_{j=0}^{n-b_n} \sum_{l=1}^{b_n-s} Z_{j+l} Z_{j+l+s} \sum_{k=l+s}^{b_n} \Delta_2 w_k \\
&= 2n^{-1} \sum_{s=1}^{b_n-1} \sum_{j=0}^{n-b_n} \sum_{l=1}^{b_n-s} Z_{j+l} Z_{j+l+s} \Delta_1 w_{s+l} \\
&= 2n^{-1} \sum_{s=1}^{b_n-1} \sum_{l=1}^{b_n-s} \sum_{j=0}^{n-b_n} Z_{j+l} Z_{j+l+s} \Delta_1 w_{s+l} \\
&= 2 \sum_{s=1}^{b_n-1} \sum_{l=1}^{b_n-s} \Delta_1 w_{s+l} \left[\gamma_n(s) - n^{-1} \left(Z_l Z_{s+l} + \cdots + Z_{l-1} Z_{s+l-1} + \cdots + Z_{n-s} Z_n \right) \right] .
\end{aligned}$$

Therefore,

$$B = 2 \sum_{s=1}^{b_n-1} \left(\sum_{l=1}^{b_n-s} \Delta_1 w_{s+l} \right) \gamma_n(s) - 2 \sum_{s=1}^{b_n-1} a_n(s) .$$

Then by (3.23) we have

$$B = 2 \sum_{s=1}^{b_n-1} \gamma_n(s) w_n(s) - 2 \sum_{s=1}^{b_n-1} a_n(s) .$$

Let

$$d_n = a_n(0) + 2 \sum_{s=1}^{b_n-1} a_n(s) .$$

Then we have

$$\hat{\sigma}^2(n) = \gamma_n(0) + 2\sum_{s=1}^{b_n-1} \gamma_n(s)w_n(s) - d_n \ ,$$

and since the lag window is assumed to be even, we can write

$$\hat{\sigma}^2(n) = \gamma_n(0) + \sum_{s=-(b_n-1)}^{b_n-1} \gamma_n(s)w_n(s) - d_n \ ,$$

and using our previous notation, $\hat{\sigma}^2(n) = 2\pi f_n(0) - d_n$. □

Lemma 8. *Suppose* (3.6) *holds with* $\psi(n) = n^\alpha \log n$ *where* $\alpha = 1/(2+\delta)$ *and Assumptions 1 and 2 hold. If*

1. *a sampling plan* b_n *and a lag window* $w_n(\cdot)$ *exist such that*

$$b_n n^{2\alpha}(\log n)^3 \left(\sum_{k=1}^{b_n} |\Delta_2 w_n(k)|\right)^2 \to 0 \text{ as } n \to \infty \text{ and} \tag{3.26}$$

$$n^{2\alpha}(\log n)^2 \sum_{k=1}^{b_n} |\Delta_2 w_n(k)| \to 0 \text{ as } n \to \infty \ . \tag{3.27}$$

Then as $n \to \infty$, $\hat{\sigma}^2(n) - \sigma_g^2\tilde{\sigma}_*^2 \to 0$ *w.p.1.*

Proof. Notice

$$\hat{\sigma}^2(n) - \sigma_g^2\tilde{\sigma}_*^2 = n^{-1}\sum_{j=0}^{n-b_n}\sum_{k=1}^{b_n} k^2\Delta_2 w_n(k)\left(\left[\bar{Y}_j(k) - \bar{Y}_n\right]^2 - \sigma_g^2\left[\bar{B}_j(k) - \bar{B}_n\right]^2\right)$$

where $\bar{B}_j(k) = k^{-1}(B(j+k) - B(j))$ and $\bar{B}_n = n^{-1}B(n)$ are defined as before. If we

let

$$A = k\left[\bar{Y}_j(k) - \sigma_g \bar{B}_j(k)\right] = \left[\left(\sum_{i=1}^{j+k} Y_i - \sigma_g B(j+k)\right) - \left(\sum_{i=1}^{j} Y_i - \sigma_g B(j)\right)\right] ,$$

$$D = B(j+k) - B(j) ,$$

$$E = k\bar{B}_n \quad \text{and}$$

$$F = k\left[\bar{Y}_n - \sigma_g \bar{B}_n\right] ,$$

then

$$\begin{aligned} k(\bar{Y}_j(k) - \bar{Y}_n) &= k\left[\bar{Y}_j(k) - \bar{Y}_n \pm \sigma_g \bar{B}_j(k) \pm \sigma_g \bar{B}_n\right] \\ &= k\left[\bar{Y}_j(k) - \sigma_g \bar{B}_j(k)\right] + \sigma_g\left[B(j+k) - B(j)\right] \\ &\quad - \sigma_g k\bar{B}_n - k\left[\bar{Y}_n - \sigma_g \bar{B}_n\right] \\ &= A + \sigma_g(D-E) - F . \end{aligned}$$

We can then rewrite

$$\begin{aligned} \left|\hat{\sigma}^2(n) - \sigma_g^2\tilde{\sigma}_*^2\right| &\le n^{-1}\sum_{j=0}^{n-b_n}\sum_{k=1}^{b_n}\left|\Delta_2 w_n(k)\left[(A+\sigma_g(D-E)-F)^2 - \sigma_g^2(D-E)^2\right]\right| \\ &\le n^{-1}\sum_{j=0}^{n-b_n}\sum_{k=1}^{b_n}|\Delta_2 w_n(k)|\left[A^2 + F^2 + 2\sigma_g|AD| + 2\sigma_g|AE|\right. \\ &\quad \left. + 2\sigma_g|AF| + 2\sigma_g|DF| + 2\sigma_g|EF|\right] . \end{aligned} \tag{3.28}$$

It suffices to show the 7 terms in (3.28) tend to 0 as $n \to \infty$. To this end, we will use the LIL type results on the increments of Brownian motion from Appendix A.1. Our assumptions say there exists a strong invariance principle, or there exists a constant

C such that for all large n

$$\left| \sum_{i=1}^{n} g(X_i) - nE_\pi g - \sigma_g B(n) \right| \leq Cn^\alpha \log n \tag{3.29}$$

where $\alpha = 1/(2+\delta)$.

1. From (3.29), we can see that

$$|A| \leq C(j+k)^\alpha \log(j+k) + C(j)^\alpha \log(j) \leq 2Cn^\alpha \log n \tag{3.30}$$

resulting in

$$n^{-1} \sum_{j=0}^{n-b_n} \sum_{k=1}^{b_n} |\Delta_2 w_n(k)| \, A^2 \leq 4C^2 n^{2\alpha} (\log n)^2 \sum_{k=1}^{b_n} |\Delta_2 w_n(k)| \to 0$$

as $n \to \infty$ by (3.27).

2. From (3.29) and the fact that $k \leq b_n \leq n$

$$|F| \leq Ckn^{\alpha-1} \log n \, , \tag{3.31}$$

resulting in

$$\begin{aligned} n^{-1} \sum_{j=0}^{n-b_n} \sum_{k=1}^{b_n} |\Delta_2 w_n(k)| \, F^2 &\leq C^2 n^{2\alpha-2} (\log n)^2 \sum_{k=1}^{b_n} k^2 \, |\Delta_2 w_n(k)| \\ &\leq C^2 b_n^2 n^{2\alpha-2} (\log n)^2 \sum_{k=1}^{b_n} |\Delta_2 w_n(k)| \to 0 \end{aligned}$$

as $n \to \infty$ by (3.27).

3. From Lemma 16

$$\begin{aligned}|D| &= |B(j+k) - B(j)| \\ &\leq \sup_{0\leq t\leq n-b_n} \sup_{0\leq s\leq b_n} |B(t+s) - B(t)| \\ &\leq (1+\epsilon)\left(2b_n\left(\log\frac{n}{b_n} + \log\log n\right)\right)^{1/2} \\ &\leq 2(1+\epsilon)b_n^{1/2}\left(\log n\right)^{1/2} \ . \end{aligned} \tag{3.32}$$

Combining this with (3.30), we can see

$$\begin{aligned}&n^{-1}\sum_{j=0}^{n-b_n}\sum_{k=1}^{b_n} |\Delta_2 w_n(k)|\, 2\sigma_g\, |AD| \\ &\qquad\leq 8C\sigma_g(1+\epsilon)b_n^{1/2}n^{\alpha}\left(\log n\right)^{3/2}\sum_{k=1}^{b_n}|\Delta_2 w_n(k)| \to 0\end{aligned}$$

as $n \to \infty$ by (3.26).

4. From Lemma 15

$$|E| \leq \sqrt{2}(1+\epsilon)kn^{-1/2}(\log\log n)^{1/2} \ . \tag{3.33}$$

Combining this with (3.30), we can see

$$\begin{aligned}&n^{-1}\sum_{j=0}^{n-b_n}\sum_{k=1}^{b_n} |\Delta_2 w_n(k)|\, 2\sigma_g\, |AE| \\ &\qquad\leq 2^{5/2}C\sigma_g(1+\epsilon)n^{\alpha-1/2}\log n(\log\log n)^{1/2}\sum_{k=1}^{b_n} k\,|\Delta_2 w_n(k)| \to 0\end{aligned}$$

as $n \to \infty$ by (3.26).

5. From (3.30) and (3.31) we can write

$$n^{-1}\sum_{j=0}^{n-b_n}\sum_{k=1}^{b_n}|\Delta_2 w_n(k)|\,2\,|AF| \leq 4C^2 n^{2\alpha-1}(\log n)^2\sum_{k=1}^{b_n}k\,|\Delta_2 w_n(k)| \to 0$$

as $n \to \infty$ by (3.27).

6. From (3.32) and (3.31) we can write

$$n^{-1}\sum_{j=0}^{n-b_n}\sum_{k=1}^{b_n}|\Delta_2 w_n(k)|\,2\sigma_g\,|DF| \leq 4\sigma_g(1+\epsilon)b_n^{1/2}n^{\alpha-1}\,(\log n)^{3/2}\sum_{k=1}^{b_n}k\,|\Delta_2 w_n(k)| \to 0$$

as $n \to \infty$ by (3.26).

7. From (3.33) and (3.31) we can write

$$n^{-1}\sum_{j=0}^{n-b_n}\sum_{k=1}^{b_n}|\Delta_2 w_n(k)|\,2\sigma_g\,|EF| \leq 2^{3/2}C\sigma_g(1+\epsilon)n^{\alpha-3/2}\log n(\log\log n)^{1/2}\sum_{k=1}^{b_n}k^2\,|\Delta_2 w_n(k)| \to 0$$

as $n \to \infty$ by (3.26).

Hence, all the terms in (3.28) tend to 0 as $n \to \infty$ and the lemma is proved. □

Lemma 9. *Let X be a geometrically ergodic Markov chain with invariant distribution π and $g : \mathsf{X} \to \mathbb{R}$ be a Borel function with $E_\pi|g|^{2+\delta+\epsilon} < \infty$ for some $\delta \geq 2$ and some $\epsilon > 0$. If Assumption 2 holds, $b_n^{-1}\log n$ stays bounded as $n \to \infty$, and $b_n^{-1}n^{2\alpha}\log n \to 0$*

as $n \to \infty$ where $\alpha = 1/(2+\delta)$, then $b_n^{-1} \sum_{i=1}^{b_n} Y_i^2$ and $b_n^{-1} \sum_{i=n-b_n+1}^{n} Y_i^2$ stay bounded as $n \to \infty$ w.p.1.

Proof. First note that if $E_\pi |g|^2 < \infty$, then $b_n^{-1} \sum_{i=1}^{b_n} Y_i^2$ stays bounded as $n \to \infty$ w.p.1 from the Ergodic Theorem.

Consider the sequence $\{Y_i^2 : i \geq 1\}$. The stated assumptions imply a Markov chain CLT holds where $\sigma_g' < \infty$ is defined as the resulting asymptotic standard deviation. Lemma 4 implies that

$$\left| \sum_{i=1}^{n} Y_i^2 - n E_\pi Y_1^2 - \sigma_g' B(n) \right| < C'(\omega) n^{2\alpha} \log n$$

for all $n > n_0$.

Then

$$\begin{aligned}
b_n^{-1} \left| \sum_{i=n-b_n+1}^{n} Y_i^2 \right| &= b_n^{-1} \left| \sum_{i=1}^{n} Y_i^2 - \sum_{i=1}^{n-b_n} Y_i^2 \right| \\
&= b_n^{-1} \left| \left(\sum_{i=1}^{n} Y_i^2 - n E_\pi Y_1^2 - \sigma_g' B(n) \right) \right. \\
&\quad - \left(\sum_{i=1}^{n-b_n} Y_i^2 - (n-b_n) E_\pi Y_1^2 - \sigma_g' B(n-b_n) \right) \\
&\quad \left. + \sigma_g' \left(B(n) - B(n-b_n) \right) + b_n E_\pi Y_1^2 \right| \\
&\leq b_n^{-1} \left(2C'(\omega) n^{\alpha'} \log n + (1+\epsilon) \left(2 b_n \left(\log \frac{n}{b_n} + \log \log n \right) \right)^{1/2} + b_n E_\pi Y_1^2 \right) \\
&= E_\pi Y_1^2 + 2C'(\omega) b_n^{-1} n^{\alpha'} \log n + O\left((b_n^{-1} \log n)^{1/2} \right) \quad w.p.1.
\end{aligned}$$

Hence, $b_n^{-1} \left| \sum_{i=n-b_n+1}^{n} Y_i^2 \right|$ stays bounded w.p.1 since $b_n^{-1} n^{\alpha'} \log n \to 0$ as $n \to \infty$. □

Lemma 10. *(Damerdji, 1991, Proposition 4.3) Suppose Assumptions 1 and 2 hold.*

1.

$$b_n n^{-1} \sum_{k=1}^{b_n} k \left|\Delta_1 w_n(k)\right| \to 0 \text{ as } n \to \infty ;$$

2. and $b_n^{-1} \sum_{i=1}^{b_n} Y_i^2$ *and* $b_n^{-1} \sum_{i=n-b_n+1}^{n} Y_i^2$ *stay bounded as* $n \to \infty$ *w.p.1.*

Then $d_n \to 0$ *as* $n \to \infty$ *w.p.1.*

Proof. Damerdji (1991) shows a proof for general processes which is expanded here for clarity. We will show $d_n \to 0$ as $n \to \infty$ w.p.1 in three steps. Recall

$$\begin{aligned} d_n = n^{-1} \Bigg(& \left[\sum_{l=1}^{b_n} \Delta w_n(l) \left(\sum_{i=1}^{l-1} Z_i^2 + \sum_{i=n-b_n+l+1}^{n} Z_i^2 \right) \right] \\ & + 2 \sum_{s=1}^{b_n-1} \left[\sum_{l=1}^{b_n-s} \Delta w_n(s+l) \left(\sum_{i=1}^{l-1} Z_i Z_{s+i} + \sum_{i=n-b_n+l+1}^{n-s} Z_i Z_{s+i} \right) \right] \Bigg) , \end{aligned} \tag{3.34}$$

where any empty sums are defined to be zero and $Z_i = Y_i - \bar{Y}_n$ for all $i = 1, 2, \ldots, n$.

1. Let us put bounds on $|d_n|$ such that for $l = 1, \ldots, b_n$

$$\sum_{i=1}^{l-1} Z_i^2 \le \sum_{i=1}^{b_n} Z_i^2 \quad \text{and} \quad \sum_{i=n-b_n+l+1}^{n} Z_i^2 \le \sum_{i=n-b_n+1}^{n} Z_i^2 .$$

Then bound the sums in the cross terms using the inequality $|ab| \le (a^2 + b^2)/2$. So, for $s = 1, \ldots, b_n$ and $l = 1, \ldots, b_n - s$ we have

$$\sum_{i=1}^{l-1} |Z_i Z_{s+i}| \le \frac{1}{2} \sum_{i=1}^{l-1} \left(Z_i^2 + Z_{s+i}^2 \right) \le \sum_{i=1}^{b_n} Z_i^2$$

and

$$\sum_{i=n-b_n+l+1}^{n-s} |Z_i Z_{s+i}| \le \sum_{i=n-b_n+1}^{n} Z_i^2 .$$

Then $|d_n|$ can be bounded by

$$\begin{aligned}
|d_n| &\leq \left(\sum_{i=1}^{b_n} Z_i^2 + \sum_{i=n-b_n+1}^{n} Z_i^2\right) n^{-1} \left(\sum_{l=1}^{b_n} |\Delta w_n(l)| + 2\sum_{s=1}^{b_n-1}\sum_{l=1}^{b_n-s} |\Delta w_n(s+l)|\right) \\
&= b_n^{-1}\left(\sum_{i=1}^{b_n} Z_i^2 + \sum_{i=n-b_n+1}^{n} Z_i^2\right) \\
&\quad \times n^{-1} b_n \left(\sum_{l=1}^{b_n} |\Delta w_n(l)| + 2\sum_{s=1}^{b_n-1}\sum_{l=1}^{b_n-s} |\Delta w_n(s+l)|\right) \\
&= b_n^{-1}\left(\sum_{i=1}^{b_n} Z_i^2 + \sum_{i=n-b_n+1}^{n} Z_i^2\right) \qquad (3.35) \\
&\quad \times n^{-1} b_n \left(\sum_{l=1}^{b_n} |\Delta w_n(l)| + 2\sum_{k=1}^{b_n} k\,|\Delta w_n(k)|\right) . \qquad (3.36)
\end{aligned}$$

Then we can see that (3.36) tends to 0 as $n \to \infty$ by condition 1. We will show in the next two steps that (3.35) is bounded implying $d_n \to 0$ as $n \to \infty$ w.p.1.

2. Here, we will use the fact that $b_n^{-1}\sum_{i=1}^{b_n} Y_i^2$ and $b_n^{-1}\sum_{i=n-b_n+1}^{n} Y_i^2$ stay bounded as $n \to \infty$ w.p.1.

$$b_n^{-1}\sum_{i=1}^{b_n} Z_i^2 = b_n^{-1}\sum_{i=1}^{b_n} Y_i^2 - 2\bar{Y}_n b_n^{-1}\sum_{i=1}^{b_n} Y_i + \left(\bar{Y}_n\right)^2 .$$

Then the first term is bounded and by the Cauchy-Schwartz inequality we have

$$\left(\sum_{i=1}^{b_n} Y_i\right)^2 \leq \sum_{i=1}^{b_n} Y_i^2$$

which implies $\sum_{i=1}^{b_n} Y_i$ is bounded resulting in a bound for the second term. Finally, the third term is bounded by the assumption of a finite mean throughout. Therefore, $b_n^{-1}\sum_{i=1}^{b_n} Z_i^2$ stays bounded for large n w.p.1 and hence the expression in d_n coming from the starting observations goes to 0.

3. Similarly, we can show $b_n^{-1} \sum_{i=n-b_n+1}^{n} Z_i^2$ stays bounded for large n w.p.1.

Thus, $d_n \to 0$ as $n \to \infty$ w.p.1. □

3.4.3 Results for Proof of Proposition 2

Lemma 11. *(Jones et al., 2006, Lemma 8) Suppose* (3.6) *holds with* $\psi(n) = n^\alpha \log n$ *where* $\alpha = 1/(2+\delta)$ *and Assumption 2 holds. If (i)* $b_n^{-1} n^{2\alpha} [\log n]^3 \to 0$ *as* $n \to \infty$ *then* $\hat{\sigma}^2_{BM} - \sigma_g^2 \tilde{\sigma}^2_{BM} \to 0$ *as* $n \to \infty$ *w.p.1.*

Proof. Recall that $X = \{X_1, X_2, \dots\}$ is a Harris ergodic Markov chain. Define the process Y by $Y_i = g(X_i) - E_\pi g$ for $i = 1, 2, 3, \dots$. Then

$$\hat{\sigma}^2_{BM} = \frac{b_n}{a_n - 1} \sum_{k=0}^{a-1} (\bar{Y}_k - \bar{Y}_n)^2$$

where $\bar{Y}_k := \frac{1}{b_n} \sum_{i=1}^{b_n} g(X_{kb_n+i})$ for $k = 0, \dots, a-1$ and $\bar{Y}_n = n^{-1} \sum_{i=1}^{n} Y_i$. Since

$$\bar{Y}_k - \bar{Y}_n = \bar{Y}_k - \bar{Y}_n \pm \sigma_g \bar{B}_k \pm \sigma_g \bar{B}_n$$

we have

$$\begin{aligned}
\left|\hat{\sigma}^2_{BM} - \sigma_g^2 \tilde{\sigma}^2_{BM}\right| \le \frac{b_n}{a_n - 1} \sum_{k=0}^{a-1} \Big[& \left(\bar{Y}_k - \sigma_g \bar{B}_k\right)^2 + \left(\bar{Y}_n - \sigma_g \bar{B}_n\right)^2 \\
& + \left|2 \left(\bar{Y}_k - \sigma_g \bar{B}_k\right) \left(\bar{Y}_n - \sigma_g \bar{B}_n\right)\right| + \left|2\sigma_g \left(\bar{Y}_k - \sigma_g \bar{B}_k\right) \bar{B}_k\right| \\
& + \left|2\sigma_g \left(\bar{Y}_k - \sigma_g \bar{B}_k\right) \bar{B}_n\right| + \left|2\sigma_g \left(\bar{Y}_n - \sigma_g \bar{B}_n\right) \bar{B}_k\right| \\
& + \left|2\sigma_g \left(\bar{Y}_n - \sigma_g \bar{B}_n\right) \bar{B}_n\right| \Big] \ .
\end{aligned}$$

Now we will consider each term in the sum and show that it tends to 0.

1. Our assumptions say that there exists a constant C such that for all large n

$$\left|\sum_{i=1}^{n} g(X_i) - nE_\pi g - \sigma_g B(n)\right| < Cn^\alpha \log n \quad w.p.1. \tag{3.37}$$

Note that

$$\bar{Y}_k - \sigma_g \bar{B}_k = \frac{1}{b_n}\left[\sum_{i=1}^{(k+1)b_n} Y_i - \sigma_g B((k+1)b_n)\right] - \frac{1}{b_n}\left[\sum_{i=1}^{kb_n} Y_i - \sigma_g B(kb_n)\right]$$

and hence by (3.37)

$$\begin{aligned}\left|\bar{Y}_k - \sigma_g \bar{B}_k\right| &\le \frac{1}{b_n}\left[\left|\sum_{i=1}^{(k+1)b_n} Y_i - \sigma_g B((k+1)b_n)\right| + \left|\sum_{i=1}^{kb_n} Y_i - \sigma_g B(kb_n)\right|\right] \\ &< \frac{2}{b_n} C n^\alpha \log n \end{aligned} \tag{3.38}$$

Then

$$\frac{b_n}{a_n - 1}\sum_{k=0}^{a-1}\left(\bar{Y}_k - \sigma_g \bar{B}_k\right)^2 < 4C^2 \frac{a_n}{a_n - 1} b_n^{-1} n^{2\alpha} (\log n)^2 \to 0$$

as $n \to \infty$ since Assumption 2 implies $a_n \to \infty$ as $n \to \infty$ and (i).

2. Apply (3.37) to obtain

$$\left|\bar{Y}_n - \sigma_g \bar{B}_n\right| = n^{-1}\left|\sum_{i=1}^{n} Y_i - \sigma_g B(n)\right| < Cn^{\alpha-1}\log n\ . \tag{3.39}$$

Then

$$\frac{b_n}{a_n - 1}\sum_{k=0}^{a-1}\left(\bar{Y}_n - \sigma_g \bar{B}_n\right)^2 < C^2 \frac{a_n}{a_n - 1}\frac{b_n}{n}\frac{(\log n)^2}{n^{1-2\alpha}} \to 0$$

as $n \to \infty$ by Assumption 2 and since $1 - 2\alpha > 0$.

3. By (3.38) and (3.39)

$$\left|2\left(\bar{Y}_k - \sigma_g \bar{B}_k\right)\left(\bar{Y}_n - \sigma_g \bar{B}_n\right)\right| < 4C^2 b_n^{-1} n^{2\alpha-1} (\log n)^2 \, .$$

Thus

$$\frac{b_n}{a_n - 1} \sum_{k=0}^{a-1} \left|2\left(\bar{Y}_k - \sigma_g \bar{B}_k\right)\left(\bar{Y}_n - \sigma_g \bar{B}_n\right)\right| < 4C^2 \frac{a_n}{a_n - 1} \frac{(\log n)^2}{n^{1-2\alpha}} \to 0$$

as $n \to \infty$ by Assumption 2 and since $1 - 2\alpha > 0$.

4. Since $b_n \geq 2$, (A.2) and (3.38) together imply

$$\begin{aligned}\left|\left(\bar{Y}_k - \sigma_g \bar{B}_k\right)\bar{B}_k\right| < 2^{3/2} C(1+\epsilon) b_n^{-1} \big[& b_n^{-1} n^{2\alpha} (\log n)^2 \log(n/b_n) \\ & + b_n^{-1} n^{2\alpha} (\log n)^2 \log\log n\big]^{1/2} \, .\end{aligned}$$

Hence

$$\begin{aligned}&\frac{b_n}{a_n - 1} \sum_{k=0}^{a-1} \left|2\sigma_g \left(\bar{Y}_k - \sigma_g \bar{B}_k\right)\bar{B}_k\right| \leq \\ &8\sigma_g C(1+\epsilon) \frac{a_n}{a_n - 1} \left[b_n^{-1} n^{2\alpha} (\log n)^2 \log(n/b_n) + b_n^{-1} n^{2\alpha} (\log n)^2 \log\log n\right]^{1/2}\end{aligned}$$

which tends to 0 as $n \to \infty$ by Assumption 2 and (i).

5. By (3.38) and (A.1)

$$\left|\left(\bar{Y}_k - \sigma_g \bar{B}_k\right)\bar{B}_n\right| < 4C(1+\epsilon) b_n^{-1} n^{-1/2+\alpha} (\log n)(\log\log n)^{1/2}$$

so that

$$\frac{b_n}{a_n - 1} \sum_{k=0}^{a-1} \left| 2\sigma_g \left(\bar{Y}_k - \sigma_g \bar{B}_k \right) \bar{B}_n \right| <$$
$$8\sigma_g C(1+\epsilon) \frac{a_n}{a_n - 1} \frac{(\log n)(\log\log n)^{1/2}}{n^{1/2-\alpha}} \to 0$$

as $n \to \infty$ by Assumption 2 and since $1/2 - \alpha > 0$.

6. Use (A.2) and (3.39) to get

$$\left| \left(\bar{Y}_n - \sigma_g \bar{B}_n \right) \bar{B}_k \right| < \sqrt{2} C(1+\epsilon) \frac{n^{\alpha-1} \log n}{b_n^{1/2}} \left[\log(n/b_n) + \log\log n \right]^{1/2}$$

and hence using Assumption 2 and (i) shows that as $n \to \infty$

$$\frac{b_n}{a_n - 1} \sum_{k=0}^{a-1} \left| 2\sigma_g \left(\bar{Y}_n - \sigma_g \bar{B}_n \right) \bar{B}_k \right| <$$
$$4\sigma_g C(1+\epsilon) \frac{a_n}{a_n - 1} \frac{b_n}{n} \left[b_n^{-1} n^{2\alpha} ((\log n)^2 \log(n/b_n) + (\log n)^2 \log\log n) \right]^{1/2} \to 0 \, .$$

7. Now (A.1) and (3.39) imply

$$\left| \left(\bar{Y}_n - \sigma_g \bar{B}_n \right) \bar{B}_n \right| < 2C(1+\epsilon) n^{-3/2+\alpha} (\log n)^{3/2} \, .$$

Hence

$$\frac{b_n}{a_n - 1} \sum_{k=0}^{a-1} \left| 2\sigma_g \left(\bar{Y}_n - \sigma_g \bar{B}_n \right) \bar{B}_n \right| < 4\sigma_g C(1+\epsilon) \frac{a_n}{a_n - 1} \frac{b_n}{n} \frac{(\log n)^{3/2}}{n^{1/2-\alpha}} \to 0$$

as $n \to \infty$ by Assumption 2 and since $1/2 - \alpha > 0$.

□

3.4.4 Results for Mean-Square Consistency

Preliminary Results

We require the *Generalized Dominated Convergence Theorem.*

Lemma 12. *(Zeidler, 1990, p. 1015) Suppose*

1. $||f_n(x)|| \leq g_n(x)$ *for almost all* $x \in M$ *and all* $n \in \mathbb{N}$ *where all the functions* $g_n, g : M \to \mathbb{R}$ *are integrable and we have the convergence* $g_n \to g$ *almost everywhere on* M *as* $n \to \infty$ *along with*

$$\int_M g_n dx \to \int_M g dx \text{ as } n \to \infty ,$$

2. $\lim_{n\to\infty} f_n(x)$ *exists for almost all* $x \in M$*, where* $f_n : M \subseteq \mathbb{R}^N \to Y$ *is measureable for all* n*.*

Then we have

$$\lim_{n\to\infty} \int_M f_n dx = \int_M \lim_{n\to\infty} f_n(x) dx .$$

Lemma 13. *(Jones et al., 2006, p. 1545-1546) Suppose* (3.6) *holds with* $\psi(n) = n^\alpha \log n$ *where* $\alpha = 1/(2+\delta)$ *and Assumption 2 holds. Then as* $n \to \infty$

$$\begin{aligned}
\left|\hat{\sigma}^2_{BM} - \sigma_g^2\tilde{\sigma}^2_{BM}\right| \leq\ & 4C^2\frac{a_n}{a_n-1}b_n^{-1}n^{2\alpha}(\log n)^2 + C^2\frac{a_n}{a_n-1}\frac{b_n}{n}\frac{(\log n)^2}{n^{1-2\alpha}} \\
& + 4C^2\frac{a_n}{a_n-1}\frac{(\log n)^2}{n^{1-2\alpha}} + 8\sigma_g C(1+\epsilon)\frac{a_n}{a_n-1} \\
& \times \left[b_n^{-1}n^{2\alpha}(\log n)^2\log(n/b_n) + b_n^{-1}n^{2\alpha}(\log n)^2\log\log n\right]^{1/2} \\
& + 8\sigma_g C(1+\epsilon)\frac{a_n}{a_n-1}\frac{(\log n)(\log\log n)^{1/2}}{n^{1/2-\alpha}} \\
& + 4\sigma_g C(1+\epsilon)\frac{a_n}{a_n-1}\frac{b_n}{n} \\
& \times \left[b_n^{-1}n^{2\alpha}((\log n)^2\log(n/b_n) + (\log n)^2\log\log n)\right]^{1/2} \\
& + 4\sigma_g C(1+\epsilon)\frac{a_n}{a_n-1}\frac{b_n}{n}\frac{(\log n)^{3/2}}{n^{1/2-\alpha}} \qquad (3.40)
\end{aligned}$$

w.p.1.

Remark 12. Notice that if $b_n^{-1}n^{2\alpha}[\log n]^3 \to 0$ as $n \to \infty$, $\left|\hat{\sigma}^2_{BM} - \sigma_g^2\tilde{\sigma}^2_{BM}\right| \to 0$ as $n \to \infty$.

Lemma 14. *Suppose* (3.6) *holds with* $\psi(n) = n^\alpha \log n$ *where* $\alpha = 1/(2+\delta)$*, Assumption 2 holds, and let* $g : \mathsf{X} \to \mathbb{R}$ *be a Borel function with* $E_\pi g^4 < \infty$*. Suppose* $\hat{\sigma}^2$ *is* $\hat{\sigma}^2_{BM}$ *or* $\hat{\sigma}^2_{OBM}$*.*

1. *If* $b_n^{-1}n^{2\alpha}[\log n]^3 \to 0$ *as* $n \to \infty$*, then*

$$\lim_{n\to\infty} E\left[\left|\hat{\sigma}^2 - \sigma_g^2\tilde{\sigma}^2\right|\right] = 0 \text{ if } E_\pi C^2 < \infty \text{ and}$$

$$\lim_{n\to\infty} E\left[\left(\hat{\sigma}^2 - \sigma_g^2\tilde{\sigma}^2\right)^2\right] = 0 \text{ if } E_\pi C^4 < \infty\,.$$

2. If $b_n^{-1} n^{1/2+\alpha} [\log n]^{3/2} \to 0$ *as* $n \to \infty$*, then*

$$\lim_{n\to\infty} E\left[\frac{n}{b_n}\left(\hat{\sigma}^2 - \sigma_g^2\tilde{\sigma}^2\right)^2\right] = 0 \text{ if } E_\pi C^4 < \infty \,.$$

Proof. We will only prove the first claim for BM as the proof of the others are similar. From Lemma 13 there exists an integer N_0 and functions g_1 and g_2 such that

$$\begin{aligned}
|\hat{\sigma}^2_{BM} - \sigma_g^2\tilde{\sigma}^2_{BM}| &= |\hat{\sigma}^2_{BM} - \sigma_g^2\tilde{\sigma}^2_{BM}| I(0 \le n \le N_0) + |\hat{\sigma}^2_{BM} - \sigma_g^2\tilde{\sigma}^2_{BM}| I(N_0 < n) \\
&\le |\hat{\sigma}^2_{BM} - \sigma_g^2\tilde{\sigma}^2_{BM}| I(0 \le n \le N_0) + [C^2 g_1(n) + C g_2(n)] I(N_0 < n) \\
&:= g_n(X_1, \ldots X_n, B(0), \ldots, B(n)) \,.
\end{aligned}$$

Now

$$\begin{aligned}
E g_n(X_1, \ldots X_n, B(0), \ldots, B(n)) = I(0 \le n \le N_0) E|\hat{\sigma}^2_{BM} - \sigma_g^2\tilde{\sigma}^2_{BM}| \\
+ [g_1(n)E(C^2) + g_2(n)E(C)] I(N_0 < n)
\end{aligned}$$

and since Lemma 23 and our assumptions on the moments of g imply

$$E|\hat{\sigma}^2_{BM} - \sigma_g^2\tilde{\sigma}^2_{BM}| \le E\hat{\sigma}^2_{BM} + \sigma_g^2 E\tilde{\sigma}^2_{BM} = E\hat{\sigma}^2_{BM} + \sigma_g^2 < \infty$$

it follows from our assumptions on the moments of C that $E|g_n| < \infty$. Next observe that as $n \to \infty$, we have $g_n \to 0$ w.p.1 and $E g_n \to 0$ by Lemma 13. From Lemma 11 we have that $|\hat{\sigma}^2_{BM} - \sigma_g^2\tilde{\sigma}^2_{BM}| \to 0$ w.p.1 as $n \to \infty$. An application of the Generalized Dominated Convergence Theorem implies, as $n \to \infty$,

$$E\left[\left|\hat{\sigma}^2_{BM} - \sigma_g^2\tilde{\sigma}^2_{BM}\right|\right] \to 0 \,.$$

The second result for BM is similar if (3.40) is multiplied through by $(n/b_n)^{1/2}$.

Then note that $b_n^{-1} n^{1/2+\alpha} (\log n)^{3/2} \to 0$ as $n \to \infty$ implies $b_n^{-3/2} n^{1/2+2\alpha} (\log n)^3 \to 0$ as $n \to \infty$.

For OBM, the results follow from Lemma 8 with the modified Bartlett lag window and the Generalized Dominated Convergence Theorem. □

Proof of Theorem 7

Damerdji (1995) shows a proof for mean-square consistency assuming (3.6) holds with $\gamma(n) = n^{1/2-\alpha'}$ where $\alpha' \leq \delta'/(24 + 12\delta')$. However, this result incorrectly simplifies results similar to Lemma 8 and Lemma 13 by only addressing the slowest converging term. This leads to a necessary condition of $E_\pi C^2 < \infty$ rather than $E_\pi C^4 < \infty$. In addition, the result incorrectly proves the result in Appendix A of Damerdji (1995) which we replace with the Generalized Dominated Convergence Theorem. Both of these issues are corrected in the statement in the proof.

Here we show the proof for geometrically ergodic chains. For ease of exposition, suppose $\tilde{\sigma}^2$ is the appropriate Brownian motion variance estimate defined in Appendix A.1.4. Both $\hat{\sigma}^2$ and $\tilde{\sigma}^2$ are still dependent on n, but we have suppressed this dependence.

Recall, the MSE of the estimator $\hat{\sigma}^2$ of σ_g^2 is given by $\text{MSE}(\hat{\sigma}^2) := E_{\sigma_g^2}[(\hat{\sigma}^2 - \sigma_g^2)^2]$. In this case, σ_g^2 is a parameter of π, so we could think of this expectation as $\text{MSE}(\hat{\sigma}^2) = E_\pi[(\hat{\sigma}^2 - \sigma_g^2)^2]$. Throughout this proof, we will suppress this dependency, but the results will only apply to stationary processes because of this dependency on π.

1. First, we will show $\lim_{n\to\infty} \frac{n}{b_n}\mathrm{Var}(\hat{\sigma}^2) = \frac{4}{3}\sigma_g^4$ as $n \to \infty$. First, one can write

$$\begin{aligned}
\mathrm{Var}(\hat{\sigma}^2) &= E\left[\left((\hat{\sigma}^2 - \sigma_g^2\tilde{\sigma}^2) + \sigma_g^2(\tilde{\sigma}^2 - E\tilde{\sigma}^2) - (E\hat{\sigma}^2 - \sigma_g^2 E\tilde{\sigma}^2)\right)^2\right] \\
&= \sigma_g^4 E\left[(\tilde{\sigma}^2 - E\tilde{\sigma}^2)^2\right] + E\left[\left((\hat{\sigma}^2 - \sigma_g^2\tilde{\sigma}^2) - E(\hat{\sigma}^2 - \sigma_g^2\tilde{\sigma}^2)\right)^2\right] \\
&\quad + 2\sigma_g^2 E\left[(\tilde{\sigma}^2 - E\tilde{\sigma}^2)\left((\hat{\sigma}^2 - \sigma_g^2\tilde{\sigma}^2) - E(\hat{\sigma}^2 - \sigma_g^2\tilde{\sigma}^2)\right)\right] \\
&= \sigma_g^4 \mathrm{Var}(\tilde{\sigma}^2) + \eta\,, \qquad (3.41)
\end{aligned}$$

   where

$$\eta = \mathrm{Var}\left(\hat{\sigma}^2 - \sigma_g^2\tilde{\sigma}^2\right) + 2\sigma_g^2 E\left[(\tilde{\sigma}^2 - E\tilde{\sigma}^2)(\hat{\sigma}^2 - \sigma_g^2\tilde{\sigma}^2)\right]\,.$$

   Using Lemma 23, (3.41) reduces to

$$\mathrm{Var}(\hat{\sigma}^2) = c\sigma_g^4\frac{b_n}{n} + o\left(\frac{b_n}{n}\right) + \eta\,,$$

   where $c = 2$ for BM and $c = 4/3$ for OBM.

   To finish the proof of this part of the theorem, we will show that $\eta \to 0$ as $n \to \infty$. We can simplify the expression of η using the Cauchy-Schwarz inequality and the fact that $\mathrm{Var}(X) \le EX^2$,

$$\begin{aligned}
|\eta| &= \left|\mathrm{Var}\left(\hat{\sigma}^2 - \sigma_g^2\tilde{\sigma}^2\right) + 2\sigma_g^2 E\left[(\tilde{\sigma}^2 - E\tilde{\sigma}^2)(\hat{\sigma}^2 - \sigma_g^2\tilde{\sigma}^2)\right]\right| \\
&\le E\left[\left(\hat{\sigma}^2 - \sigma_g^2\tilde{\sigma}^2\right)^2\right] + 2\sigma_g^2\left(E\left[\left(\hat{\sigma}^2 - \sigma_g^2\tilde{\sigma}^2\right)^2\right]\mathrm{Var}\left(\tilde{\sigma}^2\right)\right)^{1/2}\,,
\end{aligned}$$

   which tends to zero as a result of Proposition 14 and (A.12).

2. Next, we will show $\mathrm{Bias}(\hat{\sigma}^2) \to 0$ as $n \to \infty$.

Lemma 23 shows that $E[\tilde{\sigma}^2] = 1$. We can use this fact with Lemma 14 to get

$$\begin{aligned}\text{Bias}(\hat{\sigma}^2) &= E(\hat{\sigma}^2) - \sigma_g^2 \\ &\leq E\left[\left|\hat{\sigma}^2 - \sigma_g^2\tilde{\sigma}^2\right|\right] ,\end{aligned}$$

which tends to zero as $n \to \infty$.

3. Clearly, $\text{MSE}(\hat{\sigma}^2) \to 0$ as $n \to \infty$ from the first two parts of the theorem.

Proof of Theorem 8

Recall $\tilde{\sigma}^2$ is defined in Appendix A.1.4. Using (A.12) from Lemma 23, (3.41) reduces to

$$\frac{n}{b_n}\text{Var}(\hat{\sigma}^2) = c\sigma_g^4 + o(1) + \frac{n}{b_n}\eta ,$$

where $c = 2$ for BM and $c = 4/3$ for OBM.

To finish the proof, we will show that $\frac{n}{b_n}\eta \to 0$ as $n \to \infty$. Again, we can simplify the expression using the Cauchy-Schwarz inequality and the fact that $\text{Var}(X) \leq EX^2$,

$$\begin{aligned}\frac{n}{b_n}|\eta| &= \frac{n}{b_n}\left|\text{Var}\left(\hat{\sigma}^2 - \sigma_g^2\tilde{\sigma}^2\right) + 2\sigma_g^2 E\left[(\tilde{\sigma}^2 - E\tilde{\sigma}^2)(\hat{\sigma}^2 - \sigma_g^2\tilde{\sigma}^2)\right]\right| \\ &\leq E\left[\frac{n}{b_n}\left(\hat{\sigma}^2 - \sigma_g^2\tilde{\sigma}^2\right)^2\right] + 2\sigma_g^2\left(E\left[\frac{n}{b_n}\left(\hat{\sigma}^2 - \sigma_g^2\tilde{\sigma}^2\right)^2\right]\frac{n}{b_n}\text{Var}\left(\tilde{\sigma}^2\right)\right)^{1/2} ,\end{aligned}$$

which tends to zero as a result of Lemma 14 and (A.12).

Chapter 4

Subsampling

We have seen several approaches for dealing with ergodic averages and their corresponding standard errors. Very little formal attention has been given to quantities that cannot be expressed as an ergodic average. This is somewhat surprising considering empirical quantiles of the posterior distribution are commonly reported in MCMC literature. For this reason, we introduce subsampling (Politis et al., 1999) for time–series data and illustrate its use in two examples.

4.1 Introduction

Suppose we want to find the value of some functional $\theta := \theta_\pi$ where π is a probability distribution with support X. Our discussion to this point only applies when $\theta = E_\pi g$. How can we extend these ideas to characteristics of the target distribution that cannot be represented as ergodic averages? For example, θ might be the qth quantile of π. Usually in the MCMC literature, when θ is a quantile what is nearly always meant is that it is a quantile of one of the univariate marginal distributions associated with π. The natural estimate of θ is simply the sample quantile from the observed Markov chain. For a general quantity, consider

$$\hat{\theta}_n = \hat{\theta}\left(X_1, \ldots, X_n\right) \tag{4.1}$$

as the appropriate estimate of θ. As before, $\hat{\theta}_n$ is only a point estimate for the quantity of interest and we would like to assess the MCSE of $\hat{\theta}_n$.

First, consider the sampling distribution of $\hat{\theta}_n$, Carlstein (1986) provides one set of conditions for asymptotic normality for a general statistic under non-trivial dependence. Here, we will present one corollary of his results for quantiles in the context of a stationary Markov chain. The relationship between the required mixing conditions from Carlstein (1986) and those stated here can be found in Section 2.4.2. Let us define for $\kappa \in [1/n, 1]$ the statistic $Z_b^i(\kappa)$ to be the $[\kappa n]$th ordered element of $\{X_{i+1}, X_{i+2}, \ldots, X_{i+b}\}$. If we consider a subsample from X of length $b_n = b$ where we are suppressing the dependency on n (though this is not necessary in general), then Proposition 4 gives us joint asymptotic normality between the same quantile from the sample and the subsample, $Z_n^0(\kappa)$ and $Z_b^i(\kappa)$ respectively.

Proposition 4. *(Carlstein, 1986, Corollary 10) Let X_1 have an absolutely continuous strictly increasing cdf F, with derivative f. Let $F_i(X_1, X_i)$ be the joint cdf of (X_1, X_i). Define $k := F^{-1}(\kappa)$. If $f(k) > 0$ and X is geometrically ergodic, then*

$$\nu := \sum_{i=-\infty}^{\infty} \left[F_i(k,k) - \kappa^2\right] < \infty \, .$$

If $\nu > 0$ and $t_b^i := \sqrt{b}(Z_b^i(\kappa) - k)$ then for any sequence n satisfying $b/n \to \rho^2$ and $n \geq i + b \geq b \to \infty$

$$\begin{pmatrix} t_n^0 \\ t_b^i \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{pmatrix}\right) \, , \quad \textit{for all } \rho^2 \in (0,1) \, ,$$

where $\sigma^2 = \nu / f^2(k)$.

Carlstein (1986) gives a more general result that is used by Politis et al. (1999) to address these questions for a general statistic. Applying this theorem requires further

work to estimate σ^2. To this end, we will focus our attention on the block bootstrap and subsampling methods.

4.1.1 Non-overlapping Block Bootstrap

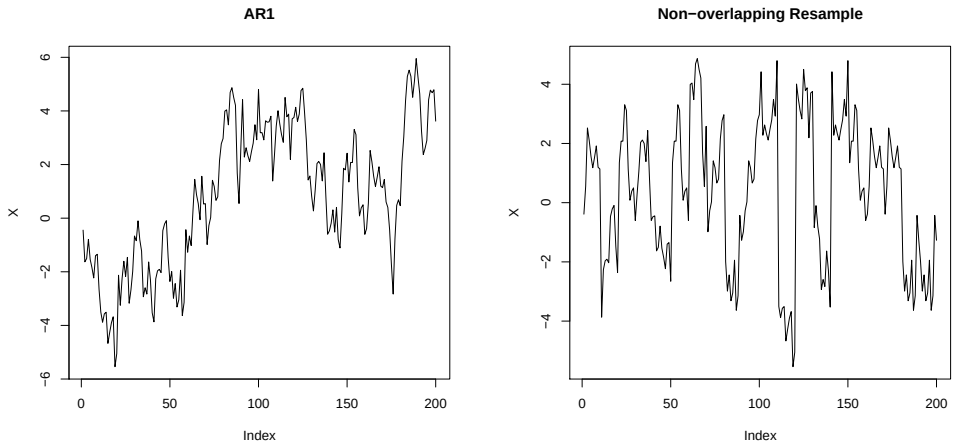
While there has been little investigation of the utility of bootstrap methods in the context of MCMC there has been a substantial amount of work on using bootstrap methods for stationary time–series. Some of this work is appropriate for use in MCMC; see e.g., Bertail and Clémençon (2006), Bühlmann (2002), Datta and McCormick (1993), Politis (2003). Efron and Tibshirani (1993) provide a thorough introduction to the bootstrap.

One of the simplest approaches is to use a non-overlapping block resampling scheme to estimate the Monte Carlo standard error of $\hat{\theta}_n$. Davison and Hinkley (1997) provide an overview of this technique. The basic idea is to split the Markov chain, X, into a non-overlapping blocks of length b, where we suppose the algorithm is run for a total of $n = ab$ iterations where $a = a_n$ and $b = b_n$ are functions of n. We will then sample the blocks independently with replacement and with equal probability and put these blocks together (end to end) to form a new series. We will use this new series to estimate θ yielding a single bootstrap replicate $\hat{\theta}^*_n$. Repeating this procedure p times results in p independent and identically distributed estimates; or

$$\left\{\hat{\theta}^*_{n,1}, \hat{\theta}^*_{n,2}, \ldots, \hat{\theta}^*_{n,p}\right\} .$$

Using the p replicates, we can appeal to classical bootstrap results to estimate the standard error of $\hat{\theta}_n$ by defining

$$\hat{\sigma}^2_B = \frac{1}{p-1}\sum_{j=1}^{p}(\hat{\theta}^*_{n,j} - \bar{\theta}^*_n)^2 \quad \text{where} \quad \bar{\theta}^*_n = \frac{1}{p}\sum_{j=1}^{p}\hat{\theta}^*_{n,j} .$$



(a) Original AR(1) Markov chain (b) Resampled AR(1) chain

Figure 4.1: Plots illustrating non-overlapping bootstrap resampling for time–series data using the AR(1) model.

We can also calculate an approximate $(1-\alpha)$-level confidence interval as

$$\left[\hat{\theta}_n - t_{p-1,\alpha/2} * \hat{\sigma}_B \quad , \quad \hat{\theta}_n + t_{p-1,\alpha/2} * \hat{\sigma}_B\right] .$$

While the simplicity of this method is appealing, there are a number of problems. For illustration purposes, consider the AR(1) model introduced in Chapter 2 with $\rho = .95$. Figure 4.1a shows a plot of $\{X_1, X_2, \ldots, X_{200}\}$ for one realization of this Markov chain. Using this data, Figure 4.1b shows a single non-overlapping bootstrap replicate with 10 batches, each of length 20. We can see the method does not retain the dependence structure of the original Markov chain. Resampling has introduced a number of very large jumps (where the blocks join) that are not seen in the original chain.

In general, resampling tends to generate sequences that are less dependent than the original. As in the AR(1) example, a chain with strong autocorrelation, block bootstrap resampling can result in unrepresentative samples. There are also exam-

ples (in time–series applications) that "can lead to catastrophically bad resampling approximations" (Davison and Hinkley, 1997, p. 397). Additionally, we found block resampling to be **extremely** computationally intensive even in relatively easy MCMC problems. With these problems in mind, the non-overlapping block bootstrap was not pursued.

4.1.2 Subsampling

This section will provide a brief overview of subsampling methods for dependent data taken from Politis et al. (1999). We will use this framework to estimate the Monte Carlo standard error of $\hat{\theta}_n$ and calculate an appropriate confidence interval. The benefit of subsampling (SUB) methods presented by Politis et al. (1999) are their generality. Here we will focus on SUB for MCMC. Section 2.4.2 outlines the connections between the mixing conditions as stated by Politis et al. (1999) and the presentation here in an MCMC context.

Subsampling differs from the traditional bootstrap in that we are taking samples (or blocks) of size b **without** replacement from the initial sample of size n. Typically, $b << n$, and in the case of time–series data, there are $n - b + 1$ such subsamples. Notice, if we assume X_1 is a draw from π, then the different blocks are identically distributed, though they are clearly still dependent. If we are interested in estimating θ, we can define the estimator based on the subsample $\{X_i, \ldots, X_{i+b-1}\}$ as

$$\theta^*_{n,b,i} = \hat{\theta}(X_i, \ldots, X_{i+b-1}) \text{ for } i = 1, \ldots, n - b + 1 \, .$$

Then define $J_{b,i}$ to be the sampling distribution of $\tau_b(\theta^*_{n,b,i} - \theta)$ where τ_b is the appropriate normalizing constant (which may be unknown). Simplifying notation slightly, define $\hat{\theta}_n := \theta^*_{n,n,1}$, the estimator from the whole sample as in (4.1). Then define $J_n := J_{n,1}$ to be sampling distribution of $\tau_n(\hat{\theta}_n - \theta)$ with τ_n equal to the appropriate

normalizing constant. The corresponding cumulative distribution function is

$$J_{b,i}(x) = \Pr\{\tau_b(\theta^*_{n,b,i} - \theta) \leq x\} \ .$$

We will approximate $J_n(x) := J_{n,1}(x)$ with the empirical distribution, namely

$$L_{n,b}(x) = \frac{1}{n-b+1} \sum_{i=1}^{n-b+1} I\{\tau_b(\theta^*_{n,b,i} - \theta) \leq x\} \ ,$$

where I is the usual indicator function on $\mathbb{Z}_+$. Then, the main assumption needed to consistently estimate J_n is as follows.

Assumption 3. (Politis et al., 1999, Assumption 4.2.1) There exists a limiting law J such that

1. J_n converges weakly to J as $n \to \infty$, and

2. for every continuity point x of J and for any sequences n, b with $n, b \to \infty$ and $b/n \to 0$, we have $\frac{1}{n-b+1} \sum_{i=1}^{n-b+1} J_{b,i}(x) \to J(x)$.

The first condition states that our estimator, properly normalized, has a limiting distribution. Politis et al. (1999) state "it is hard to conceive of any asymptotic theory free of such a requirement." While this may be the case, this is not always a trivial assumption to verify. If $\theta = E_\pi g$, then we can appeal to a CLT, but if we are interested in the case where θ is a quantile (or some other function), it is not entirely clear how to verify this. Currently, this is an open question, though Proposition 4 from Carlstein (1986) may provide an initial direction.

The second condition states that, for large n, the distribution functions for the subsamples will on average be close to the distribution function for the entire sample. Notice that if we assume stationarity, the second part of the assumption is not necessary.

Proposition 5 is stated without proof and follows directly from Theorem 4.2.1 of Politis et al. (1999). This result allows us to construct a theoretically valid general confidence interval.

Proposition 5. *Suppose Assumption 3 holds and that* $\tau_b/\tau_n \to 0$, $b/n \to 0$, *and* $b \to \infty$ *as* $n \to \infty$. *Also assume that* X *is geometrically ergodic.*

1. *If* x *is a continuity point of* $J(\cdot)$, *then* $L_{n,b}(x) \to J(x)$ *in probability.*

2. *If* $J(\cdot)$ *is continuous, then* $\sup_x |L_{n,b}(x) - J(x)| \to 0$ *in probability.*

3. *For* $\alpha \in (0,1)$, *let*

$$c_{n,b}(1-\alpha) = \inf\{x : L_{n,b}(x) \geq 1-\alpha\} .$$

Correspondingly, define

$$c_{\pi}(1-\alpha) = \inf\{x : J(x) \geq 1-\alpha\} .$$

If $J(\cdot)$ *is continuous at* $c_{\pi}(1-\alpha)$, *then*

$$Pr\{\tau_n(\hat{\theta}_n - \theta) \leq c_{n,b}(1-\alpha)\} \to 1-\alpha \text{ as } n \to \infty .$$

Thus, the asymptotic coverage probability of the interval $[\hat{\theta}_n - \tau_n^{-1}c_{n,b}(1-\alpha), \infty)$ is the nominal level $1-\alpha$. Politis et al. (1999) further extend this result for two-sided symmetric and equal-tailed intervals, where

$$[\hat{\theta}_n - \tau_n^{-1}c_{n,b}(1-\alpha/2) \quad , \quad \hat{\theta}_n + \tau_n^{-1}c_{n,b}(\alpha/2)] , \tag{4.2}$$

is a level $1-\alpha$ two-sided equal-tailed interval.

Next, we will consider estimating the variance of $\hat{\theta}_n$. First, we will define the estimator

$$\hat{\sigma}^2_{SUB} = \frac{\tau_b^2}{\tau_n^2} \frac{n}{n-b+1} \sum_{i=1}^{n-b+1} (\theta^*_{n,b,i} - \bar{\theta}^*_{n,b,\cdot})^2 \,, \tag{4.3}$$

where

$$\bar{\theta}^*_{n,b,\cdot} = \frac{1}{n-b+1} \sum_{i=1}^{n-b+1} \theta^*_{n,b,i} \,.$$

Remark 13. If $\theta = E_\pi g$, then the estimator from SUB in (4.3) is asymptotically equivalent to the estimator from OBM in (3.12).

Remark 14. Estimating τ_n when the limiting distribution converges slower than a $\sqrt{n}$-rate is an open question. Politis et al. (1999) give a method to estimate τ_n but there has been no investigation of this in the context of MCMC.

The following theorem gives conditions for the consistency of $\hat{\sigma}^2_{SUB}$. Define the normalized estimator as $T_n := \tau_n(\hat{\theta}_n - E(\hat{\theta}_n))$.

Proposition 6. *Let X be a Harris ergodic Markov chain with invariant distribution π. Further suppose X is geometrically ergodic. Assume that $\tau_b/\tau_n \to 0$, $b/n \to 0$, and $b \to \infty$ as $n \to \infty$. Also assume that*

1. *$\{(T_n)^4\}$ are uniformly integrable,*
2. *$\frac{n}{n-b+1} \sum_{i=1}^{n-b+1} Var(\tau_b \theta^*_{n,b,i}) \to \sigma^2$,*

Then $\hat{\sigma}^2_{SUB}$ is a consistent estimator of σ^2.

The proof follows directly from Lemma 4.6.1 of Politis et al. (1999). Again, the obvious question is how to verify the necessary assumptions? This again is an open research question, and a clear direction for future work.

In our examples we take $b_n = \lfloor n^{1/2} \rfloor$ as in our analysis using BM. We note that in our experience subsampling requires approximately the same computational effort

as BM (when programmed efficiently). Both of these methods are much less computationally intensive than the traditional bootstrap. For now, we will restrict our use of subsampling to when θ is a quantile.

4.2 Stopping the Simulation

In this section we consider two formal approaches to terminating the simulation when estimating general quantities. The first is based on calculating MCSEs and is a generalization of the approach used previously. The second is based on the Gelman-Rubin diagnostic (GRD) introduced in Chapter 1. As we have previously seen, GRD and MCSEs are used to stop the simulation in a similar manner. After n iterations either the value of the GRD or MCSE is calculated and if it isn't sufficiently small then we continue to run the chain until either value has met a user-specified criterion.

4.2.1 Fixed-Width Methods

In the previous chapters we suggested stopping the simulation when the MCSE of $\bar{g}_n$ is sufficiently small. In this section, we simply generalize this approach to $\hat{\theta}_n$. Of course, we may have to check whether this criterion is met many times, hence the procedure remains sequential.

Given an estimate of the Monte Carlo standard error of $\hat{\theta}_n$, say $\hat{\sigma}^2/\sqrt{n}$, we can form a confidence interval for θ. If this interval is too large then the value of n is increased and simulation continues until the interval is sufficiently small. The procedure is terminated when

$$t_* \frac{\hat{\sigma}^2_{SUB}}{\sqrt{n}} + p(n) \leq \epsilon \tag{4.4}$$

where t_* is an appropriate quantile, $p(n) = \epsilon I(n \leq n^*)$ where, $n^* > 0$ is fixed, I is

the usual indicator function on $\mathbb{Z}_+$ and $\epsilon > 0$ is the user-specified half-width. The role of p is to ensure that the simulation is not terminated prematurely due to a poor estimate of σ_g^2. If $\theta = E_\pi g$, there exist conditions such that this procedure will result in asymptotically valid confidence intervals. The immediate question here is whether these results can be extended to the general case.

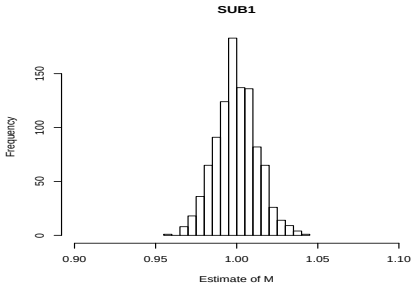
Non-parametric Approach

Previously, we proposed stopping the simulation when the half-width in (4.4) is below a pre-specified value. Alternatively, we could consider a non-parametric approach when using SUB methods. Using subsampling, we can calculate a non-parametric $1-\alpha$ confidence interval as in (4.2). Using these estimates, we can run the simulation until the length of the non-parametric confidence interval is below a pre-specified threshold as an alternative the approach in (4.4). In our examples, a non-parametric approach to stopping the simulation was not implemented.

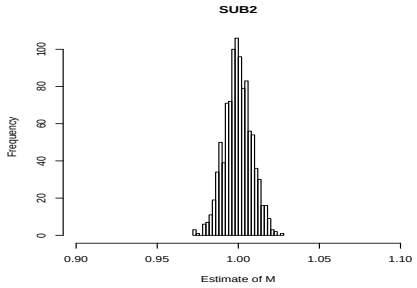
Toy Example Revisited

The toy example introduced in Section 1.3.2 will be used to illustrate the use of subsampling. Consider estimating the median of the conditional distribution of $\mu|y$, denoted M and the median of the conditional distribution of $\lambda|y$, denoted L. A routine calculation yields $M = 1$ (see Section 1.7.1). However, L cannot be easily calculated. Instead, the `integrate` function was used in R to calculate $L = 1.6780871$. (This calculation yielded an absolute error of 5.5×10^{-6} for $\Pr_\pi(\lambda < L)$ and 3.3×10^{-6} for $\Pr_\pi(\lambda > L)$.) Again, the output from the Gibbs sampler introduced earlier will be used estimate M and L. Further, subsampling will be used to assess the Monte Carlo error.

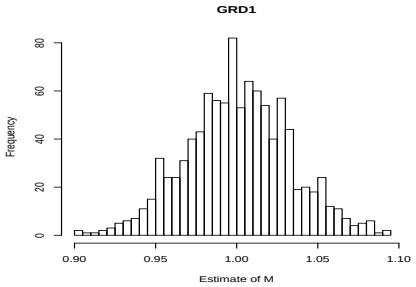
Consider estimating M and L with $\hat{M}$ and $\hat{L}$, the sample median of the output for each variable from the Gibbs sampler and using subsampling methods to calculate



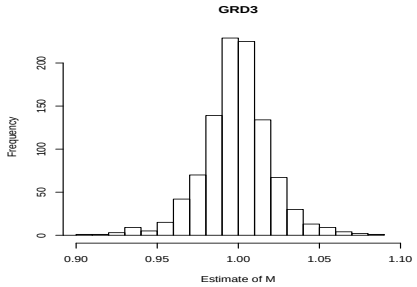
(a) SUB1, with a cutoff of $\epsilon = 0.06$.



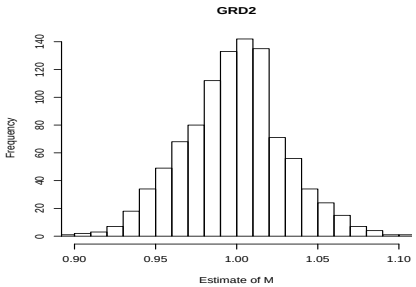
(b) SUB2, with a cutoff of $\epsilon = 0.04$.



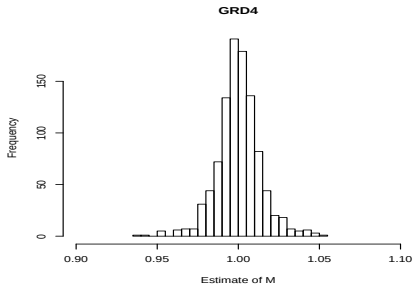
(c) GRD1, 2 chains and $\delta = 1.1$.



(d) GRD3, 2 chains and $\delta = 1.005$.



(e) GRD2, 4 chains and $\delta = 1.1$.



(f) GRD4, 4 chains and $\delta = 1.005$.

Figure 4.2: Histograms from 1000 replications estimating M for the toy example of Section 1.3.2 with SUB and GRD. Simulation sample sizes are given in Table 4.3.

an MCSE. We performed 1000 independent replications of the following procedure. Each replication of the Gibbs sampler was started from $\bar{y}$ and run for a minimum of 400 iterations. A replication was terminated when

$$t_{n-b_n} \frac{\hat{\sigma}_{SUB}}{\sqrt{n}} + \epsilon I(n < 400) \leq \epsilon \tag{4.5}$$

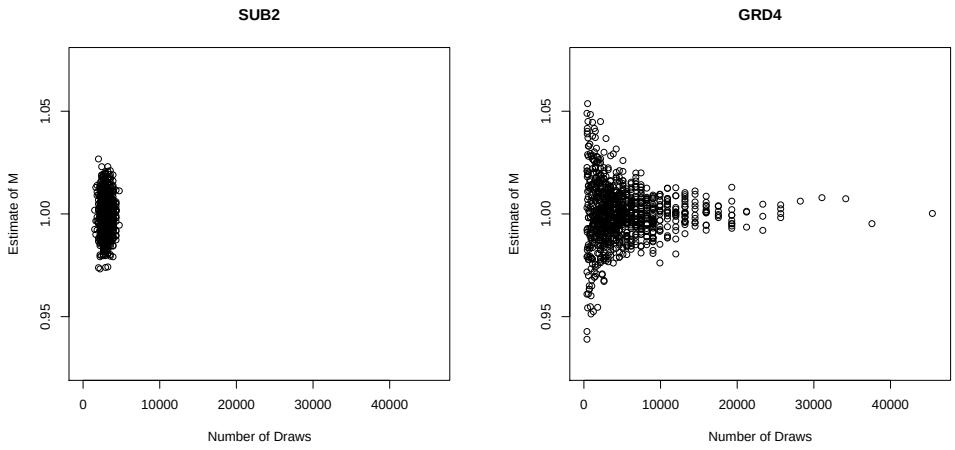
for each of the parameters of interest. Again, t_{n-b_n} is an appropriate quantile from Student's t distribution with $n - b_n$ degrees of freedom and $\epsilon > 0$ is the user-specified half-width. If the maximum half-width was not less than the cutoff, then 10% of the current chain length was added to the simulation before checking again. We used two settings for the cutoff, $\epsilon = 0.06$ and $\epsilon = 0.04$ denoted SUB1 and SUB2 respectively.

Figures 4.2a and 4.2b show the results from the simulation for estimating M. We can see that this proposed procedure performs well. Looking only at SUB2, Figure 4.3a shows the total length of chain used varies from 1509 to 4722 resulting in "good" estimates of M with a limited number of draws. Plots for estimating L lead to similar conclusions and are therefore omitted.

Table 4.1 summarizes the observed proportion of parametric confidence intervals based on (4.5) and a non-parametric confidence intervals based on (4.2) containing the true values of M and L. We can see all the results are close to the nominal 0.95 level. Here, the parametric intervals have a higher coverage in every case, though some of the non-parametric intervals are closer to the nominal value.

4.2.2 The Gelman-Rubin Diagnostic

Recall the Gelman-Rubin diagnostic (Gelman and Rubin, 1992; Brooks and Gelman, 1998) introduced in Chapter 1. For practitioners, this seems to be the most popular method for assessing the output of MCMC algorithms. This method is only appropriate when $\theta = E_\pi g$, hence SUB applies much more generally.



(a) $\hat{M}$ vs. n for SUB2 (b) $\hat{M}$ vs. n for GRD4

Figure 4.3: Estimating M for the toy example of Section 1.3.2. Estimates of M versus number of draws for the SUB2 and GRD4 settings.

	Parametric C.I.	
	M	L
SUB1	0.952 (0.0068)	0.946 (0.0072)
SUB2	0.965 (0.0058)	0.959 (0.0063)
	Non-Parametric C.I.	
	M	L
SUB1	0.931 (0.008)	0.917 (0.0087)
SUB2	0.949 (0.007)	0.932 (0.008)

Table 4.1: Summary of the proportion of replications when the true parameter value fell within a parametric and non-parametric confidence interval for estimating M and L for the Toy Example.

The starting point for GRD is the simulation of m independent parallel Markov chains having invariant distribution π, each of length $2l$. Thus the total simulation effort is $2lm$. The first l simulations are usually discarded and the inference is based on the last l simulations.

It is unclear how to implement GRD when θ is a quantile. The most obvious method is using the same stopping criteria used when trying to estimate $E_\pi g$. Specifically, based on the user provides a cutoff, $\delta > 0$, the simulation continues until

$$\hat{R}_{0.975} + p(n) \leq \delta \ ,$$

where $\hat{R}_{0.975}$ is defined in Section 1.4.2. As with fixed-width methods, the role of $p(n)$ is to ensure that we do not stop the simulation prematurely due to a poor estimate. By requiring a minimum total simulation effort of $n^* = 2lm$, we are effectively setting $p(n) = \delta I(n \leq n^*)$ where n indexes the total simulation effort. Again, we are left with the same practical implementation issues as when $\theta = E_\pi g$ (see Section 1.4.2). In addition, there is no reason to believe that $\hat{R} \approx 1$ implies that the quantile of $g(X_{ij})$ is a good estimate of true quantile of interest.

Toy Example

As before, the initial values when implementing GRD should be drawn from an "over-dispersed" distribution. However, we can sequentially sample the exact distribution of interest (see Section 1.7.1) to obtain starting values for the GRD method. Since we have started with a draw from the stationary distribution, the resulting marginal distribution for each X_i is the same as the stationary distribution.

To perform the GRD calculations, we took multiple chains starting from different draws from the sequential sampler. The multiple chains were started such that the total simulation effort was 400 draws, then the $\hat{R}_{0.975}$ was calculated for each param-

Method	Chains	Stopping Rule	MSE for M	S.E.	MSE for L	S.E.
SUB1	1	0.06	0.000165	7.6e-06	0.000911	4e-05
SUB2	1	0.04	7.09e-05	3.2e-06	0.000373	1.7e-05
GRD1	2	1.1	0.00103	4.6e-05	0.00541	0.00027
GRD2	4	1.1	0.000998	4.9e-05	0.00512	0.00023
GRD3	2	1.005	0.000461	3.1e-05	0.00238	0.00016
GRD4	4	1.005	0.000179	1.2e-05	0.000969	8e-05

Table 4.2: Summary table for all settings and estimated mean-squared-error for estimating M and L for the toy example of Section 1.3.2.

eter of interest and compared to a pre-specified value, ϵ. If the largest $\hat{R}_{0.975} < \epsilon$, the simulation was stopped. Otherwise, 10% of the current chain length was added to each chain before $\hat{R}_{0.975}$ was recalculated. This continued until the maximum $\hat{R}_{0.975}$ was below ϵ. This simulation procedure was repeated independently 1000 times starting from the same starting points. The settings considered were all combinations of 2 and 4 chains with $\delta_1 = 1.1$ proposed by Gelman et al. (2004) and $\delta_2 = 1.005$ in the hope of obtaining more precise estimates. Table 4.2 lists the different settings.

Upon completion of the simulation, $\hat{M}$ and $\hat{L}$ were recorded. Figures 4.2c-4.2f show histograms of $\hat{M}$ for each of the four settings using GRD. We can see that all the settings center around the true value of 1, and the use of the more stringent cut-off of 1.005 provides better estimates. Increasing the number of chains seems to have little impact on the quality of estimation using a cut-off of 1.1. However, with a cut-off of 1.005, we can see the histograms are narrower when more chains are used. The histograms of $\hat{L}$ are not shown, but careful examination leads to the same conclusions.

Let's focus our attention on the case of GRD4 where we have 4 chains and a cut-off of 1.005. Figure 4.3b show a plot of $\hat{M}$ versus the total number of draws in the chains,

Method	Prop. at Min.	S.E.	Prop. ≤ 1000	S.E.	N	S.E.
SUB1	0	0	0.075	0.0083	1385	8.7
SUB2	0	0	0	0	3059	16
GRD1	0.576	0.016	0.987	0.0036	469	4.1
GRD2	0.587	0.016	0.993	0.0026	471	4.2
GRD3	0.062	0.0076	0.363	0.015	2300	83.5
GRD4	0.01	0.0031	0.083	0.0087	5365	150.5

Table 4.3: Summary of the proportion (and standard error) of the observed estimates which were based on the minimum number (400) of draws, less than or equal to 1000 draws, and the average total simulation effort for the toy example in Section 1.3.2.

N. The plot clearly shows as the total number of draws increase, the quality of the estimate increases. The plot also illustrate that the total number of draws was highly variable using GRD. Similar to the results estimating means in Chapter 1, this implies that we are likely to run a simulation either too long or too short. Table 4.3 shows the proportion of the 1000 simulations which were stopped at their minimums and the proportion with less than 1000 total draws. This clearly shows that premature stopping was evident in all the settings for GRD, but was particularly bad using only 2 chains or a cut-off of 1.1. For example, looking at Figure 4.2f we can see GRD4 with the smallest cut-off and the most chains still has a number of the estimates are far from the actual value of 1.

4.3 Examples

4.3.1 Toy Example Continued

In this section, we will directly compare SUB to GRD in terms of mean-squared error (MSE) and chain length to estimate M and L. To this end we ran 1000 independent replications of the simulation procedure outlined above for each method. Table 4.3

shows the percentage of time the estimates of M and L were based on both the minimum and 1000 draws or less. We can see from this, that using GRD as a stopping rule tends to stop the simulation prematurely in a significant portion of the simulations.

In our experience, the time it took to run one replication with a similar number of draws was approximately equivalent for GRD and SUB. Therefore, we will use N, the total chain length, as a measure of the computational effort. Looking closely at SUB2 and GRD4, we can see that the average number of draws are 3059 and 5365 from Table 4.3. In this case, SUB2 is using less than 60% of draws required for GRD4. However, if we look at the quality of the estimation of M and L by comparing MSEs, we can see that the MSEs for GRD4 are 2.5 times higher than SUB2 for estimating M, and 2.6 times higher for estimating L (see Table 4.2). Again, it is clear the results using SUB are superior to those using GRD. In addition, SUB allows us to work with the actual parameter we are interested in, and provides a method to estimate the standard errors.

However, remember for this toy example we are using a sequential sampler to draw from the true distribution implying no burn-in is needed at all to implement GRD method (in this specific setting). Taking this into account, we can recalculate the estimates ignoring burn-in by using the entire chain length. For estimating M, we can calculate the MSE of 9.99×10^{-5} for GRD4 with a standard error of $7.8 \times 10{-}6$, resulting in GRD4 having a MSE 1.4 times larger than SUB2. For estimating L, the MSE from GRD4 is 0.000473 with a standard error of $3.6 \times 10{-}5$, resulting in GRD4 having an MSE 1.3 times higher than that of SUB2. This difference is not significant based on the standard errors, but if the simulation efforts were equal, the results would certainly be significant. Now, we can see SUB is still a superior method for estimating M and L because of the *variable total sample size* related to using GRD.

i	y	x	i	y	x
1	−0.2649	−0.0083	11	−0.3447	0.0298
2	−0.1552	0.4994	12	1.7768	1.6694
3	−2.5194	−0.1883	13	0.5059	0.5349
4	−0.8693	−0.3308	14	0.1383	0.9383
5	−0.7272	−0.7908	15	2.0452	0.2940
6	−3.0486	−1.9435	16	−1.9861	−0.4860
7	0.5511	0.0880	17	−0.3182	−0.2256
8	1.7590	1.9856	18	−2.3956	−1.9250
9	−1.0116	−0.5404	19	−2.5586	−1.0142
10	0.0094	0.3795	20	1.2208	0.7291

Table 4.4: Simulated data for block Gibbs example.

4.3.2 Block Gibbs Numerical Example

The goal of this example is to compare the finite sample properties obtained using BM to SUB. To this end, we will estimate the posterior means for β, λ_R and λ_D (denoted $E(\beta|y)$, $E(\lambda_R|y)$, and $E(\lambda_D|y)$ respectively), calculate the MCSEs with the two methods, and compare the results. In the case of estimating means, the estimate from SUB is simply equal to the estimate of OBM. Additionally, we will illustrate the use of SUB estimation of the median and first quartile for the three distributions of interest. The medians will be denoted $\beta^{(0.50)}$, $\lambda_R^{(0.50)}$, and $\lambda_D^{(0.50)}$. The first quartiles will be denoted $\beta^{(0.25)}$, $\lambda_R^{(0.25)}$, and $\lambda_D^{(0.25)}$. Application of the Markov chain CLT and BM require geometric ergodicity and finite moment conditions. Geometric ergodicity for the block Gibbs sampler was shown in Section 2.3.1 and, for now, we will assume the proper moment conditions have been met for our settings to implement BM and OBM. The theoretical justification for SUB is a future research direction.

First, we simulated the data in Table 4.4 with the model from Section 2.3.1 with $n = 20$, $\lambda_R = \lambda_D = 4$ and covariate $X \sim \mathrm{N}(0, I_{20})$. Then, using the block Gibbs sampler with prior settings $r_1 = r_2 = d_1 = d_2 = 4$, we estimated the true value for

Method	$E(\beta\|y)$	S.E.	$E(\lambda_R\|y)$	S.E.	$E(\lambda_D\|y)$	S.E.
BM	1.1897	0.0003	1.2106	0.0003	1.5417	0.0004
SUB	1.1891	0.0012	1.2113	0.0013	1.5398	0.0016
Method	$\beta^{(0.50)}$	S.E.	$\lambda_R^{(0.50)}$	S.E.	$\lambda_D^{(0.50)}$	S.E.
SUB	1.1917	0.0015	1.1463	0.0014	1.4695	0.0017
Method	$\beta^{(0.25)}$	S.E.	$\lambda_R^{(0.25)}$	S.E.	$\lambda_D^{(0.25)}$	S.E.
SUB	0.7916	0.0016	0.8677	0.0012	1.1533	0.0015

Table 4.5: Estimates for the nine quantities in the hierarchical example in Section 2.3.1. Calculations from BM are based on 4×10^6 iterations in the chain and subsampling calculations are based on 250,000 iterations in the chain. The standard errors (S.E.) for each of the observed quantities are included.

each of the three mean parameters, $E(\beta|y)$, $E(\lambda_R|y)$, and $E(\lambda_D|y)$, based on 4×10^6 iterations in the Markov chain. We are going to consider these estimates to be the "truth". Table 4.5 shows the numerical values of the parameters. In addition, we calculated the MCSE of these estimates using BM. Similarly, we estimated the true value for the median and first quartile parameters using 250,000 iterations from the Markov chain and calculated the MCSEs of these estimates using SUB. It's easy to see that the MCSEs are very small for BM and relatively small for SUB. These will be kept in mind when selecting ϵ for the fixed-width procedure that follows.

Consider the Markov chain $\{\xi_1, \ldots, \xi_m\}$. We estimated the posterior means for β, λ_R and λ_D with $\bar{\beta}_m$, $\bar{\lambda}_{R_m}$ and $\bar{\lambda}_{D_m}$ respectively and calculated the MCSE. Using BM, we performed 500 independent replications of a procedure similar to that employed in the toy example. First, each replication of the Gibbs sampler was started from the same point and run for a minimum of 400 iterations. A replication was terminated when the half-width of a 95% interval for each parameter of interest (three total) was below a pre-specified cutoff. The equation for a half-width using BM is given in (3.4). If the largest standard error was not less than the cutoff, then 10% of the current chain length was added to the chain before checking again. We used three settings

Method	ϵ	Mean half-width	S.E.	N	S.E.
BM1	0.06	0.056394	0.00013	754	7.4
BM2	0.04	0.037746	7.6e-05	1683	15.4
BM3	0.02	0.019078	3.1e-05	6646	47.2
SUB1	0.06	0.05813	5.6e-05	946	6.8
SUB2	0.04	0.038801	3.4e-05	2088	14.1
SUB3	0.02	0.019478	1.6e-05	8142	38.6

Table 4.6: Summary table for all settings considered in the hierarchical example in Section 2.3.1. This table also gives the mean observed value of the half-width and number of iterations. Both are reported with standard errors in the additional column.

for the cutoff for each method, $\epsilon_1 = 0.06$, $\epsilon_2 = 0.04$, and $\epsilon_3 = 0.02$. These settings will be denoted BM1, BM2, and BM3 respectively for BM.

Using subsampling, we employed the same basic procedure, but in addition to estimating the posterior means, we estimated $\beta^{(0.50)}$, $\lambda_R^{(0.50)}$, $\lambda_D^{(0.50)}$, $\beta^{(0.25)}$, $\lambda_R^{(0.25)}$, and $\lambda_D^{(0.25)}$, nine total quantities. A replication was terminated when the half-width for a 95% interval for each of the nine parameters was below the pre-specified cutoff. The SUB half-widths were calculated as in (4.5). We used the same three cutoffs as above, and denoted them SUB1, SUB2, and SUB3.

First, we will directly compare BM to SUB for estimating the three posterior means using the MSEs and the overall chain length. Looking at histograms (not shown) for each situation, we see they are centered around the calculated "truth". Table 4.6 summarizes the mean largest half-width and mean iterations for each setting when the replication was terminated. We can see, for equal ϵ settings, that subsampling ran slightly longer than BM. This is based on the fact that we are estimating nine parameters with subsampling and only three with BM. In this example, the subsampling processing time was appreciably slower than BM. This is most likely a result of inefficient code, though this issue warrants further attention.

Method	ϵ	MSE for $E(\beta\|y)$	S.E.	MSE for $E(\lambda_R\|y)$	S.E.	MSE for $E(\lambda_D\|y)$	S.E.
BM1	0.06	0.000597	4.1e-05	0.000511	3.1e-05	0.000933	6.1e-05
BM2	0.04	0.000259	1.6e-05	0.000264	1.9e-05	0.000404	2.5e-05
BM3	0.02	5.83e-05	3.7e-06	5.5e-05	3.4e-06	9.19e-05	5.8e-06
SUB1	0.06	0.00044	2.7e-05	0.000419	2.6e-05	0.000691	4.4e-05
SUB2	0.04	0.000203	1.4e-05	0.00021	1.3e-05	0.000314	2.1e-05
SUB3	0.02	5.3e-05	3.5e-06	4.83e-05	2.9e-06	8.61e-05	5e-06
Method	ϵ	MSE for $\beta^{(0.50)}$	S.E.	MSE for $\lambda_R^{(0.50)}$	S.E.	MSE for $\lambda_D^{(0.50)}$	S.E.
SUB1	0.06	0.000609	4.1e-05	0.000516	3.4e-05	0.000781	4.9e-05
SUB2	0.04	0.00024	1.6e-05	0.000239	1.6e-05	0.000342	2.3e-05
SUB3	0.02	7.03e-05	4.7e-06	5.93e-05	3.7e-06	0.000107	5.9e-06
Method	ϵ	MSE for $\beta^{(0.25)}$	S.E.	MSE for $\lambda_R^{(0.25)}$	S.E.	MSE for $\lambda_D^{(0.25)}$	S.E.
SUB1	0.06	0.000802	5e-05	0.000446	3e-05	0.000645	4.2e-05
SUB2	0.04	0.000351	2.3e-05	0.000224	1.5e-05	0.000278	1.9e-05
SUB3	0.02	9.11e-05	5.6e-06	4.52e-05	2.7e-06	7.6e-05	4.9e-06

Table 4.7: Summary of replications for estimating the nine parameters of interest for the Hierarchical example in Section 2.3.1 based on 500 independent replications. Standard errors (S.E.) are listed for each of the quantities.

Table 4.7 shows the calculated MSEs for the two different methods. Notice that for equal settings for ϵ, we can see that the MSEs from BM and SUB are approximately equal. For $\epsilon_1 = 0.06$ and $\epsilon_2 = 0.04$, confidence intervals for the MSE of BM (based on the standard error) do not contain observed MSE for SUB. However, subsampling is using more iterations for equal ϵ because it is estimating nine quantities rather than three. For $\epsilon_3 = 0.02$, the two methods are not statistically different, based on the same reasoning. We can also compare the BM and SUB by looking at the univariate parametric confidence intervals for each replication. For 17 of the possible 18 combinations, the proportion of intervals containing the true parameter is within two standard errors of the nominal value of 0.95. The exception is BM1 when estimating $E(\lambda_D|y)$. In this case the proportion of intervals is 0.922 with a standard error of 0.012, so the estimate is within 2.5 standard errors of the mean. This is still a very reasonable result considering we have 18 total combinations.

The final results using BM and SUB are very similar in this case. The benefit of using BM is the moderately faster computational time, while SUB will enjoy a lower asymptotic variance (because of the relationship with OBM shown in Section 3.2.3). In the case of estimating ergodic averages, both methods enjoy checkable theoretical conditions to ensure consistent estimators. However, as mentioned, BM will not allow us to calculate MCSEs for estimating quantiles, this will require SUB.

Now, we will illustrate the use of subsampling for estimating univariate quantiles of the posterior distribution. The six parameters of interest are $\beta^{(0.50)}$, $\lambda_R^{(0.50)}$, $\lambda_D^{(0.50)}$, $\beta^{(0.25)}$, $\lambda_R^{(0.25)}$, and $\lambda_D^{(0.25)}$. Again, histograms (not shown) for each situation show nothing alarming, with the estimates centered around the true value in each case. Table 4.7 shows the MSEs for estimating the quantile parameters are on the same order of magnitude as those from estimating the mean parameters. Looking at the univariate parametric confidence intervals, the observed coverage probabilities of all 18 were within two standard errors of the nominal 0.95 value. For this example, sub-

sampling is doing the "right" thing for estimating means and quantiles, and provides a method to calculate MCSEs in both of these cases.

4.3.3 Discussion

As we have seen in our examples, subsampling methods seem to work well estimating quantiles in finite sample settings. There are a number of future directions for this research. A substantial project would be to extend the ideas for ergodic averages in Chapter 3 to quantiles, and further to general quantities, via subsampling methods. Specifically, can the work from Politis et al. (1999) be translated into checkable conditions for MCMC practitioners? The use of subsampling also allows calculating non-parametric confidence intervals as in (4.2). Do these intervals ensure asymptotically valid confidence intervals using fixed-width methods? Further research is also needed in the case of establishing checkable conditions for limiting distributions for quantities that cannot be expressed as ergodic averages. Here, the work of Carlstein (1986) provides an excellent starting point, but more work is necessary. Finally, and potentially most importantly, there is a culture in MCMC literature of rarely reporting MCSEs for the quantities of interest, including quantiles. Hence, further work is necessary in applying the results in this thesis to important real problems seen by practitioners.

Appendix A

Supplementary Material

A.1 Brownian Motion

Brownian motion is central to some of the proofs throughout this discussion. Recall that Brownian motion is a continuous process with independent, stationary, and normally distributed increments. Throughout this thesis, we will let $B = \{B(t), t \geq 0\}$ denote a standard Brownian motion process, hence $B(t) - B(s) \sim N(0, t-s)$ for all $0 \leq s < t$. For ease of exposition, define $\bar{B}_j(k) := k^{-1}(B(j+k) - B(j))$ and $\bar{B}_n := n^{-1}B(n)$ throughout this section. Further define $U_i := B(i) - B(i-1)$ as the increments of Brownian motion between times i and $i-1$, then $U_1, \ldots, U_n$ are i.i.d. $N(0,1)$.

In this work, we will be particularly interested in Law of Iterated Logarithm (LIL) type results from Csörgő and Révész (1981) on the increments of Brownian motion. The following two lemmas will be used in future sections.

Lemma 15. *(Csörgő and Révész, 1981) For all $\epsilon > 0$ and for almost all sample paths there exists $n_0(\epsilon)$ such that for all $n \geq n_0$*

$$|B(n)| < (1+\epsilon)\left[2n \log\log n\right]^{1/2} . \tag{A.1}$$

Lemma 16. *(Damerdji, 1994, Lemma 7.1.2) Suppose Assumption 2 holds, then for all $\epsilon > 0$ and for almost all sample paths, there exists $n_0(\epsilon)$ such that for all $n \geq n_0$*

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

Corollary 3. *(Damerdji, 1994, p. 508) For all $\epsilon > 0$ and for almost all sample paths, there exists $n_0(\epsilon)$ such that for all $n \geq n_0$*

$$|\bar{B}_j(a_n)| \leq \sqrt{2}(1+\epsilon)a_n^{-1/2}\left(\log\frac{n}{a_n} + \log\log n\right)^{1/2} , \tag{A.2}$$

where $\bar{B}_j(a_n) = a_n^{-1}(B((j+1)a_n) - B(ja_n))$.

A.1.1 Results for Spectral Variance

Define

$$\tilde{\sigma}_*^2 = n^{-1}\sum_{j=0}^{n-b_n}\sum_{k=1}^{b_n}\alpha_n(k)(\bar{B}_j(k) - \bar{B}_n)^2 .$$

For ease of exposition, let $T_i = U_i - \bar{B}_n$ for $i = 1, \ldots, n$ and further define

$$\begin{aligned}
\tilde{\gamma}_n(s) &= n^{-1}\sum_{t=1}^{n-s} U_t U_{t+s} \text{ for } s = 1, \ldots, n , \\
\tilde{f}_n(0) &= \frac{1}{2\pi}\sum_{s=-(b_n-1)}^{b_n-1} w_n(s)\tilde{\gamma}_n(s) \text{ , and} \\
\tilde{d}_n &= n^{-1}\left(\left[\sum_{l=1}^{b_n}\Delta_1 w_n(l)\left(\sum_{i=1}^{l-1}T_i^2 + \sum_{i=n-b_n+l+1}^{n}T_i^2\right)\right]\right. \\
&\quad \left. + 2\sum_{s=1}^{b_n-1}\left[\sum_{l=1}^{b_n-s}\Delta_1 w_n(s+l)\left(\sum_{i=1}^{l-1}T_iT_{s+i} + \sum_{i=n-b_n+l+1}^{n-s}T_iT_{s+i}\right)\right]\right) .
\end{aligned}$$

Proposition 7. *Suppose Assumptions 1 and 2 hold. Further assume*

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

2. $b_n n^{-1} \log n \to 0$ *as* $n \to \infty$,

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

4. *and* $b_n^{-1} \log n$ *stays bounded as* $n \to \infty$.

Then $\tilde{\sigma}_*^2 \to 1$ *as* $n \to \infty$.

Proof. Proposition 3 shows,

$$\tilde{\sigma}_*^2 = 2\pi \tilde{f}_n(0) - \tilde{d}_n \, .$$

Lemma 17 shows $2\pi \tilde{f}_n(0) \to 1$ and Lemma 19 shows $\tilde{d}_n \to 0$ as $n \to \infty$. Thus, combining all of these yield the desired result. □

Lemma 17. *(Damerdji, 1991, Theorem 4.1) Suppose Assumptions 1 and 2 hold. Further assume*

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

2. $b_n n^{-1} \log n \to 0$ *as* $n \to \infty$.

Then $2\pi \tilde{f}_n(0) \to 1$ *as* $n \to \infty$ *w.p.1.*

Proof. This proof appears in Damerdji (1991) but it is expanded on here for clarity.

First, notice that

$$\begin{aligned} 2\pi \tilde{f}_n(0) &= \sum_{i=-(b_n-1)}^{b_n-1} w_n(i) \tilde{\gamma}_n(i) \\ &= \tilde{\gamma}_n(0) + 2 \sum_{i=1}^{b_n-1} w_n(i) \tilde{\gamma}_n(i) \, . \end{aligned}$$

Since $U_1, \ldots, U_n$ are i.i.d. $N(0,1)$, the classical Strong Law of Large Numbers implies that $\tilde{\gamma}_n(0) \to 1$ as $n \to \infty$ w.p.1. Hence, it suffices to show that $\sum_{i=1}^{b_n-1} w_n(i)\tilde{\gamma}_n(i) \to 0$ as $n \to \infty$ w.p.1. Then we can write for $i \leq n$

$$\begin{aligned}
\tilde{\gamma}_n(i) &= n^{-1} \sum_{t=1}^{n-i} (U_t - \bar{B}_n)(U_{t+i} - \bar{B}_n) \\
&= n^{-1} \left[\sum_{t=1}^{n-i} U_t U_{t+i} + (n-i)\, \bar{B}_n^2 - \bar{B}_n \sum_{t=1}^{n-i} (U_t + U_{t+i}) \right] \\
&= n^{-1} \sum_{t=1}^{n-i} U_t U_{t+i} - \left(1 + \frac{i}{n}\right) \bar{B}_n^2 + 2\bar{B}_n^2 - n^{-1}\bar{B}_n \sum_{t=1}^{n-i} (U_t + U_{t+i}) \\
&= n^{-1} \sum_{t=1}^{n-i} U_t U_{t+i} - \left(1 + \frac{i}{n}\right) \bar{B}_n^2 + \bar{B}_n n^{-1} \left[B(i) + (B(n) - B(n-i)) \right] ,
\end{aligned}$$

resulting in

$$\begin{aligned}
\sum_{i=1}^{b_n-1} w_n(i)\tilde{\gamma}_n(i) = \sum_{i=1}^{b_n-1} w_n(i) n^{-1} \sum_{t=1}^{n-i} U_t U_{t+i} - \bar{B}_n^2 \sum_{i=1}^{b_n-1} w_n(i) \left(1 + \frac{i}{n}\right) \\
+ \bar{B}_n n^{-1} \sum_{i=1}^{b_n-1} w_n(i) \left[B(i) + (B(n) - B(n-i)) \right] . \qquad \text{(A.3)}
\end{aligned}$$

We will now show that each of the three terms in (A.3) tend to 0 separately.

1. Here we will show that

$$\sum_{i=1}^{b_n-1} w_n(i) n^{-1} \sum_{t=1}^{n-i} U_t U_{t+i} \to 0 \text{ as } n \to \infty \quad w.p.1. \qquad \text{(A.4)}$$

To this end, we will use a Borel-Cantelli argument with condition 1. It suffices to show

$$E\left[\left(\sum_{i=1}^{b_n-1} w_n(i) n^{-1} \sum_{t=1}^{n-i} U_t U_{t+i} \right)^{2a} \right] = O\left(\frac{b_n}{n}\right)^a , \qquad \text{(A.5)}$$

where the symbol "O" is the Big-Oh notation, which means there exists n_0 and

a constant $C > 0$ such that

$$\left| E\left[\left(\sum_{i=1}^{b_n-1} w_n(i) n^{-1} \sum_{t=1}^{n-i} U_t U_{t+i} \right)^{2c} \right] \right| \leq C(b_n/n)^c$$

for all $n > n_0$. Then since

$$\sum_{m=1}^{\infty} C \left(\frac{b_{n_0+m}}{n_0+m} \right)^c < \infty$$

using condition 1, (A.4) holds via Borel-Cantelli.

Now we will show that (A.5) holds for any positive integer $c \geq 2$ using the following lemma.

Lemma 18. *(Damerdji, 1991, Corollary 4) For a sequence of i.i.d. standard normal variables $\{U_n : n \geq 1\}$ where $A = \sum_j \sum_k a_{jk} u_j u_k$, then for $c \geq 2$ we have*

$$E\left[|A - EA|^{2c}\right] \leq K(c) \left(\sum_j \sum_k a_{jk}^2 \right)^c ,$$

for some constant $K(c)$ depending only on c.

If we then let $D(n)$ be the $n \times n$ matrix with entries,

$$D_{i,j}(n) = \begin{cases} n^{-1} w_n(j-i) & \text{for } i < j \ , \\ 0 & \text{otherwise.} \end{cases}$$

Define $U := (U_1, \ldots, U_n)^T$, then by construction of the matrix $D(n)$

$$\begin{aligned} A(n) &= U'D(n)U \\ &= \sum_{t=1}^{n} U_t \sum_{i=1}^{t-1} n^{-1} w_n(i) U_{t-i} \\ &= \sum_{i=1}^{n-1} \sum_{t=i+1}^{n} n^{-1} w_n(i) U_{t-i} U_t \\ &= \sum_{i=1}^{n-1} w_n(i) n^{-1} \sum_{t=1}^{n-i} U_t U_{t+i} \\ &= \sum_{i=1}^{b_n-1} w_n(i) n^{-1} \sum_{t=1}^{n-i} U_t U_{t+i} \end{aligned}$$

where the last step results from the fact that $w_n(s) = 0$ for all $|s| \geq b_n$.

Recall $U_1, \ldots, U_n$ are i.i.d. $N(0,1)$, then $EA(n) = 0$ since $EU_tU_{t+i} = 0$ for $i \geq 1$, and we can apply Lemma 18 as follows

$$E\left[\left(\sum_{i=1}^{b_n-1} w_n(i) n^{-1} \sum_{t=1}^{n-i} U_t U_{t+i}\right)^{2c}\right] \leq K(c) \left(\sum_j \sum_k a_{jk}^2\right)^c$$

where the coefficients a_{jk} are the elements of the matrix $D(n)$, and $K(c)$ is a constant in c. Now we can see that

$$\sum_j \sum_k a_{jk}^2 = n^{-2} \sum_{i=1}^{b_n-1} (n-i) w_n^2(i) \leq n^{-1} \sum_{i=1}^{b_n-1} w_n^2(i) \leq b_n n^{-1}$$

and hence, we have shown the relationship in (A.5), implying the first term in (A.3) tends to 0 as $n \to \infty$.

2. From Assumption 1, the fact that $1 + i/n \leq 2$, and (A.1), it follows that

$$\begin{aligned} 0 \leq \bar{B}_n^2 \left| \sum_{i=1}^{b_n - 1} w_n(i) \left(1 + \frac{i}{n}\right) \right| &\leq 2 b_n \bar{B}_n^2 \\ &\leq 2 b_n n^{-2} (1+\epsilon)^2 2n \log\log n = 4(1+\epsilon)^2 b_n n^{-1} \log\log n \to 0 \ , \end{aligned}$$

as $n \to \infty$ by condition 2.

3. Again, we will use the LIL type results from Appendix A.1, specifically we will apply Lemma 16. For $1 \leq i \leq b_n$,

$$\begin{aligned} |B(i) + (B(n) - B(n-i))| &\leq \sup_{0 \leq s \leq b_n} |B(s)| + \sup_{0 \leq s \leq b_n} |B(n) - B(n-s)| \\ &\leq \sup_{0 \leq t \leq n - b_n} \sup_{0 \leq s \leq b_n} |B(t+s) - B(t)| \\ &\quad + \sup_{0 \leq s \leq b_n} |B(n) - B(n-s)| \\ &\leq 2(1+\epsilon) \left(2 b_n \left(\log \frac{n}{b_n} + \log\log n \right) \right)^{1/2} \ . \end{aligned}$$

Therefore, from Lemma 15 and the above relationship,

$$\begin{aligned} &\left| \bar{B}_n n^{-1} \sum_{i=1}^{b_n - 1} w_n(i) \left[B(i) + (B(n) - B(n-i)) \right] \right| \\ &\qquad \leq 4 n^{-2} (1+\epsilon)^2 \left(n \log\log n \right)^{1/2} b_n \left(b_n \left(\log \frac{n}{b_n} + \log\log n \right) \right)^{1/2} \\ &\qquad = O(b_n^{3/2} n^{-3/2} \log n) \ . \end{aligned}$$

Hence, the above will go to 0 as a result of Assumption 2 and condition 2.

□

Lemma 19. *(Damerdji, 1991, Proposition 4.3) Suppose Assumptions 1 and 2 hold. If*

1. $$b_n n^{-1} \sum_{k=1}^{b_n} k \left|\Delta_1 w_n(k)\right| \to 0 \text{ as } n \to \infty$$

2. *and* $b_n^{-1} \log n$ *stays bounded as* $n \to \infty$.

Then $\tilde{d}_n \to 0$ *as* $n \to \infty$ *w.p.1.*

Proof. Recall $T_i = U_i - \bar{B}_n$ for $i = 1, \ldots, n$ and

$$\tilde{d}_n = n^{-1} \left(\left[\sum_{l=1}^{b_n} \Delta w_n(l) \left(\sum_{i=1}^{l-1} T_i^2 + \sum_{i=n-b_n+l+1}^{n} T_i^2 \right) \right] + 2 \sum_{s=1}^{b_n-1} \left[\sum_{l=1}^{b_n-s} \Delta w_n(s+l) \left(\sum_{i=1}^{l-1} T_i T_{s+i} + \sum_{i=n-b_n+l+1}^{n-s} T_i T_{s+i} \right) \right] \right) ,$$

We will procede similarly to the proof of Lemma 10.

1. We can bound $|\tilde{d}_n|$ as before

$$\tilde{d}_n \leq b_n^{-1} \left(\sum_{i=1}^{b_n} T_i^2 + \sum_{i=n-b_n+1}^{n} T_i^2 \right) \tag{A.6}$$

$$\times\, n^{-1} b_n \left(\sum_{l=1}^{b_n} |\Delta w_n(l)| + 2 \sum_{k=1}^{b_n} k \left|\Delta w_n(k)\right| \right) . \tag{A.7}$$

   Again, (A.7) tends to 0 as $n \to \infty$ by condition 1. Hence, it suffices to show that (A.6) is bounded w.p.1.

2. The classical strong law of large numbers implies the sums $\sum_{i=1}^{b_n} U_i$ and $\sum_{i=1}^{b_n} U_i^2$ stay bounded w.p.1. Then we can apply the same argument as above to show the contribution from the starting effects stay bounded for $\tilde{d}_n$.

3. Finally, we are left to show that the sums contributed from the terminating end effects stay bounded, or $b_n^{-1}\sum_{i=n-b_n+1}^{n} T_i^2 < \infty$ as $n \to \infty$. Using the same technique as before, it suffices to show $b_n^{-1}\sum_{i=n-b_n+1}^{n} U_i$ and $b_n^{-1}\sum_{i=n-b_n+1}^{n} U_i^2$ stay bounded. We have $b_n^{-1}\sum_{i=n-b_n+1}^{n} U_i = b_n^{-1}(B(n) - B(n-b_n))$, which stays bounded by Lemma 16 and condition 2. As for $b_n^{-1}\sum_{i=n-b_n+1}^{n} U_i^2$, we will use the following strong invariance principle. Komlós et al. (1975), Komlós et al. (1976), and Major (1976) have shown in the i.i.d. case, that if $E[\exp|tX_1|] < \infty$ in a neighborhood of $t = 0$, then

$$S_n - n\mu = \sigma B(n) + O(\log n) \quad w.p.1$$

where the $\log n$ rate is extremely sharp. Since $U_1, \ldots, U_n$ are i.i.d. $N(0,1)$ the sequence $U_1^2, \ldots U_n^2$ are i.i.d. χ_1^2. This implies that

$$\left| \sum_{i=1}^{n} U_i^2 - n - 2B(n) \right| \leq C' \log n \ ,$$

resulting in

$$\begin{aligned} b_n^{-1}\left|\sum_{i=n-b_n+1}^{n} U_i^2\right| &= b_n^{-1}\left|\sum_{i=1}^{n} U_i^2 - \sum_{i=1}^{n-b_n} U_i^2\right| \\ &= b_n^{-1}\Bigg|\left(\sum_{i=1}^{n} U_i^2 - n - 2B(n)\right) \\ &\quad - \left(\sum_{i=1}^{n-b_n} U_i^2 - (n-b_n) - 2B(n-b_n)\right) \\ &\quad + 2\left(B(n) - B(n-b_n)\right) + b_n\Bigg| \\ &\leq b_n^{-1}\left(2C'\log n + (1+\epsilon)\left(2b_n\left(\log\frac{n}{b_n} + \log\log n\right)\right)^{1/2} + b_n\right) \\ &= 1 + 2C' b_n^{-1}\log n + O\left((b_n^{-1}\log n)^{1/2}\right) \quad w.p.1. \end{aligned}$$

Hence, $b_n^{-1}\left|\sum_{i=n-b_n+1}^{n} U_i^2\right|$ stays bounded w.p.1 from condition 2.

Thus the proof is complete. □

A.1.2 Results for Overlapping Batch Means

As shown previously, OBM is a special case of a SV estimator. In this section, we will show $\tilde{\sigma}^2(n) \to 1$ w.p.1 where

$$\tilde{\sigma}^2(n) = b_n n^{-1}\sum_{j=0}^{n-b_n}\left(\bar{B}_j(k) - \bar{B}_n\right)^2$$

under less stringent conditions than Proposition 7.

The following lemma from Kendall and Stuart (1977) is required.

Lemma 20. *If $Z \sim \chi_v^2$, then for all positive integers r there exists a constant $K := K(r)$ such that $E\left[(Z-v)^{2r}\right] \leq Kv^r$.*

Proposition 8. *Assume Assumption 2 and (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 $n \to \infty$; and (c) $n^{-3/2} b_n^2 \to 0$ as $n \to \infty$, then as $n \to \infty$, $\tilde{\sigma}^2(n) \to 1$ w.p.1.*

Proof. Let $w_n(k) = 1 - |k|/b_n$ for $|k| < b_n$ and 0 otherwise, the modified Bartlett lag window. Consider $\alpha_n(k) := k^2 \Delta_2 w_n(k)$ is a sequence of weights where $\Delta_2 w_n(k) = w_n(k-1) - 2w_n(k) + w_n(k+1)$. Recall that

$$\begin{aligned}\tilde{\sigma}^2(n) &= b_n n^{-1} \sum_{j=0}^{n-b_n} \left(\bar{B}_j(k) - \bar{B}_n\right)^2 \\ &= n^{-1} \sum_{j=0}^{n-b_n} \sum_{k=1}^{b_n} \alpha_n(k) \left(\bar{B}_j(k) - \bar{B}_n\right)^2 \ .\end{aligned}$$

Using this representation, Proposition 3 in Section 3.4.2 shows there exists a sequence $\tilde{d}_n$ due to some end effects, such that

$$\tilde{\sigma}^2(n) = 2\pi \tilde{f}_n(0) - \tilde{d}_n \ .$$

Lemma 17 shows $2\pi \tilde{f}_n(0) \to 1$ as $n \to \infty$ w.p.1. Using the modified Bartlett lag window, Lemma 21 shows $\tilde{d}_n \to 0$ as $n \to \infty$ w.p.1. Combining these yield the desired result. □

Lemma 21. *Suppose Assumption 2 holds and $w_n(k) = 1 - |k|/b_n$ for $|k| < b_n$ and 0 otherwise. If*

1. *there exists a constant $c \geq 1$ such that $\sum_n (b_n/n)^c < \infty$;*
2. *$b_n^{-1} \log n$ stays bounded as $n \to \infty$;*
3. *and $n^{-3/2} b_n^2 \to 0$ as $n \to \infty$.*

Then $\tilde{d}_n \to 0$ as $n \to \infty$ w.p.1.

Proof. Recall $T_i = U_i - \bar{B}_n$ for $i = 1, \ldots, n$. Using the Bartlett lag window and the definition of $\tilde{d}_n$,

$$\begin{aligned}
|\tilde{d}_n| &= n^{-1} \left| \left[\sum_{l=1}^{b_n} \Delta w_n(l) \left(\sum_{i=1}^{l-1} T_i^2 + \sum_{i=n-b_n+l+1}^{n} T_i^2 \right) \right] \right. \\
&\quad \left. + 2 \sum_{s=1}^{b_n-1} \left[\sum_{l=1}^{b_n-s} \Delta w_n(s+l) \left(\sum_{i=1}^{l-1} T_i T_{s+i} + \sum_{i=n-b_n+l+1}^{n-s} T_i T_{s+i} \right) \right] \right| \\
&\leq n^{-1} \left| \sum_{l=1}^{b_n} b_n^{-1} \left(\sum_{i=1}^{l-1} T_i^2 + \sum_{i=n-b_n+l+1}^{n} T_i^2 \right) \right| \qquad \text{(A.8)} \\
&\quad + n^{-1} \left| 2 \sum_{s=1}^{b_n-1} \sum_{l=1}^{b_n-s} b_n^{-1} \left(\sum_{i=1}^{l-1} T_i T_{s+i} + \sum_{i=n-b_n+l+1}^{n-s} T_i T_{s+i} \right) \right| . \qquad \text{(A.9)}
\end{aligned}$$

We will show that (A.8) and (A.9) each tend to zero implying the desired result.

First consider (A.8),

$$\begin{aligned}
n^{-1} \left| \sum_{l=1}^{b_n} b_n^{-1} \left(\sum_{i=1}^{l-1} T_i^2 + \sum_{i=n-b_n+l+1}^{n} T_i^2 \right) \right| &\leq n^{-1} \left| \sum_{l=1}^{b_n} b_n^{-1} \left(\sum_{i=1}^{b_n-1} T_i^2 + \sum_{i=n-b_n+2}^{n} T_i^2 \right) \right| \\
&= n^{-1} \left| \sum_{i=1}^{b_n-1} T_i^2 + \sum_{i=n-b_n+2}^{n} T_i^2 \right| \\
&= n^{-1} \left| \sum_{i=1}^{b_n-1} \left(U_i - \bar{B}_n \right)^2 + \sum_{i=n-b_n+2}^{n} \left(U_i - \bar{B}_n \right)^2 \right| \\
&\leq n^{-1} \left| \sum_{i=1}^{b_n-1} U_i^2 + \sum_{i=n-b_n+2}^{n} U_i^2 \right| \\
&\quad + n^{-1} \left| 2\bar{B}_n \left(\sum_{i=1}^{b_n-1} U_i + \sum_{i=n-b_n+2}^{n} U_i \right) \right| \\
&\quad + n^{-1} \left| 2(b_n - 1)\bar{B}_n^2 \right| .
\end{aligned}$$

Now we show that each of the three terms above tend to zero.

1. Since $U_1^2, \ldots U_n^2$ are i.i.d. χ_1^2, $\sum_{i=1}^{b_n-1} U_i^2 + \sum_{i=n-b_n+2}^{n} U_i^2 \sim \chi_{2(b_n-1)}^2$. By Lemma 20

we have

$$E\left[\left(\sum_{i=1}^{b_n-1} U_i^2 + \sum_{i=n-b_n+2}^{n} U_i^2 - 2(b_n-1)\right)^{2c}\right] \leq K\left(2(b_n-1)\right)^c \ ,$$

and

$$E\left[\left(n^{-1}\left(\sum_{i=1}^{b_n-1} U_i^2 + \sum_{i=n-b_n+2}^{n} U_i^2\right) - \frac{2(b_n-1)}{n}\right)^{2c}\right] \leq K\left(\frac{2(b_n-1)}{n^2}\right)^c \ .$$

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

$$\sum_n K\left(\frac{2(b_n-1)}{n^2}\right)^c < \infty \ .$$

A Borel-Cantelli argument results in

$$\left(n^{-1}\left(\sum_{i=1}^{b_n-1} U_i^2 + \sum_{i=n-b_n+2}^{n} U_i^2\right) - \frac{2(b_n-1)}{n}\right)^{2c} \to 0 \quad w.p.1 \text{ as } n \to \infty \ .$$

Hence

$$n^{-1}\left|\sum_{i=1}^{b_n-1} U_i^2 + \sum_{i=n-b_n+2}^{n} U_i^2\right| \to 0 \quad w.p.1 \text{ as } n \to \infty$$

since $b_n/n \to 0$ as $n \to \infty$.

2. Notice

$$n^{-1}\left|2\bar{B}_n\left(\sum_{i=1}^{b_n-1} U_i + \sum_{i=n-b_n+2}^{n} U_i\right)\right| \leq 2\left|\bar{B}_n\right| n^{-1}\sum_{i=1}^{n}|U_i| \ .$$

By the classical strong law, $n^{-1}\sum_{i=1}^{n}|U_i| \to \sqrt{2/\pi}$ w.p.1 since $|U_i|$ are i.i.d.

half-normal distributions. Combining this with Lemma 15 imply

$$n^{-1}\left|2\bar{B}_n\left(\sum_{i=1}^{b_n-1}U_i+\sum_{i=n-b_n+2}^{n}U_i\right)\right|\leq 4/\sqrt{\pi}n^{-1/2}(1+\epsilon)\left[\log\log n\right]^{1/2}\ ,$$

which tends to 0 as $n\to\infty$.

3. Using Lemma 15

$$n^{-1}\left|2(b_n-1)\bar{B}_n^2\right|\leq(1+\epsilon)^2 4n^{-2}\left[\log\log n\right](b_n-1)\ ,$$

which tends to 0 since $b_n/n\to 0$ as $n\to\infty$.

Next, consider (A.9)

$$\begin{aligned}
&n^{-1}\left|2\sum_{s=1}^{b_n-1}\sum_{l=1}^{b_n-s}b_n^{-1}\left(\sum_{i=1}^{l-1}T_iT_{s+i}+\sum_{i=n-b_n+l+1}^{n-s}T_iT_{s+i}\right)\right|\\
&\quad=n^{-1}\left|2\sum_{s=1}^{b_n-1}\sum_{l=1}^{b_n-s}b_n^{-1}\left(\sum_{i=1}^{l-1}\left(U_i-\bar{B}_n\right)\left(U_{s+i}-\bar{B}_n\right)\right.\right.\\
&\qquad\left.\left.+\sum_{i=n-b_n+l+1}^{n-s}\left(U_i-\bar{B}_n\right)\left(U_{s+i}-\bar{B}_n\right)\right)\right|\\
&\quad\leq n^{-1}\left|2\sum_{s=1}^{b_n-1}\sum_{l=1}^{b_n-s}b_n^{-1}\left(\sum_{i=1}^{l-1}U_iU_{s+i}+\sum_{i=n-b_n+l+1}^{n-s}U_iU_{s+i}\right)\right|\\
&\qquad+n^{-1}\left|2\sum_{s=1}^{b_n-1}\sum_{l=1}^{b_n-s}\bar{B}_nb_n^{-1}\left(\sum_{i=1}^{l-1}(-U_i-U_{s+i})+\sum_{i=n-b_n+l+1}^{n-s}(-U_i-U_{s+i})\right)\right|\\
&\qquad+n^{-1}\left|2\sum_{s=1}^{b_n-1}\sum_{l=1}^{b_n-s}\bar{B}_n^2b_n^{-1}2(b_n-1)\right|\ .
\end{aligned}$$

Again, we show that each of the three terms above tend to zero.

1. Let $C(b_n)$ be the $(b_n - 1) \times (b_n - 1)$ matrix with entries,

$$C_{i,j}(b_n) = \begin{cases} 2n^{-1}b_n^{-1}(b_n - i) & \text{for } i > j \ , \\ 0 & \text{otherwise.} \end{cases}$$

Further let $D(b_n)$ be the $(b_n - 1) \times (b_n - 1)$ matrix with entries,

$$D_{i,j}(b_n) = \begin{cases} 2n^{-1}b_n^{-1}j & \text{for } i > j \ , \\ 0 & \text{otherwise.} \end{cases}$$

Define $U := (U_1, \ldots, U_{b_n-1}, U_{n-b_n+2}, \ldots, U_n)^T$, then by construction

$$\begin{aligned} A(b_n) &= U' \begin{pmatrix} C & 0 \\ 0 & D \end{pmatrix} U \\ &= n^{-1}2 \sum_{s=1}^{b_n-1} \sum_{l=1}^{b_n-s} b_n^{-1} \left(\sum_{i=1}^{l-1} U_i U_{s+i} + \sum_{i=n-b_n+l+1}^{n-s} U_i U_{s+i} \right) . \end{aligned}$$

Recall $U_1, \ldots, U_{b_n-1}, U_{n-b_n+2}, \ldots, U_n$ are i.i.d. $N(0,1)$ resulting in $EA(b_n) = 0$. Thus, we can apply Lemma 18

$$\begin{aligned} E \left[\left(n^{-1}2 \sum_{s=1}^{b_n-1} \sum_{l=1}^{b_n-s} b_n^{-1} \left(\sum_{i=1}^{l-1} U_i U_{s+i} + \sum_{i=n-b_n+l+1}^{n-s} U_i U_{s+i} \right) \right)^{2c} \right] \\ \leq K(c) \left(\sum_j \sum_k a_{jk}^2 \right)^c \end{aligned}$$

where the coefficients a_{jk} are the elements of the matrix

$$\begin{pmatrix} C & 0 \\ 0 & D \end{pmatrix}$$

and $K(c)$ is a constant in c. Since $b_n^{-1}(b_n - i) < 1$ for all i in C and $b_n^{-1} j < 1$ for all j in D we have

$$\sum_j \sum_k a_{jk}^2 \leq n^{-2} 4 \sum_{i=1}^{b_n-1} \sum_{j=1}^{b_n-1} 1^2 = n^{-2}(b_n - 1)^2.$$

Hence,

$$E\left[\left(n^{-1} 2 \sum_{s=1}^{b_n-1} \sum_{l=1}^{b_n-s} b_n^{-1} \left(\sum_{i=1}^{l-1} U_i U_{s+i} + \sum_{i=n-b_n+l+1}^{n-s} U_i U_{s+i}\right)\right)^{2c}\right] \leq K(c) \left(\frac{b_n - 1}{n}\right)^{2c}.$$

Then by a Borel-Cantelli argument,

$$n^{-1} \left| 2 \sum_{s=1}^{b_n-1} \sum_{l=1}^{b_n-s} b_n^{-1} \left(\sum_{i=1}^{l-1} U_i U_{s+i} + \sum_{i=n-b_n+l+1}^{n-s} U_i U_{s+i}\right)\right|$$

tends to 0 as $n \to \infty$ since there exists constant $c \geq 1$ such that $\sum_n (b_n/n)^c < \infty$.

2. Notice

$$
\begin{aligned}
& n^{-1}\left|2\sum_{s=1}^{b_n-1}\sum_{l=1}^{b_n-s}\bar{B}_n b_n^{-1}\left(\sum_{i=1}^{l-1}(-U_i-U_{s+i})+\sum_{i=n-b_n+l+1}^{n-s}(-U_i-U_{s+i})\right)\right| \\
& \leq n^{-1}2\sum_{s=1}^{b_n-1}\sum_{l=1}^{b_n-s}|\bar{B}_n|b_n^{-1}\left(\sum_{i=1}^{l-1}(|U_i|+|U_{s+i}|)+\sum_{i=n-b_n+l+1}^{n-s}(|U_i|+|U_{s+i}|)\right) \\
& \leq n^{-1}2\sum_{s=1}^{b_n-1}\sum_{l=1}^{b_n-s}|\bar{B}_n|b_n^{-1}\left(2\sum_{i=1}^{b_n}|U_i|+2\sum_{i=n-b_n+1}^{n}|U_i|\right) \\
& < n^{-1}2b_n^2|\bar{B}_n|b_n^{-1}\left(2\sum_{i=1}^{b_n}|U_i|+2\sum_{i=n-b_n+1}^{n}|U_i|\right) \\
& = n^{-1}4b_n^2|\bar{B}_n|b_n^{-1}\left(\sum_{i=1}^{b_n}|U_i|+\sum_{i=n-b_n+1}^{n}|U_i|\right) .
\end{aligned}
$$

Then from Lemma 15

$$
n^{-1}4b_n^2|\bar{B}_n|b_n^{-1}\left(\sum_{i=1}^{b_n}|U_i|+\sum_{i=n-b_n+1}^{n}|U_i|\right)\leq n^{-3/2}4b_n^2(1+\epsilon)\left[2\log\log n\right]^{1/2}K \tag{A.10}
$$

where $K=b_n^{-1}\left(\sum_{i=1}^{b_n}|U_i|+\sum_{i=n-b_n+1}^{n}|U_i|\right)$. If K stays bounded w.p.1 then (A.10) tends to 0 as $n\to\infty$ since $n^{-3/2}b_n^2\to 0$ as $n\to\infty$.

Hence, it suffices to show that K stays bounded w.p.1. First notice the classical strong law of large numbers imples $b_n^{-1}\sum_{i=1}^{b_n}|U_i|$ stays bounded w.p.1. We will show $b_n^{-1}\sum_{i=n-b_n+1}^{n}|U_i|$ stays bounded using the following strong invariance principle. Komlós et al. (1975), Komlós et al. (1976), and Major (1976) have shown in the i.i.d. case, that if $E[\exp|tX_1|]<\infty$ in a neighborhood of $t=0$, then

$$
S_n-n\mu=\sigma B(n)+O(\log n)\quad w.p.1
$$

where the $\log n$ rate is extremely sharp. We will apply this to the i.i.d. sequence $\{|U_i|:i\geq 1\}$ with a half-normal distribution. (Recall the $E|U_i|=\sqrt{2/\pi}$ and

$\mathrm{Var}|U_i| = 1 - 2/\pi$.) Then

$$\left| \sum_{i=1}^{n} |U_i| - n\sqrt{2/\pi} - (1 - 2/\pi)B(n) \right| \leq C' \log n \ ,$$

resulting in

$$\begin{aligned}
b_n^{-1} \sum_{i=n-b_n+1}^{n} |U_i| &= b_n^{-1} \left| \sum_{i=1}^{n} |U_i| - \sum_{i=1}^{n-b_n} |U_i| \right| \\
&= b_n^{-1} \Bigg| \left(\sum_{i=1}^{n} |U_i| - n\sqrt{2/\pi} - (1 - 2/\pi)B(n) \right) \\
&\quad - \left(\sum_{i=1}^{n-b_n} |U_i| - (n - b_n)\sqrt{2/\pi} - (1 - 2/\pi)B(n - b_n) \right) \\
&\quad + (1 - 2/\pi)\left(B(n) - B(n - b_n)\right) + b_n\sqrt{2/\pi} \Bigg| \\
&\leq b_n^{-1} \left(2C' \log n + (1 + \epsilon) \left(2b_n \left(\log \frac{n}{b_n} + \log\log n \right) \right)^{1/2} + b_n\sqrt{2/\pi} \right) \\
&= \sqrt{2/\pi} + 2C' b_n^{-1} \log n + O\left((b_n^{-1} \log n)^{1/2} \right) \quad w.p.1.
\end{aligned}$$

Hence, $b_n^{-1} \sum_{i=n-b_n+1}^{n} |U_i|$ stays bounded w.p.1 since $b_n^{-1} \log n$ stays bounded as $n \to \infty$.

3. Using Lemma 15 we have

$$\begin{aligned}
n^{-1} \left| 2 \sum_{s=1}^{b_n - 1} \sum_{l=1}^{b_n - s} \bar{B}_n^2 b_n^{-1} 2(b_n - 1) \right| &\leq n^{-1} 4 \bar{B}_n^2 b_n^2 \\
&\leq n^{-2} 8(1 + \epsilon)^2 \left[\log\log n \right] b_n^2 \ ,
\end{aligned}$$

which tends to 0 as $n \to \infty$ since $b_n/n \to 0$ as $n \to \infty$.

□

A.1.3 Results for Batch Means

Recall that $B = \{B(t), t \geq 0\}$ denotes a standard Brownian motion. Define

$$\tilde{\sigma}^2_{BM} = \frac{b_n}{a_n - 1} \sum_{k=0}^{a_n - 1} (\bar{B}_k - \bar{B}_n)^2$$

where $\bar{B}_k = b^{-1}(B((k+1)b) - B(kb))$ for $k = 0, \ldots, a_n - 1$ and $\bar{B}_n = n^{-1}B(n)$.

Lemma 22. *(Damerdji, 1994, Proposition 3.1) Assume Assumption 2 and there exists a constant $c \geq 1$ such that $\sum_n (b_n/n)^c < \infty$, then as $n \to \infty$, $\tilde{\sigma}^2_{BM} \to 1$ w.p.1.*

Proof. Notice that

$$\begin{aligned}
\tilde{\sigma}^2_{BM} &= \frac{b_n}{a_n - 1} \sum_{k=0}^{a_n - 1} (\bar{B}_k - \bar{B}_n)^2 \\
&= \frac{a_n}{a_n - 1} \left(\frac{1}{a_n} \sum_{k=0}^{a_n - 1} b_n \bar{B}_k^2 - b_n \bar{B}_n^2 \right) .
\end{aligned}$$

Using this representation, we will the first term above goes to 1 while the second terms tends to 0 as $n \to \infty$.

1. Properties of Brownian motion imply for all $k = 0, \ldots, a_n - 1$ that $\bar{B}_k$ are i.i.d. $N(0, 1/b_n)$ and hence $b_n \bar{B}_k^2$ are i.i.d. χ^2_1. Then $\sum_{k=0}^{a_n - 1} b_n \bar{B}_k^2 \sim \chi^2_{a_n}$. Therefore, by Lemma 20 we have

$$E\left[\left(\sum_{k=0}^{a_n - 1} b_n \bar{B}_k^2 - a_n \right)^{2c} \right] \leq K \left(a_n\right)^c ,$$

and

$$E\left[\left(\frac{1}{a_n} \sum_{k=0}^{a_n - 1} b_n \bar{B}_k^2 - 1 \right)^{2c} \right] \leq K \left(\frac{b_n}{n} \right)^c .$$

Then by a Borel-Cantelli argument,

$$\frac{1}{a_n}\sum_{k=0}^{a_n-1} b_n\bar{B}_k^2 \to 1 \quad w.p.1 \text{ as } n\to\infty \ .$$

2. Similar properties of Brownian motion imply $n\bar{B}_n^2 \sim \chi_1^2$, hence $E\left[b_n\bar{B}_n^2\right] = b_n/n$. By a Borel-Cantelli argument, if $\sum_n (b_n/n) < \infty$, then $b_n\bar{B}_n^2 \to 0$ as $n\to\infty$ w.p.1.

   More generally, we can use Lemma 20 to show

$$E\left[\left(n\bar{B}_n^2 - 1\right)^{2c}\right] \le K \ ,$$

   and therefore

$$E\left[\left(b_n\bar{B}_n^2 - \frac{b_n}{n}\right)^{2c}\right] \le K\left(\frac{b_n}{n}\right)^{2c} \ .$$

   Then by a Borel-Cantelli argument, $b_n\bar{B}_n^2 \to 0$ as $n\to\infty$ w.p.1.

□

A.1.4 Results for Mean Square Consistency

Consider the Brownian motion estimator for OBM,

$$\tilde{\sigma}_{OBM}^2 = \frac{nb_n}{(n-b_n)(n-b_n+1)}\sum_{j=0}^{n-b_n}(\bar{B}_j(b_n) - \bar{B}_n)^2 \ .$$

Further define, the Brownian motion estimator for BM,

$$\tilde{\sigma}_{BM}^2 = \frac{b_n}{a_n-1}\sum_{k=0}^{a_n-1}(\bar{B}_k - \bar{B}_n)^2 \ ,$$

where $\bar{B}_k := b_n^{-1}\left(B((k+1)b_n) - B(kb_n+1)\right)$ for $k = 0,\ldots,a_n-1$.

Lemma 23. *(Damerdji, 1995, p. 285 and Lemma 2) Suppose Assumption 2 holds, then*

$$E\left[\tilde{\sigma}^2_{OBM}\right] = E\left[\tilde{\sigma}^2_{BM}\right] = 1\ , \tag{A.11}$$

$$\frac{n}{b_n} Var\left[\tilde{\sigma}^2_{OBM}\right] = \frac{4}{3} + o\left(1\right)\ ,\ and \tag{A.12}$$

$$\frac{n}{b_n} Var\left[\tilde{\sigma}^2_{BM}\right] = 2 + o\left(1\right)\ . \tag{A.13}$$

Proof. The proof for $\tilde{\sigma}^2_{BM}$ is straight forward and is therefore omitted.

First, define $U_i := B(i) - B(i-1)$ as the increments of Brownian motion between times i and $i-1$ and recall that U_i are i.i.d. $N(0,1)$ for all $i = 1, \ldots, n$. Notice that $\bar{B}_j(b_n) - \bar{B}_n$ can be written as a linear combination of i.i.d. normal distributions for all $j = 1, \ldots, n-b+1$,

$$\bar{B}_j(b_n) - \bar{B}_n = \frac{(n-b)}{nb} \sum_{i=j}^{j+b} U_i - \frac{1}{n} \sum_{i=1}^{j-1} U_i - \frac{1}{n} \sum_{i=j+b+1}^{n} U_i\ ,$$

where any empty sums are defined to be zero.

Then it is easy to see the $E[\bar{B}_j(b_n) - \bar{B}_n] = 0$ for all $j = 1, \ldots, n-b+1$ and

$$\begin{aligned} \text{Var}\left[\bar{B}_j(b_n) - \bar{B}_n\right] &= \text{Var}\left[\frac{(n-b)}{nb} \sum_{i=j}^{j+b} U_i - \frac{1}{n} \sum_{i=1}^{j-1} U_i - \frac{1}{n} \sum_{i=j+b+1}^{n} U_i\right] \\ &= \left(\frac{n-b}{nb}\right)^2 b + \frac{n-b}{n^2} \\ &= \frac{n-b}{bn} \end{aligned} \tag{A.14}$$

for all $j = 1, \ldots, n-b+1$.

Now we can calculate $E\left[\tilde{\sigma}^2_{OBM}\right]$

$$\begin{aligned} E\left[\tilde{\sigma}^2_{OBM}\right] &= E\left[\frac{nb}{(n-b)(n-b+1)}\sum_{j=1}^{n-b+1}(\bar{B}_j(b_n)-\bar{B}_n)^2\right] \\ &= \frac{nb}{(n-b)(n-b+1)}\sum_{j=1}^{n-b+1}E\left[(\bar{B}_j(b_n)-\bar{B}_n)^2\right] \\ &= \frac{nb}{(n-b)(n-b+1)}\sum_{j=1}^{n-b+1}\frac{n-b}{bn} \\ &= 1\,. \end{aligned} \tag{A.15}$$

Calculating $\mathrm{Var}\left[\tilde{\sigma}^2_{OBM}\right]$ will require first calculating $E\left[\tilde{\sigma}^4_{OBM}\right]$. First notice that,

$$\begin{aligned} \tilde{\sigma}^4_{OBM} &= \frac{n^2b^2}{(n-b)^2(n-b+1)^2}\left[\sum_{j=1}^{n-b+1}(\bar{B}_j(b_n)-\bar{B}_n)^2\right]^2 \\ &= \frac{n^2b^2}{(n-b)^2(n-b+1)^2}\left[\sum_{j=1}^{n-b+1}(\bar{B}_j(b_n)-\bar{B}_n)^4\right. && \text{(A.16)} \\ &\quad + 2\sum_{s=1}^{b-1}\sum_{j=1}^{n-b+1-s}(\bar{B}_j(b_n)-\bar{B}_n)^2(\bar{B}_{j+s}(b_n)-\bar{B}_n)^2 && \text{(A.17)} \\ &\quad \left.+2\sum_{s=b}^{n-b}\sum_{j=1}^{n-b+1-s}(\bar{B}_j(b_n)-\bar{B}_n)^2(\bar{B}_{j+s}(b_n)-\bar{B}_n)^2\right]\,. && \text{(A.18)} \end{aligned}$$

Then we can let A be the summation in (A.16), B be the double summation in (A.17), and C be the double summation in (A.18). We will calculate the expectation of these separately. First recall that $\bar{B}_j(b_n)-\bar{B}_n \sim N(0,(n-b)/bn)$ from (A.14), then it's

easy to see

$$\begin{aligned} EA &= \sum_{j=1}^{n-b+1} E\left[(\bar{B}_j(b_n) - \bar{B}_n)^4\right] \\ &= \sum_{j=1}^{n-b+1} 3\left(\frac{n-b}{bn}\right)^2 \\ &= 3(n-b+1)\left(\frac{n-b}{bn}\right)^2 . \end{aligned} \tag{A.19}$$

We will calculate EB by first calculating

$$E\left[(\bar{B}_j(b_n) - \bar{B}_n)^2(\bar{B}_{j+s}(b_n) - \bar{B}_n)^2\right] = E\left[Z_1^2 Z_2^2\right]$$

where $Z_1 := (\bar{B}_j(b_n) - \bar{B}_n)$ and $Z_2 := (\bar{B}_{j+s}(b_n) - \bar{B}_n)$ for all $j = 1, \ldots, n-b+1-s$ and $s = 1, \ldots, b-1$. Notice that both Z_1 and Z_2 are linear combinations of i.i.d. standard normal random variables. We will calculate the joint normal distribution of $Z := (Z_1, Z_2)^T$. Recall that $U := (U_1, \ldots, U_n)^T$, then

$$U \sim N(0, I_n) \ ,$$

and if D is a $m \times n$ matrix, then the random vector $Z = DU$ is jointly normal with

$$Z \sim N\left(0, DD^T\right) \ .$$

Then we can calculate the distribution of Z as

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma\right) \ ,$$

where

$$\Sigma = \begin{pmatrix} \frac{n-b}{bn} & (b-s)\left(\frac{n-b}{nb}\right)^2 - \frac{2s}{n}\left(\frac{n-b}{nb}\right) + \frac{n-b-s}{n^2} \\ (b-s)\left(\frac{n-b}{nb}\right)^2 - \frac{2s}{n}\left(\frac{n-b}{nb}\right) + \frac{n-b-s}{n^2} & \frac{n-b}{bn} \end{pmatrix}$$
$$= \begin{pmatrix} \frac{n-b}{bn} & \frac{nb-ns-b^2}{nb^2} \\ \frac{nb-ns-b^2}{nb^2} & \frac{n-b}{bn} \end{pmatrix} .$$

To calculate the desired expectation, we will iterate the expectation. To this end, we will need to calculate the conditional distribution of $Z_1|Z_2$ and the marginal distribution of Z_2. Recall that if

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right) ,$$

and Σ_{22} is non-singular, then the conditional distribution of $X_1|X_2$ is

$$X_1|X_2 \sim N\left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}\left(X_2 - \mu_2\right), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right) .$$

In our case, $\frac{n-b}{bn}$ is clearly non-singular since $b < n$, and we can calculate the conditional distribution of $Z_1|Z_2$ and the marginal distribution of Z_2 as

$$Z_1|Z_2 \sim N\left(\frac{b(n-b)-ns}{b(n-b)}Z_2, \frac{2bs(n-b)-ns^2}{b^3(n-b)}\right)$$
$$Z_2 \sim N\left(0, \frac{n-b}{bn}\right) .$$

Now we can calculate the desired expectation using an iterated expectation as follows

$$\begin{aligned}
E\left[Z_1^2 Z_2^2\right] &= E_{Z_2}\left[E_{Z_1|Z_2}\left[Z_1^2 Z_2^2 \mid Z_2\right]\right] \\
&= E_{Z_2}\left[Z_2^2 E_{Z_1|Z_2}\left[Z_1^2 \mid Z_2\right]\right] \\
&= E_{Z_2}\left[Z_2^2\left(\left(\frac{b(n-b)-ns}{b(n-b)}Z_2\right)^2 + \frac{2bs(n-b)-ns^2}{b^3(n-b)}\right)\right] \\
&= \left(\frac{b(n-b)-ns}{b(n-b)}\right)^2 E_{Z_2}\left[Z_2^4\right] + \frac{2bs(n-b)-ns^2}{b^3(n-b)} E_{Z_2}\left[Z_2^2\right] \\
&= \left(\frac{b(n-b)-ns}{b(n-b)}\right)^2 3\left(\frac{n-b}{bn}\right)^2 + \frac{2bs(n-b)-ns^2}{b^3(n-b)}\left(\frac{n-b}{bn}\right) \\
&= \frac{3(b(n-b)-ns)^2}{b^4 n^2} + \frac{2bs(n-b)-ns^2}{b^4 n} \\
&= \frac{3(b(n-b)-ns)^2 + 2nbs(n-b)-n^2 s^2}{b^4 n^2}\,.
\end{aligned}$$

Next, we can calculate

$$\begin{aligned}
EB &= E\left[\sum_{s=1}^{b-1}\sum_{j=1}^{n-b+1-s}(\bar{B}_j(b_n)-\bar{B}_n)^2(\bar{B}_{j+s}(b_n)-\bar{B}_n)^2\right] \\
&= \sum_{s=1}^{b-1}\sum_{j=1}^{n-b+1-s} E\left[(\bar{B}_j(b_n)-\bar{B}_n)^2(\bar{B}_{j+s}(b_n)-\bar{B}_n)^2\right] \\
&= \sum_{s=1}^{b-1}\sum_{j=1}^{n-b+1-s} E\left[Z_1^2 Z_2^2\right] \\
&= \sum_{s=1}^{b-1}\sum_{j=1}^{n-b+1-s}\left[\frac{3(b(n-b)-ns)^2+2nbs(n-b)-n^2s^2}{b^4n^2}\right] \\
&= \sum_{s=1}^{b-1}(n-b+1-s)\left[\frac{3(b(n-b)-ns)^2+2nbs(n-b)-n^2s^2}{b^4n^2}\right] \\
&= \frac{1}{b^4n^2}\sum_{s=1}^{b-1}(n-b+1-s)\left[3b^2(n-b)^2-4nbs(n-b)+2n^2s^2\right] \\
&= \frac{1}{b^4n^2}\left[3b^2(n-b)^2(n-b+1)(b-1)-4nb(n-b)(n-b+1)\sum_{s=1}^{b-1}s\right. \\
&\qquad + 2n^2(n-b+1)\sum_{s=1}^{b-1}s^2-3b^2(n-b)^2\sum_{s=1}^{b-1}s \\
&\qquad \left. + 4nb(n-b)\sum_{s=1}^{b-1}s^2-2n^2\sum_{s=1}^{b-1}s^3\right] .
\end{aligned}$$

Then by mathematical fact

$$\begin{aligned}
\sum_{s=1}^{b-1}s &= \frac{(b-1)b}{2} , \\
\sum_{s=1}^{b-1}s^2 &= \frac{(b-1)b(2b-1)}{6} , \text{ and} \\
\sum_{s=1}^{b-1}s^3 &= \frac{(b-1)^2b^2}{4} .
\end{aligned}$$

Resulting in

$$\begin{aligned}
EB &= \frac{1}{b^4n^2}\Bigg[3b^2(n-b)^2(n-b+1)(b-1) - 4nb(n-b)(n-b+1)\sum_{s=1}^{b-1} s \\
&\quad + 2n^2(n-b+1)\sum_{s=1}^{b-1} s^2 - 3b^2(n-b)^2\sum_{s=1}^{b-1} s \\
&\quad + 4nb(n-b)\sum_{s=1}^{b-1} s^2 - 2n^2\sum_{s=1}^{b-1} s^3\Bigg] \\
&= \frac{1}{b^4n^2}\Bigg[3b^2(n-b)^2(n-b+1)(b-1) - 4nb(n-b)(n-b+1)\frac{(b-1)b}{2} \\
&\quad + 2n^2(n-b+1)\frac{(b-1)b(2b-1)}{6} - 3b^2(n-b)^2\frac{(b-1)b}{2} \\
&\quad + 4nb(n-b)\frac{(b-1)b(2b-1)}{6} - 2n^2\frac{(b-1)^2b^2}{4}\Bigg] \\
&= \frac{1}{6b^3n^2}\left[18b(n-b)^2(n-b+1)(b-1) - 12nb(n-b)(n-b+1)(b-1)\right. \\
&\quad + 2n^2(n-b+1)(b-1)(2b-1) - 9b^2(n-b)^2(b-1) \\
&\quad \left. + 4nb(n-b)(b-1)(2b-1) - 3n^2b(b-1)^2\right] .
\end{aligned}$$

Now we are going to rearrange terms from 2 parts of the above expression

$$\begin{aligned}
&2n^2(n-b+1)(b-1)(2b-1) - 3n^2b(b-1)^2 \\
&\quad = 2n^3(b-1)(2b-1) - 2n^2(b-1)^2(2b-1) - 3n^2b(b-1)^2 \\
&\quad = 2n^3(b-1)(2b-1) - n^2(b-1)^2(7b-2) .
\end{aligned}$$

We will also rearrange the 4 remaining terms,

$$
\begin{aligned}
&18b(n-b)^2(n-b+1)(b-1) - 12nb(n-b)(n-b+1)(b-1) \\
&\qquad - 9b^2(n-b)^2(b-1) + 4nb(n-b)(b-1)(2b-1) \\
&= 18b(n-b)(n-b+1)^2(b-1) - 18b(n-b)(n-b+1)(b-1) \\
&\qquad - 12b(n-b)(n-b+1)^2(b-1) - 12b(n-b)(n-b+1)(b-1)^2 \\
&\qquad - 9b^2(n-b)^2(b-1) + 4nb(n-b)(b-1)(2b-1) \\
&= 6b(n-b)(n-b+1)^2(b-1) - 18b(n-b)(n-b+1)(b-1) \\
&\qquad - 12b(n-b)(n-b+1)(b-1)^2 \\
&\qquad - 9b^2(n-b)^2(b-1) + 4nb(n-b)(b-1)(2b-1) \\
&= 6b(n-b)(n-b+1)^2(b-1) - 18b(n-b)(n-b+1)(b-1) \\
&\qquad - 12b^2(n-b)(n-b+1)(b-1) + 12b(n-b)(n-b+1)(b-1) \\
&\qquad - 9b^2(n-b)(n-b+1)(b-1) + 9b^2(n-b)(b-1) \\
&\qquad + 8b^2(n-b)(n-b+1)(b-1) + 8b^2(n-b)(b-1)^2 - 4nb(n-b)(b-1) \\
&= 6b(n-b)(n-b+1)^2(b-1) - 13b^2(n-b)(n-b+1)(b-1) \\
&\qquad + 9b^2(n-b)(b-1) + 8b^2(n-b)(b-1)^2 \\
&\qquad - 4nb(n-b)(b-1) - 6b(n-b)(n-b+1)(b-1) \\
&= 6b(n-b)(n-b+1)^2(b-1) - 13b^2(n-b)(n-b+1)(b-1) \\
&\qquad + 9b^2(n-b)(b-1) + 8b^2(n-b)(b-1)^2 \\
&\qquad - 4b(n-b)^2(b-1) - 4b^2(n-b)(b-1) \\
&\qquad - 6b(n-b)^2(b-1) - 6b(n-b)(b-1) \\
&= 6b(n-b)(n-b+1)^2(b-1) - 13b^2(n-b)(n-b+1)(b-1) \\
&\qquad - 10b(n-b)^2(b-1) + b(n-b)(b-1)\left[9b + 8b(b-1) - 4b - 6\right] \\
&= 6b(n-b)(n-b+1)^2(b-1) - 13b^2(n-b)(n-b+1)(b-1) \\
&\qquad - 10b(n-b)^2(b-1) + b(n-b)(b-1)\left[8b^2 - 3b - 6\right] \ .
\end{aligned}
$$

Then we can present the result as in Damerdji (1995),

$$\begin{aligned}
EB = \frac{1}{6b^3n^2}\big[&18b(n-b)^2(n-b+1)(b-1) - 12nb(n-b)(n-b+1)(b-1) \\
&+ 2n^2(n-b+1)(b-1)(2b-1) - 9b^2(n-b)^2(b-1) \\
&+ 4nb(n-b)(b-1)(2b-1) - 3n^2b(b-1)^2\big] \\
= \frac{1}{6b^3n^2}\big[&6b(n-b)(n-b+1)^2(b-1) - 13b^2(n-b)(n-b+1)(b-1) \\
&- 10b(n-b)^2(b-1) + b(n-b)(b-1)\left[8b^2-3b-6\right] \\
&+2n^3(b-1)(2b-1) - n^2(b-1)^2(7b-2)\big] \ . \qquad \text{(A.20)}
\end{aligned}$$

Similar to the previous expectation, we will calculate the EC by first calculating

$$E\left[(\bar{B}_j(b_n) - \bar{B}_n)^2(\bar{B}_{j+s}(b_n) - \bar{B}_n)^2\right] = E\left[Z_1^2 Z_2^2\right]$$

where $Z_1 = (\bar{B}_j(b_n) - \bar{B}_n)$ and $Z_2 = (\bar{B}_{j+s}(b_n) - \bar{B}_n)$ for all $j = 1, \ldots, n-b+1-s$ and $s = b, \ldots, n-b$. Notice that both Z_1 and Z_2 are linear combinations of i.i.d. standard normal random variables. We can again calculate the joint normal distribution of Z_1 and $Z-2$ as

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{n-b}{bn} & -\frac{1}{n} \\ -\frac{1}{n} & \frac{n-b}{bn} \end{pmatrix}\right) \ ,$$

resulting in the conditional distribution of $Z_1|Z_2$ and the marginal distribution of Z_2 as

$$\begin{aligned}
Z_1|Z_2 &\sim N\left(\frac{-b}{n-b}Z_2, \left[\frac{n-b}{bn} - \frac{bn}{n^2(n-b)}\right]\right) \\
Z_2 &\sim N\left(0, \frac{n-b}{bn}\right) \ .
\end{aligned}$$

Then we can calculate the desired expectation using an iterated expectation as follows

$$\begin{aligned}
E\left[Z_1^2 Z_2^2\right] &= E_{Z_2}\left[E_{Z_1|Z_2}\left[Z_1^2 Z_2^2 \mid Z_2\right]\right] \\
&= E_{Z_2}\left[Z_2^2 E_{Z_1|Z_2}\left[Z_1^2 \mid Z_2\right]\right] \\
&= E_{Z_2}\left[Z_2^2\left(\left(\frac{-b}{n-b}Z_2\right)^2 + \left[\frac{n-b}{bn} - \frac{bn}{n^2(n-b)}\right]\right)\right] \\
&= \left(\frac{-b}{n-b}\right)^2 E_{Z_2}\left[Z_2^4\right] + \left[\frac{n-b}{bn} - \frac{bn}{n^2(n-b)}\right] E_{Z_2}\left[Z_2^2\right] \\
&= \left(\frac{-b}{n-b}\right)^2 3\left(\frac{n-b}{bn}\right)^2 + \left[\frac{n-b}{bn} - \frac{bn}{n^2(n-b)}\right]\frac{n-b}{bn} \\
&= \frac{2}{n^2} + \left(\frac{n-b}{bn}\right)^2
\end{aligned}$$

Finally, we can calculate the EC,

$$\begin{aligned}
EC &= E\left[\sum_{s=b}^{n-b}\sum_{j=1}^{n-b+1-s}(\bar{B}_j(b_n) - \bar{B}_n)^2(\bar{B}_{j+s}(b_n) - \bar{B}_n)^2\right] \\
&= \sum_{s=b}^{n-b}\sum_{j=1}^{n-b+1-s} E\left[(\bar{B}_j(b_n) - \bar{B}_n)^2(\bar{B}_{j+s}(b_n) - \bar{B}_n)^2\right] \\
&= \sum_{s=b}^{n-b}\sum_{j=1}^{n-b+1-s} E\left[Z_1^2 Z_2^2\right] \\
&= \sum_{s=b}^{n-b}\sum_{j=1}^{n-b+1-s}\left[\frac{2}{n^2} + \left(\frac{n-b}{bn}\right)^2\right] \\
&= \left[\frac{2}{n^2} + \left(\frac{n-b}{bn}\right)^2\right]\sum_{s=b}^{n-b}(n-b+1-s) \\
&= \left[\frac{2}{n^2} + \left(\frac{n-b}{bn}\right)^2\right]\left[(n-2b+1)(n-b+1) - \sum_{s=b}^{n-b} s\right] \\
&= \left[\frac{2}{n^2} + \left(\frac{n-b}{bn}\right)^2\right]\left[(n-2b+1)(n-b+1) - \frac{n(n-2b+1)}{2}\right] .
\end{aligned}$$

We can further factor the second part into a more desireable form

$$\begin{aligned}
&\left[(n-2b+1)(n-b+1) - \frac{n(n-2b+1)}{2}\right] \\
&\quad = \left[(n-b+1)^2 - b(n-b+1) - \frac{n(n-b+1)}{2} + \frac{nb}{2}\right] \\
&\quad = \left[(n-b+1)^2 - b(n-b+1) - \frac{(n-b+1)^2}{2} - \frac{(b-1)(n-b+1)}{2} + \frac{nb}{2}\right] \\
&\quad = \left[\frac{(n-b+1)^2}{2} - b(n-b+1) - \frac{(b-1)(n-b+1)}{2} + \frac{nb}{2}\right] \\
&\quad = \left[\frac{(n-b+1)^2}{2} - \frac{(3b-1)(n-b+1) + b(n-b+1) + b(b-1)}{2}\right] \\
&\quad = \frac{1}{2}\left[(n-b+1)^2 - (2b-1)(n-b+1) + b(b-1)\right] \ .
\end{aligned}$$

Then we can plug in the above result to get

$$\begin{aligned}
EC &= \left[\frac{2}{n^2} + \left(\frac{n-b}{bn}\right)^2\right]\left[(n-2b+1)(n-b+1) - \frac{n(n-2b+1)}{2}\right] \\
&= \frac{1}{2}\left[\frac{2}{n^2} + \left(\frac{n-b}{bn}\right)^2\right]\left[(n-b+1)^2 - (2b-1)(n-b+1) + b(b-1)\right] \ . \quad \text{(A.21)}
\end{aligned}$$

Combining the results from (A.19), (A.20), and (A.21), we get

$$\begin{aligned}
E\left[\tilde{\sigma}^4_{OBM}\right] &= \frac{n^2b^2}{(n-b)^2(n-b+1)^2}\left[EA+2EB+2EC\right] \\
&= \frac{n^2b^2}{(n-b)^2(n-b+1)^2}\left[3(n-b+1)\left(\frac{n-b}{bn}\right)^2\right. \\
&\quad + \frac{1}{3b^3n^2}\left[6b(n-b)(n-b+1)^2(b-1) - 13b^2(n-b)(n-b+1)(b-1)\right. \\
&\quad - 10b(n-b)^2(b-1) + b(n-b)(b-1)\left[8b^2-3b-6\right] \\
&\quad \left.+2n^3(b-1)(2b-1) - n^2(b-1)^2(7b-2)\right] \\
&\quad \left.+\left[\frac{2}{n^2}+\left(\frac{n-b}{bn}\right)^2\right]\left[(n-b+1)^2-(2b-1)(n-b+1)+b(b-1)\right]\right] \\
&= 1 + \frac{2n^3(b-1)(2b-1)}{3m(n-b)^2(n-b+1)^2} + \frac{2(b-1)}{(n-b)} - \frac{2(b-2)}{(n-b+1)} + \frac{(b-1)(3b-10)}{3(n-b+1)^2} \\
&\quad + \frac{2b^2}{(n-b)^2} - \frac{13b(b-1)}{3(n-b)(n-b+1)} + \frac{(b-1)\left[8b^2-3b-6\right]}{3(n-b)(n-b+1)^2} \\
&\quad - \frac{2b^2(2b-1)}{(n-b)^2(n-b+1)} - \frac{n^2(b-1)^2(7b-2)}{3b(n-b)^2(n-b+1)^2} + \frac{2b^3(b-1)}{(n-b)^2(n-b+1)^2}\,,
\end{aligned}$$

and so

$$E\left[\tilde{\sigma}^4_{OBM}\right] = 1 + \frac{4b}{3n} + o\left(\frac{b}{n}\right)\,. \tag{A.22}$$

The "little-o" notation can be alternatively written as for all $\epsilon > 0$, there exists a $n_0(\epsilon)$ such that

$$\left|E\left[\tilde{\sigma}^4_{OBM}\right] - 1 - \frac{4b}{3n}\right| < \epsilon\left|\frac{b}{n}\right|\,,$$

for all $n > n_0(\epsilon)$.

Finally, we can calculate $\mathrm{Var}\left[\tilde{\sigma}^2_{OBM}\right]$ from (A.15) and (A.22)

$$\begin{aligned}\mathrm{Var}\left[\tilde{\sigma}^2_{OBM}\right] &= E\left[\tilde{\sigma}^4_{OBM}\right] - \left(E\left[\tilde{\sigma}^2_{OBM}\right]\right)^2 \\ &= 1 + \frac{4b}{3n} + o\left(\frac{b}{n}\right) - 1^2 \\ &= \frac{4b}{3n} + o\left(\frac{b}{n}\right) \ .\end{aligned}$$

□

References

Alexopoulos, C. and Goldsman, D. (2004). To batch or not to batch? *ACM Trans. Model. Comput. Simul.*, 14(1):76–114.

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*. John Wiley & Sons.

Anderson, T. W. (1994). *The Statistical Analysis of Time Series (Classics Edition)*. Wiley-Interscience.

Athreya, K. B., Doss, H., and Sethuraman, J. (1996). On the convergence of the Markov chain simulation method. *The Annals of Statistics*, 24(1):69–100.

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

Bertail, P. and Clémençon, S. (2006). Regenerative block-bootstrap for Markov chains. *Bernoulli*, 12:689–712.

Bratley, P., Fox, B. L., and Schrage, L. E. (1987). *A Guide to Simulation*. Springer–Verlag, New York.

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

Bühlmann, P. (2002). Bootstraps for time series. *Statistical Science*, 17:52–72.

Carlstein, E. (1986). Asymptotic normality for a general statistic from a stationary sequence. *The Annals of Probability*, 14:1371–1379.

Chen, M.-H. and Shao, Q.-M. (2001). Propriety of posterior distribution for dichotomous quantal response models. *Proceedings of the American Mathematical Society*, 129(1):293–302.

Chen, M.-H., Shao, Q.-M., and Ibrahim, J. G. (2000). *Monte Carlo Methods in Bayesian Computation*. Springer-Verlag Inc.

Chien, C.-H., Goldsman, D., and Melamed, B. (1997). Large-sample results for batch means. *Management Science*, 43:1288–1295.

Christensen, O. F., Moller, J., and Waagepetersen, R. P. (2001). Geometric ergodicity of Metropolis-Hastings algorithms for conditional simulation in generalized linear mixed models. *Methodology and Computing in Applied Probability*, 3:309–327.

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

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

Csáki, E. and Csörgő, M. (1995). On additive functionals of Markov chains. *Journal of Theoretical Probability*, 8:905–919.

Csörgő, M. and Révész, P. (1981). *Strong Approximations in Probability and Statistics*. Academic Press.

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.

Datta, S. and McCormick, W. P. (1993). Regeneration-based bootstrap for Markov chains. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 21:181–193.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge University Press.

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

Doukhan, P., Massart, P., and Rio, E. (1994). The functional central limit theorem for strongly mixing processes. *Annales de l'Institut Henri Poincare, Section B, Calcul des Probabilities et Statistique*, 30:63–82.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman & Hall Ltd.

Finley, A. O., Banerjee, S., and Carlin, B. P. (2007). spBayes: an R package for univariate and multivariate hierarchical point-referenced spatial models. *Journal of Statistical Software*, 19.

Fishman, G. S. (1996). *Monte Carlo: Concepts, Algorithms, and Applications*. Springer, New York.

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

Flegal, J. M. and Jones, G. L. (2008). Asymptotic properties for batch means and spectral variance estimators in Markov chain Monte Carlo. Technical report, University of Minnesota, School of Statistics.

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

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

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

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

Geyer, C. J. (1992). Practical Markov chain Monte Carlo (with discussion). *Statistical Science*, 7:473–511.

Geyer, C. J. (1999). Likelihood inference for spatial point processes. In Barndorff-Nielsen, O. E., Kendall, W. S., and van Lieshout, M. N. M., editors, *Stochastic Geometry: Likelihood and Computation*, pages 79–140. Chapman & Hall/CRC, Boca Raton.

Geyer, C. J. and Thompson, E. A. (1995). Annealing Markov chain Monte Carlo with applications to ancestral inference. *Journal of the American Statistical Association*, 90:909–920.

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.

Haran, M., Bhat, K., Molineros, J., and De Wolf, E. (2008). Estimating the risk of a crop epidemic from coincident spatiotemporal processes. Technical report, The Pennsylvania State University, Department of Statistics.

Hoaglin, D. C. and Andrews, D. F. (1975). The reporting of computation-based results in statistics. *The American Statistician*, 29:122–126.

Hobert, J. P. and Geyer, C. J. (1998). Geometric ergodicity of Gibbs and block Gibbs samplers for a hierarchical random effects model. *Journal of Multivariate Analysis*, 67:414–430.

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

Ibragimov, I. A. and Linnik, Y. V. (1971). *Independent and Stationary Sequences of Random Variables*. Walters-Noordhoff, The Netherlands.

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5:299–314.

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

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

Johnson, A. A. and Jones, G. L. (2008). Gibbs sampling for a Bayesian hierarchical version of the general linear mixed model. Technical report, University of Minnesota, School of Statistics.

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.

Kendall, M. G. and Stuart, A. (1977). *The Advanced Theory of Statistics. Vol. I: Distribution Theory (4th Ed); Vol. 2: Inference and Relationship (3rd Ed)*. Charles Griffin & Co.

Komlós, J., Major, P., and Tusnády, G. (1975). An approximation of partial sums of independent RVs and the sample DF. I (Ref: 76V34 p33-58). *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 32:111–131.

Komlós, J., Major, P., and Tusnády, G. (1976). An approximation of partial sums of independent RV's, and the sample DF. II (Ref: 75V32 p111-131). *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 34:33–58.

Latuszynski, K. (2008). MCMC (e-a)-approximation under drift condition with application to Gibbs samplers for a hierarchical random effects model. Technical report, Warsaw School of Economics, Department of Mathematical Statistics.

L'Ecuyer, P., Simard, R., Chen, E. J., and Kelton, W. D. (2002). An objected-oriented random-number package with many long streams and substreams. *Operations Research*, 50:1073–1075.

Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer, New York.

Liu, J. S. and Wu, Y. N. (1999). Parameter expansion for data augmentation. *Journal of the American Statistical Association*, 94:1264–1274.

Major, P. (1976). The approximation of partial sums of independent RV's. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 35:213–220.

Marchev, D. and Hobert, J. P. (2004). Geometric ergodicity of van Dyk and Meng's algorithm for the multivariate Student's t model. *Journal of the American Statistical Association*, 99:228–238.

Marinari, E. and Parisi, G. (1992). Simulated tempering: A new Monte Carlo scheme. *Europhysics Letters*, 19:451–458.

Meketon, M. S. and Schmeiser, B. (1984). Overlapping batch means: something for nothing? In *WSC '84: Proceedings of the 16th conference on Winter simulation*, pages 226–230, Piscataway, NJ, USA. IEEE Press.

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

Meyn, S. P. and Tweedie, R. L. (1993). *Markov Chains and Stochastic Stability*. Springer-Verlag, London.

Meyn, S. P. and Tweedie, R. L. (1994). Computable bounds for geometric convergence rates of Markov chains. *The Annals of Applied Probability*, 4:981–1011.

Mira, A. and Tierney, L. (2002). Efficiency and convergence properties of slice samplers. *Scandinavian Journal of Statistics*, 29:1–12.

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

Nummelin, E. (1984). *General Irreducible Markov Chains and Non-negative Operators*. Cambridge University Press, London.

Nummelin, E. and Tuominen, P. (1982). Geometric ergodicity of Harris recurrent Markov chains with applications to renewal theory. *Stochastic Processes and their Applications*, 12:187–202.

Philipp, W. and Stout, W. (1975). Almost sure invariance principles for partial sums of weakly dependent random variables. *Memoirs of the American Mathematical Society*, 2:1–140.

Politis, D. N. (2003). The impact of bootstrap methods on time series analysis. *Statistical Science*, 18:219–230.

Politis, D. N., Romano, J. P., and Wolf, M. (1999). *Subsampling*. Springer-Verlag Inc.

Priestley, M. B. (1981). *Spectral Analysis and Time Series. (Vol. 1): Univariate Series*. Academic Press.

Robert, C. P. (1995). Convergence control methods for Markov chain Monte Carlo algorithms. *Statistical Science*, 10:231–253.

Robert, C. P. and Casella, G. (1999). *Monte Carlo Statistical Methods*. Springer, New York.

Roberts, G. O. (1996). Markov chain concepts related to sampling algorithms. In Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. E., editors, *Markov Chain Monte Carlo in Practice*, pages 45–57. Chapman & Hall, Boca Raton.

Roberts, G. O. and Polson, N. G. (1994). On the geometric convergence of the Gibbs sampler. *Journal of the Royal Statistical Society,* Series B, 56:377–384.

Roberts, G. O. and Rosenthal, J. S. (1997). Geometric ergodicity and hybrid Markov chains. *Electronic Communications in Probability*, 2:13–25.

Roberts, G. O. and Rosenthal, J. S. (1999). Convergence of slice sampler Markov chains. *Journal of the Royal Statistical Society,* Series B, 61:643–660.

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

Roberts, G. O. and Tweedie, R. L. (1999). Bounds on regeneration times and convergence rates for Markov chains. *Stochastic Processes and their Applications*, 80:211–229. Corrigendum (2001) 91:337-338.

Roberts, G. O. and Tweedie, R. L. (2001). Corrigendum to "Bounds on regeneration times and convergence rates for Markov chains". *Stochastic Processes and their Applications*, 91:337–338.

Rosenthal, J. S. (1995). Minorization conditions and convergence rates for Markov chain Monte Carlo. *Journal of the American Statistical Association*, 90:558–566.

Rosenthal, J. S. (1996). Analysis of the Gibbs sampler for a model related to James-Stein estimators. *Statistics and Computing*, 6:269–275.

Roy, V. and Hobert, J. P. (2007). Convergence rates and asymptotic standard errors for Markov chain Monte Carlo algorithms for Bayesian probit regression. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 69(4):607–623.

Song, W. T. and Schmeiser, B. W. (1995). Optimal mean-squared-error batch sizes. *Management Science*, 41:110–123.

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

van Dyk, D. A. and Meng, X.-L. (2001). The art of data augmentation (with discussion). *Journal of Computational and Graphical Statistics*, 10:1–50.

Welch, P. D. (1987). On the relationship between batch means, overlapping means and spectral estimation. In *WSC '87: Proceedings of the 19th conference on Winter simulation*, pages 320–323, New York, NY, USA. ACM.

Zeidler, E. (1990). *Nonlinear functional analysis and its applications. II/B: Nonlinear monotone operators*. Spring-Verlag, New York, second edition.