UNIVERSITY OF CALIFORNIA
RIVERSIDE

Visualizing Monte Carlo Error and Terminating Markov Chain Monte Carlo Simulation

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

Doctor of Philosophy

in

Applied Statitstics

by

Nathan Lane Robertson

September 2019

Dissertation Committee:

Dr. James M. Flegal, Chairperson
Dr. Subir Ghosh
Dr. Evangelos Papalexakis

The Dissertation of Nathan Lane Robertson is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

The text of this dissertation, in part, contains material from the paper "New Visualizations for Monte Carlo Simulations." The co-author James Flegal listed in that work directed and supervised the research which forms the basis for this dissertation. The co-author Dootika Vats provided guidance and an example. The co-author Galin Jones provided editing and guidance on assumptions.

This work has been in part supported by Mathematical Research Corporation.

To tea and coffee for all the support.

ABSTRACT OF THE DISSERTATION

Visualizing Monte Carlo Error and Terminating Markov Chain Monte Carlo Simulation

by

Nathan Lane Robertson

Doctor of Philosophy, Graduate Program in Applied Statitstics
University of California, Riverside, September 2019
Dr. James M. Flegal, Chairperson

Markov chain Monte Carlo (MCMC) is a sampling technique that allows for estimating features of intractable probability distributions. Output analysis of MCMC samples aims to assess the quality of the sampler and the resulting estimates. We provide an example based overview of current best practices using visualizations and the estimation of features of interest using Markov chain central limit theorems. Most features of interest are functionals of expectations or quantiles. The estimation of quantiles has thus-far been limited to univariate theory and marginal limiting distributions, while the estimation of expectations enjoys multivariate approaches through a joint limiting distribution. In this work, we provide an extension to jointly estimate combinations of functionals of expectations and quantiles through a joint limiting distribution for the Monte Carlo error. We use this limiting distribution to establish a procedure for finding sets of simultaneous intervals forming confidence regions of approximately correct coverage probabilities. These simultaneous intervals motivate a class of visualizations for Monte Carlo errors for a broad class of estimation procedures for which a multivariate normal limiting distribution holds. Finally, we consider MCMC sample size through sequential stopping rules which terminate simula-

tion once the Monte Carlo errors become suitably small. We develop a general sequential stopping rule for combinations of expectations and quantiles from Markov chain output and provide a simulation study to illustrate the validity.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Markov chain Monte Carlo (MCMC) sampling is a tool to estimate features of a distribution, π, by generating correlated samples from an approximation of π. This scenario frequently arises in a Bayesian setting where models often have complex dependency structures between parameters. A Bayesian practitioner may particularly be interested in credible intervals for each parameter which entails the estimation of multiple quantiles. Often, little attention is provided to the error of the estimation procedure other than to report univariate Monte Carlo standard errors based on marginal distributions, ignoring dependencies among parameters. An improvement is to provide confidence regions for multivariate output. However, currently there does not exist an approach for the multivariate estimation of quantiles for the marginal distributions. Further, visualization of confidence regions is problematic for multi-dimensional quantities as well as practical interpretation. Current visualizations tend to highlight credible intervals of marginal distributions or prediction intervals such as seen in the software "Stan" (Stan Development Team, 2018) or R package (R Core Team, 2013) "tidybayes" (Kay, 2018).

We aim to improve upon this "Stan"dard by providing a visualization which incorporates jointly estimated error in a marginal friendly interpretation. A successful visualization will work for any combination of quantiles and expectations as well as highlight the quality of the estimation. Creating this visualization requires three tasks: (i) develop our joint limiting distribution, a general Markov chain central limit theorem (MC CLT) for the Monte Carlo error of combinations of expectations which is transformed using a Bahadur representation of a quantile, (ii) estimate the covariance of the distribution of our limiting distribution and (iii) generate simultaneous confidence intervals with an overall confidence level.

This visualization will motivate a new sequential stopping rule based on confidence regions formed by our simultaneous intervals. We prove a general sequential stopping rule for MCMC simulations for confidence regions satisfying certain limiting conditions. We then specify conditions on the Markov chain and choose confidence regions which satisfy the more general conditions. An important step in developing this sequential stopping rule is establishing the strong consistency of any estimators present in the confidence region. In particular, we need strongly consistent estimators for the covariance of our limiting distribution which requires estimating two parts: the covariance of a Markov chain CLT for expectations and a transformation matrix containing density values of the marginal distributions of interest.

1.1 Components for Building Visualizations

Markov chain CLTs provide a means to estimate the Monte Carlo standard errors for expectations, based on the variance or covariance matrix of the Normal limiting distribution. These standard errors may then be used to calculate confidence intervals to assess the quality of the estimate. Jones (2004) examines various conditions that establish the existence of a Markov Chain CLT. In particular, he examines the link between mixing sequences and ergodicity leading to some fairly weak conditions. Vats et al. (2019) expand upon the fairly weak polynomial ergodic condition to a multivariate setting. Doss et al. (2014) consider the limiting distribution for the Monte Carlo error of univariate quantiles of functionals. Under the assumption of independent identically distributed data, Ferguson (1998) develops a joint limiting distribution for the mean and a quantile of a random variable, however, this has not been extended for MCMC samples or other cases of dependent sampling.

Quantile estimation is often performed nonparametrically through the use of an order statistic estimator. Several techniques have been used to developed normal limiting distributions for this estimator such as the Bahadur representation of a quantile (Bahadur, 1966) or empirical processes (Doss et al., 2014). While originally developed for independent identically distributed (i.i.d.) random variables, the Bahadur representation has several extensions to cases of dependent sequences such as in Ghosh (1971) and Sen (1968). Extensions to stationary processes are established for uniformly mixing processes in Sen (1972) which are weakened to strongly mixing processes in Wang et al. (2011) and Yoshihara (1995). The conditions in Wang et al. (2011) are generally weaker than in Yoshihara (1995)

except for an assumption that the probability the sequence repeats a value in consecutive times is zero. This extra assumption does not hold for Metropolis-Hastings algorithms and proves to be of limited use to our work.

Confidence regions in multivariate settings often take one of two approaches, a region of minimum volume or simultaneous intervals on the marginal distributions. A minimum volume region represents the smallest set of values for which the given probability level is attained, which takes an elliptical form in the case of a multivariate normal distribution. The location of this ellipsoid is determined by the mean vector while the shape is determined by the covariance matrix. Simultaneous intervals are typically based upon a Bonferroni correction which approximates the overall confidence level of independent marginal intervals, creating a hyper-rectangular confidence region of probability at least the overall confidence level. The location of the hyper-rectangle is determined by the mean vector while the diagonals of the covariance matrix, the marginal variances, determine the length of each side. The ellipsoid confidence region covers less of the support but at the cost of having to consider contour plots to visualize. The hyper-rectangle, however, can be visualized by looking at just the marginal plots but covers a larger portion of the support and has an unknown overall confidence level. For this reason, we propose a method to incorporate the covariance information to determine appropriate marginal confidence levels yielding an overall confidence level for cases of a joint multivariate normal distribution.

1.2 Components for Building a Sequential Stopping Rule

Several estimators have been developed for Markov chain CLT covariance matrices of expectations. Many of these estimators have roots in stationary processes, where dependency structures provide challenges to estimating these variances. The applicability to general dependence structures makes the estimators particularly appealing in the MCMC setting in which they may be applied under slightly different assumptions. The covariance of our joint limiting distribution is composed of two components, a Markov chain CLT covariance matrix for joint expectations and a transformation matrix related to the Bahadur representation of a quantile. We discuss the form of this matrix further in Chapter 4. To estimate the joint expectation covariance matrix, some common families of estimators include regenerative sampling, spectral variance and batch means estimators as discussed by Seila (1982), Vats et al. (2018), and Vats et al. (2019), respectively. Regenerative sampling estimators have fallen out of favor due to modest gains over the comparably cheap batch means estimator as discussed in Jones et al. (2006). Additionally, regenerations occur increasingly infrequently as the dimension of the chain increases requiring substantially more samples. Spectral variance estimators tend to outperform batch means estimators in asymptotic efficiency for certain lag windows and truncation points. Furthermore, work has gone into developing optimal truncation points for window functions in Liu and Flegal (2018a). The increased quality of estimation comes with a substantial increase in computational cost, however batch means estimators have similar performance when samples sizes become large. A new family of batch means estimators, the weighted batch means estimator introduced by Liu and Flegal (2018b), incorporates window functions into the batch

means procedure trading a modest increase in computational cost for significant gains in estimation quality. Furthermore, Vats and Flegal (2018) introduce a new window function, the lugsail window, to address concerns of downward bias known to plague batch means estimators. A thourough study of the performance of weighted batch means and spectral variance estimators in Liu and Flegal (2018b) informs our decision to limit this work to the family of weighted batch means estimators.

Assessing the standard error of quantile estimates requires estimating the density of the corresponding quantile at said point. Several classes of density estimators have been proposed of which kernel density estimators (Rosenblatt, 1956) are one of the most popular and well studied. Silverman (1986) provides a review of several, but a non-exhaustive list, of these estimators. We restrict our conversation to the class of kernel density estimators. Particularly of interest to us is the property of strong consistency and identifying sufficient conditions on a Markov chain. Early work by Ryzin (1969) only establishes the strong consistency of kernel density estimators for independent data and is not sufficient for our work. We turn to stationary process literature and the result of Meyn and Tweedie (1993) which establishes conditions for which Markov chains may behave like a stationary process. Takahata (1980) establishes strong consistency for univariate weakly dependent sequences via a law of the iterated logarithm. Masry and Györfi (1987) provides multivariate density estimation for assumption of asymptotically uncorrelated processes using mixingale theory. Roussas (1988) establishes strong consistency under different conditions for strong mixing processes.

The classic question of sample size determination, or stopping time, has a different context in the case of simulated procedures. The collection of samples becomes a matter

of computational resources and time instead of a physical collection of data. This makes collecting additional samples to improve the certainty of estimation comparatively cheap. Some early attempts at answering the question of whether enough samples have been collected took the form of checking various diagnostics for non-convergence of the Markov chain. Once the diagnostics fail to detect non-convergence estimation may then take place. One of these diagnostics, the Heidelberger-Welch diagnostic Heidelberger and Welch (1983), applies to general stationary processes of one-dimension and considers the process satisfactory once the confidence interval length for a mean is below some threshold based weighted by the magnitude of the estimate, or relative magnitude. Glynn and Whitt (1992) formalize this process for fixed-width and relative magnitude procedures to parameters of general stochastic sequences satisfying a functional central limit theorem (FCLT). This provided the first sequential stopping rule which is applicable to MCMC as formalized for univariate sequences in Jones et al. (2006). Flegal and Gong (2015) weight the fixed-width using the posterior standard deviation as an improvement for Bayesian problems. Doss et al. (2014) consider the estimation of quantiles. Early multivariate stopping rules consider termination once each dimension of the chain has reached its stopping time and taking the largest as provided by Gong and Flegal (2016). Vats et al. (2019) expand the univariate results to a multivariate stopping rule for expectations based on the volume of the confidence region. We extend the results of Vats et al. (2019) to jointly estimated expectations and quantiles of functionals of the chain in Chapter 4.

Chapter 2 provides some background theory on Markov chains and MCMC simulations needed to develop our class of visualizations and sequential stopping rule. We next present the current state of MCMC output analysis in Chapter 3. We provide an example

to illustrate certain properties of MCMC samplers and a second example to serve as a guide for what should be expected in an analysis. Chapter 4 contains the submitted paper "New Visualizations for Monte Carlo Simulations" (Robertson et al., 2019) which develops a joint limiting distribution for combinations of expectations and quantiles from Monte Carlo simulation and provide a class of visualizations to analyze the resulting Monte Carlo error. In Chapter 5 we develop a sequential stopping rule based on our joint limiting distribution and provide a simulation study as evidence for the validity of our methods. We conclude with a discussion of future work in Chapter 6.

Chapter 2

Markov Chains and MCMC

To facilitate our development of MCMC output analysis tools, we present some Markov chain theory and set notation that will be used throughout this work. We review concepts of Markov chain converge, CLTs, and current output analysis tools that will be built upon in subsequent chapters. We conclude with a review of some common MCMC sampling algorithms.

2.1 Markov Chain Theory

Let $\{X_i\}_{i=1}^{\infty} \in \mathsf{X}^d$ be a Harris recurrent Markov chain with invariant distribution π. Essentially, Harris recurrence provides that any set of positive measure may be reached with positive probability must occur infinitely often. A formal definition for Harris recurrent is beyond the scope of this work and may be found in Meyn and Tweedie (1993). The Markov transition kernel defines the probability of transitioning to state A given that the chain had value x, i.e. $P(x, A) = Pr\{X_{j+1} \in A | X_j = x\}$. The n-step transition kernel defines the

probability of this transition in n steps, i.e.

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

This transition probability only depends on the past through the state at time j. We need a more general definition for cases where X_j is from some probability distribution instead of a fixed value. When this initial probability measure is the invariant distribution the process is stationary, but this is generally uncommon in MCMC applications. For an initial probability measure $\lambda(\cdot)$ on $\mathcal{B}(X)$, we define

$$P^n(\lambda, A) = \int_{\mathsf{X}} P^n(x, A)\lambda(dx).$$

The dependence upon the previous value in the chain creates a correlation structure providing challenges when making inference about π.

2.1.1 Ergodicity

Many properties about the Markov chain are characterized by the behavior of $P^n(x, A)$ as $n \to \infty$. The total variation norm,

$$TV(P^n) = \lim_{n\to\infty} ||P^n(\lambda, \cdot) - \pi(\cdot)|| = 2 \lim_{n\to\infty} \sup_A |P^n(x, A) - \pi(A)|,$$

measures how quickly the n-step transition kernel converges to the invariant distribution. Bounds placed on $TV(P^n)$ act as conditions from which much of the theory is derived.

Consider the bound

$$TV(P^n) \leq M(x)t^n. \tag{2.1}$$

When (2.1) holds for some $t < 1$ and a nonnegative bounded function $M(x)$ the chain is said to be uniformly ergodic. This is the strongest condition we consider but is perhaps unreasonable to assume for general state space Markov chains. A weaker form of ergodicity, geometrically ergodic, relaxes the condition on $M(x)$ to merely be a non-negative function. The last form of ergodicity we consider is the weaker polynomial ergodicity. A chain is considered polynomial ergodic of order m when there exists a non-negative function $M(x)$ and $m \geq 0$ such that

$$TV(P^n) \leq M(x)n^{-m}. \tag{2.2}$$

Uniform ergodicity implies geometric ergodicity which in turn implies polynomial ergodicity. Hence, any results requiring polynomial ergodicity will hold for other forms. Establishing geometric or polynomial ergodicity is often quite challenging and problem specific. Ergodicity is established on a theoretical level often using drift and minorization conditions (Meyn and Tweedie, 1993) and not based on the output of a chain. Geometric ergodicity has been shown to be satisfied for many Gibbs samplers and Metropolis-Hasting algorithms (see e.g. Acosta et al., 2014; Doss and Hobert, 2010; Hobert and Geyer, 1998; Jarner and Hansen, 2000; Jarner and Roberts, 2002; Johnson et al., 2013; Jones and Hobert, 2004; Jones et al., 2012; Marchev and Hobert, 2004; Roberts and Polson, 1994; Tan and Hobert, 2009), indicating that geometric ergodicity is often a reasonable assumption.

2.1.2 Mixing Sequences

A second way to characterize the asymptotic behavior of a sequence of random variables is through mixing conditions. These measure how quickly the dependence of between X_k and X_{k+n} decays as $n \to \infty$ and provide a notion of how much stickiness exists in a sequence. A slowly decaying dependence causes large amounts of correlation in the sequence hindering the ability to explore the state space. By placing a notion on the rate of decay of dependence we may develop a notion of how well the state space is explored.

Let $\mathcal{F}_j^{j+k}$ be the σ-algebra generated by $X_j, \ldots, X_{j+k}$, $\sigma(X_j, \ldots, X_{j+k})$, and $\mathcal{F}_n^{\infty}$ be the tail σ-algebra, $\sigma(X_n, X_{n+1}, \ldots)$. A sequence is considered strongly mixing, or $\alpha-$mixing, if

$$\alpha(n) = \sup_{k \geq 1} \sup_{A \in \mathcal{F}_1^k, B \in \mathcal{F}_{k+n}^{\infty}} |P(A \cap B) - P(A)P(B)| \overset{n \to \infty}{\longrightarrow} 0. \tag{2.3}$$

The next mixing condition we condiser requires defining $L_2(\mathcal{F})$, the set of square integrable functions over $\mathcal{F}$. A sequence is considered asymptotically uncorrelated, or $\rho-$mixing, if

$$\rho(n) = \sup\{\mathrm{Cor}(U, V) : U \in L_2(\mathcal{F}_1^k), V \in L_2(\mathcal{F}_{k+n}^{\infty}), k \geq 1\} \overset{n \to \infty}{\longrightarrow} 0. \tag{2.4}$$

A sequence is considered uniformly mixing, or $\phi-$mixing, if

$$\phi(n) = \sup_{k \geq 1} \sup_{A \in \mathcal{F}_1^k, B \in \mathcal{F}_{k+n}^{\infty}} |P(B|A) - P(B)| \overset{n \to \infty}{\longrightarrow} 0. \tag{2.5}$$

Uniformly mixing implies asymptotically uncorrelated and asymptotically uncorrelated implies strongly mixing. A summary of the relationship between various mixing conditions

may be found in Bradley (1986, 2005). The various links between mixing conditions and various forms of ergodicity are discussed by Jones (2004) and used to establish several conditions for the existence of a Markov chain CLT. The most relevant of these results to this work provides that a geometrically ergodic Markov chain that satisfies detailed balance is an asymptotically uncorrelated sequence. A kernel P is said to satisfy detailed balance with respect to π if

$$\pi(dx)P(x, dy) = \pi(dy)P(y, dx).$$

A Markov chain that satisfies detailed balance is reversible. This guaranties any transition which occurs has a nonzero probability of occurring in reverse during the next transition.

2.2 Estimation

Frequently features of π that are of interest may be expressed as expectations or quantiles of a function of the Markov chain $\{X_j\}$. Posterior means and variances fit into the former while posterior credible intervals are an example of the latter.

2.2.1 Expectations

Let $g : \mathsf{X} \to \mathbb{R}^p$ and θ_g be a feature of π such that

$$\theta_g = E[g(X)] = \int_{\mathsf{X}} g(x)\pi(dx). \tag{2.6}$$

We may then estimate $\theta_g$with

$$\bar{g}_n = \sum_{i=1}^{n} g(X_i). \tag{2.7}$$

When $E[|g|] < \infty$, the Birkhoff ergodic theorem (Birkhoff, 1931) provides that $\bar{g}_n$ is a strongly consistent estimator for $E_\pi[\bar{g}_n] = \theta_g$, i.e. $\bar{g}_n \to \theta_g$ with probability one. Under certain conditions a Markov chain central limit theorem (MC CLT) holds, i.e.

$$\sqrt{n}(\bar{g}_n - \theta_g) \xrightarrow{d} N(0, \Sigma) \tag{2.8}$$

as $n \to \infty$ (Jones, 2004). Typically the assumptions for a Markov chain CLT consist of a moment condition on the function g and either a mixing condition or a form of ergodicity. For example, Vats et al. (2019) establish a CLT under the fairly week conditions there exists a $\delta > 0$ such that $E|g|^{2+\delta} < \infty$ and $\{X_j\}$ is polynomial ergodic of order $m > (2+\delta)/\delta$. These results provide limiting distributions for moments but not quantiles of interest.

The structure of Σ is particularly noteworthy in that it accounts for the dependence of the sampling procedure in addition to an estimate of the covariance of $g(X_j)$. That is

$$\Sigma = \text{Cov}(g(X_j), g(X_j)) + \sum_{i=1}^{\infty} \left[\text{Cov}(g(X_j), g(X_{j+i})) + \text{Cov}(g(X_j), g(X_{j+i}))'\right]. \tag{2.9}$$

It will be convenient to denote the first term

$$\Sigma_g = \text{Cov}(g(X_j), g(X_j)). \tag{2.10}$$

This is the covariance that would result if the sequence $g(X)$ were independent identically distributed.

Several estimators has been proposed to estimate Σ (see Section 1.2). We limit our discussion to the batch means estimator which has been shown to be a strongly consistent

estimator with increasing batch sizes (Vats et al., 2019). Let $b = \lfloor \sqrt{n} \rfloor$ be the batch size equal a be the number of batches such that $n = ab$. Taking $Y = g(X)$ for simplicity of notation, the mean for the k^{th} batch is

$$\bar{Y}_k(b) = b^{-1} \sum_{t=1}^{b} Y_{kb+t},$$

while the overall mean is

$$\bar{Y} = a^{-1} \sum_{k=1}^{a} \bar{Y}_k(b).$$

The batch means approach treats the mean of each batch as if it were an observation and calculates the sample covariance matrix. The batch means estimator with batch size b is

$$\hat{\Sigma} = \frac{b}{a-1} \sum_{k=0}^{a-1} (\bar{Y}_k(b) - \bar{Y})(\bar{Y}_k(b) - \bar{Y})'. \tag{2.11}$$

2.2.2 Quantiles

Let $W \sim \pi$, $h : \mathsf{X} \to \mathbb{R}$, and $V = h(W)$. We consider estimation of quantiles of V, the distribution of a functional on π. Letting F_V and f_V be the distribution and density functions of V, the quantile is defined as

$$\xi_q = \inf\{v : F_V(v) \geq q\}. \tag{2.12}$$

This may be estimated with

$$\hat{\xi}_q = h(X)_{\lfloor nq \rfloor : n}, \tag{2.13}$$

the $\lfloor nq \rfloor^{th}$ order statistic of $h(X)_i$. Doss et al. (2014) establish the strong consistency of $\hat{\xi}_q$. When $\{X_i\}$ is polynomial ergodic of order $m > 1$ a MC CLT holds,

$$\sqrt{n}(\hat{\xi}_q - \xi_q) \xrightarrow{d} N\left(0, \frac{\sigma^2}{f_V(\xi_q)^2}\right), \tag{2.14}$$

where σ^2 is the variance of the MC CLT for the Monte Carlo error

$$\frac{1}{n}\sum_{i=1}^{n} I(h(X_i) > \hat{\xi}_q) - (1-q).$$

This Monte Carlo error is a special case of (2.8) using an indicator function as the choice of g. The result in (2.14) has thus far only been established for a single quantile of a univariate distribution V. When multiple quantiles from V are of interest, such as when estimating the endpoints of a credible interval, or when quantiles from different marginal distributions, each quantity must either be considered separately or with a multiple correction such as Bonferroni. In Chapter 4 we develop a general joint limiting distribution for combinations of expectations and quantiles.

2.3 Effective Sample Size

Due to the dependence of a Markov chain, the variability of the Monte Carlo error is higher than if independent identically distributed samples were collected from π. Recall $\Sigma_g = \text{Cov}(g(X), g(X)^T)$. One might wish to know how many i.i.d. samples our dependent Monte Carlo sample is equivalent to. To measure this, we consider the effective sample size

(ESS) (Vats et al., 2019) defined as

$$\text{ESS} = n \left(\frac{|\Sigma_g|}{|\Sigma|} \right)^{1/p}. \tag{2.15}$$

To estimate (2.15) we must estimate Σ using an estimator such as batch means (2.11) and Σ_g. The sample covariance estimator is appropriate for estimating Σ_g, i.e.

$$\hat{\Sigma}_g = \frac{1}{n} \sum_{j=1}^{n} (Y_j - \bar{Y})(Y_j - \bar{Y})'.$$

The choice of n versus $n-p$ in the denominator is largely irrelevant when $n >> p$ as should be the case for a reasonable number of MCMC samples.

ESS for estimating quantiles is similarly defined. Let $\sigma^2_{Ig} = \text{Var}(I(g(X) \leq \xi_q))$. Then effective sample size for quantiles reduces to the ratio of variances of the limiting distributions for estimating the indicator functions,

$$\text{ess} = n \frac{\sigma^2_{Ig}}{\sigma^2}. \tag{2.16}$$

2.4 Sequential Stopping Rules

Consider estimating θ_g using simulations from the Markov chain $\{X_j\}_{j=1}^n$. Under a Markov chain CLT the Monte Carlo error becomes arbitrarily small as $n \to \infty$ and simulation may stop once this error becomes suitably small. Consider an error tolerance ϵ and a confidence level $1-\alpha$. Assume there exists a CLT for the Monte Carlo error $\bar{g}_n - \theta_g$ and let $\chi^2_{1-\alpha,p}$ be the $1-\alpha$ quantile of a χ^2_p distribution. Then there exists a $1-\alpha$ confidence

region

$$C_\alpha^E(n) = \{\bar{g}_n : n(\bar{g}_n - \theta_g)^T \Sigma^{-1} (\bar{g}_n - \theta_g) \leq \chi^2_{1-\alpha,p}\}. \tag{2.17}$$

In practice Σ is unknown and must be estimated. Accounting for the uncertainty in estimating Σ leads to an improved confidence region

$$C_\alpha^T(n) = \{\bar{g}_n : n(\bar{g}_n - \theta_g)^T \hat{\Sigma}^{-1} (\bar{g}_n - \theta_g) \leq T^2_{1-\alpha,p,\nu(n)}\}, \tag{2.18}$$

where the degrees of freedom $\nu(n)$ of the Hotelling's T^2 distribution depends upon the estimator of Σ and is an increasing function of n. Since $T^2_{p,\nu(n)} \overset{n\to\infty}{\longrightarrow} \chi^2_p$, $C_\alpha^T(n) \to C_\alpha^E(n)$ as $n \to \infty$. The corresponding volume of the confidence region is

$$\text{Vol}(C_\alpha^T(n)) = \frac{2\pi^{p/2}}{p\Gamma(p/2)} \left(\frac{T^2_{1-\alpha,p,\nu(n)}}{n} \right)^{p/2} |\hat{\Sigma}|^{1/2}, \tag{2.19}$$

with $\text{Vol}(C_\alpha^E(n))$ defined similarly.

Let K be a metric of the estimation process. Define a stopping time

$$t^*(\epsilon) = \inf\{n : (\text{Vol}(C_\alpha^T(n))^{1/p} + a(n) \leq \epsilon \hat{K}\}, \tag{2.20}$$

where $a(n)$ is a function $a(n) = o(n^{-1})$. This stopping time, called a fixed-width relative metric sequential stopping time, represents the smallest n such that the p^{th} volume of the confidence region becomes smaller than the prescribed error tolerance weighted by the metric K. Theorem 1 from Vats et al. (2019) formalizes this under fairly weak assumptions.

Theorem 1 *Let $g : \mathsf{X} \to \mathbb{R}^p$ be such that $E_F||g||^{2+\delta} < \infty$ for some $\delta > 0$ and let X be an F-invariant polynomial ergodic Markov chain of order $k > (1+\epsilon_1)(1+2/\delta)$ for some $\epsilon_1 > 0$. If $\hat{K} \to K$ with probability 1 and $\hat{\Sigma} \to \Sigma$ with probability 1, as $n \to \infty$, then, as $\epsilon \to 0, t^*(\epsilon) \to \infty$ and $Pr\{\theta_g \in C_\alpha^T(t^*(\epsilon))\} \to 1-\alpha$.*

The univariate case reduces to stopping once the length of the confidence interval becomes less than $\epsilon\hat{K}$,

$$t^*(\epsilon) = \inf\{n : 2z_{1-\gamma/2}\frac{\hat{\sigma}}{\sqrt{n}} + a(n) \leq \epsilon\hat{K}\},$$

and provides some motivation for the choice of relative metric K. The relative metric provides a weighting of ϵ to determine what is considered reasonably small. When the notion of small is known, that is for cases when ϵ does not need to be weighted, it may be reasonable to consider the identity metric $K = 1$. Another choice of metric from operations research literature for general stochastic simulations is the relative magnitude metric (Glynn and Whitt, 1992; Heidelberger and Welch, 1983) where $K = |\theta_g|$ and $|\cdot|$ is the absolute value or the Euclidean norm in higher dimensions. This choice of metric leans on the idea that the amount of acceptable error should be related to the magnitude of the quantity of interest. This also helps to solve the issue of different stopping times for chains sampling distributions with different units such as millimeters versus meters. A third metrics, the relative standard deviation metric, proposed by Flegal and Gong (2015) is motivated by the Bayesian setting where K is chosen to be the posterior standard deviation. Vats et al. (2019) consider a generalization setting $K = \sigma_g$, or $K = |\Sigma_g|^{1/2p}$ in the multivariate case, to extend the metric to cases where a functional of the invariant distribution is of interest.

This choice of metric appeals to the sense of choosing reasonably small based on how much variance there is in the quantity of interest. Choosing $K = |\Sigma_g|^{1/2p}$ in (2.20) is equivalent to terminating simulation once the ESS has reached some threshold related to ϵ and p (Vats et al., 2019), providing another intuition for the reasonability of this metric.

2.5 Sampling a Markov Chain

Various algorithms exist for sampling a Markov chains with two of the most popular being the Metropolis-Hastings algorithm and the Gibbs sampler.

2.5.1 Metropolis-Hastings

The Metropolis-Hastings (MH) algorithm generates samples from a probability distribution by simulating a draw from a proposal distribution and accepting this draw into the sample with an acceptance probability based on the proposed value and the previous value of the Markov chain. If the proposed value is rejected, then the previous value is taken as the new value. Formally, let $F(\cdot|x)$ be the proposal distribution with density $f(\cdot|x)$. The Metropolis-Hasting algorithm generates x_{i+1} given current state x_i as follows.

1. Sample x^* from $f(\cdot|x_i)$
2. Let $r(x_i, x^*) = \min\left\{\frac{\pi(x^*)f(x_i|x^*)}{\pi(x_i)f(x^*|x_i|)}, 1\right\}$
3. Set $x_{i+1} = \begin{cases} x^* & \text{w.p. } r(x_i, x^*) \\ x_i & \text{w.p. } 1 - r(x_i, x^*) \end{cases}$

Since the acceptance probability includes $\pi(\cdot)$ in both the numerator and denominator the normalizing constant cancels causing π to only need to be known up to that constant. The

choice of proposal distribution greatly affects the quality of the sampler. More specifically, the choice of proposal determines the mixing properties as discussed above. We illustrate these properties in Chapter 3.

A special MH algorithm that is particularly popular is the Metropolis-Hasting Random Walk which selects $f(\cdot|x)$ to be symmetric about x. Often $f(\cdot|x)$ is chosen to be normal. The sampling algorithm may be rewritten in the following form.

1. Sample ϵ from $f(\cdot|0)$
2. Set $x^* = x_i + \epsilon$
3. Let $r(x_i, x^*) = \min\left\{\frac{\pi(x^*)}{\pi(x_i)}, 1\right\}$
4. Set $x_{i+1} = \begin{cases} x^* & \text{w.p. } r(x_i, x^*) \\ x_i & \text{w.p. } 1 - r(x_i, x^*) \end{cases}$

These types of samplers are particularly useful in that the variance of the proposal distribution may be tuned to find an adequate acceptance probability. This may often be easier than specifying new shapes as in the standard MH algorithm.

2.5.2 Gibbs Sampler

Another commonly used algorithm is the Gibbs sampler. In elaborate models of high dimension finding an appropriate proposal distribution for a MH sampler can be particularly challenging. The Gibbs sampler is a common choice to handle these types of problems as it generates samples from a joint distribution by sampling from a series of conditional distributions.

Consider sampling from a distribution $f(y), y \in \mathcal{Y}_1 \times \ldots \times \mathcal{Y}_m$ for which each conditional distribution $f_j(y_j|y_1, ..., y_{j-1}, y_{j+1}, ..., y_m)$ may be simulated from. A sequential scan Gibbs sampler generates sample $y^{(i+1)}$ given current state $y^{(i)}$ as follows.

1. Sample y_1^* from $f_1(\cdot|y_2, ..., y_m)$

2. Sample y_2^* from $f_2(\cdot|y_1^*, y_3, ..., y_m)$

...

j. Sample y_j^* from $f_j(\cdot|y_1^*, ..., y_{j-1}^*, y_{j+1}, ...y_m)$

...

m. Sample y_m^* from $f_m(\cdot|y_1^*, ..., y_{m-1}^*)$

m+1. Set $y^{(i+1)} = (y_1^*, ..., y_m^*)$

The order of updates for the various y_j is up to the practitioner with the order often chosen according to the influence of the starting values. The order of updates need not even be fixed as in the case of the random scan Gibbs sampler in which $y_1, ..., y_m$ are sampled in a random order each update. The random scan is not commonly used in practice but has a place in the theory as it satisfies detailed balance (Jones, 2004). Other variants of the Gibbs sampler may update y_j as a block of random variables when f_j is multivariate. Additionally, various sampling procedures may be used to update an individual y_j such as a MH update when $f(y_j|\ldots)$ is challenging to sample from. This particular update is referred to as a Metropolis within Gibbs update. These component wise samplers may offer superior convergence rates over full dimensional updates (Johnson et al., 2013). Particularly, component wise updates provide larger improvements when little correlation exists between components.

Chapter 3

MCMC Output Analysis

The conclusions drawn from an analysis are only be as good as the MCMC sampler the analysis is built on and the quality of estimation. To assess the quality of the sampler and estimation we may look at the output of the sample and estimates. More specifically, the quality of the sampler may be assessed by examining trace plots and acceptance rates while the quality of the estimates may be assessed by examining the standard errors or measures related to the standard errors.

We illustrate the concerns of output analysis through the following example. Let X be a random variable distributed according to a mixture of 3 normal distributions, with density

$$f_X(x) = .3f_1(x; 1, 2.5) + .5f_2(x; 5, 4) + .2f_3(x; 11, 3). \tag{3.1}$$

where $f_j(x; \mu_j, \sigma_j^2)$ is the density of the j^{th} mixture component with mean μ_j and variance σ_j^2. We consider estimating the mean and variance of this distribution using a MH random walk with proposal distribution $N(0, \gamma^2)$. A more thorough explanation of a MH sampler

may be found in Section 2.5.1. In this case we have an analytic solution available

$$E[X] = \sum_{i=1}^{3} a_i \mu_i = 5,$$

and

$$\text{Var}(X) = \sum_{i=1}^{3} a_i(\sigma_i^2 + \mu_i^2) - \left(\sum_{i=1}^{3} a_i \mu_i \right)^2 = 15.35.$$

3.1 Sampler Quality

In the following we consider various settings of the MH random walk to illustrate the usefulness of output analysis. Let γ^2 be the variance of the proposal distribution and X_1 be the starting value of the Markov chain. We consider the following cases.

- Moderate Mixing/Poor Start: $\gamma^2 = 9$ and $X_1 = 50$
- Moderate Mixing/Good Start: $\gamma^2 = 9$ and $X_1 = 0$
- Poor Mixing: $\gamma^2 = 10,000$ and $X_1 = 0$
- Good Mixing: $\gamma^2 = 100$ and $X_1 = 0$

These cases will be sufficient to illustrate what we look for in assessing the quality of output from an MCMC sample and will be referenced in the discussion below.

3.1.1 Starting Values

In some special cases a Markov chain may be able to start with an initial value from the invariant distribution, π, of interest. This may occur in cases for which a linchpin sampler (Acosta et al., 2014) is available, but is generally unavailable for situations in which

MCMC is the appropriate tool. In practice starting values must be chosen which affects the estimates. However, these estimates will be asymptotically unbiased due to the ergodic theorem. In this sense more samples taken may make up for a poor starting value with the quantity of additional samples related to how extreme the poor value. A poor starting value is a starting value in the extreme tails of the distribution away from the bulk of the density. Figure 3.1 pictures an example of a poor and reasonable starting value. The resulting estimates of the mean after 1,000 samples are 4.54 and 6.26 for a reasonable and bad starting value respectively. Both estimates are reasonably far from the true value due to the small sample size, however, the poor starting value has a noticeable effect on the estimation as it introduces substantial finite sample bias.

"Any point you don't mind having in a sample is a good starting point" (Charles Geyer). While simple, these words provide a useful starting place for a discussion on choosing a starting value. As long as a value is in the state space it should eventually be visited given infinite samples and may be considered a reasonable starting value given the result of Meyn and Tweedie (1993) that convergence of the Markov kernel for an initial distribution implies convergence for every initial distribution. However, as demonstrated in Figure 3.1 some starting values were superior to others. This occurs due to how likely the vales are to occur compared to how many samples have been taken. Our poor starting value causes a disproportionate amount of the sample to be in an extreme tail of the distribution. As more samples are collected this proportion will even out over time. This motivates choosing starting values within the bulk of the density. A posterior mode estimate fulfills this conditions, and usually a rough maximum likelihood estimate will as well. A

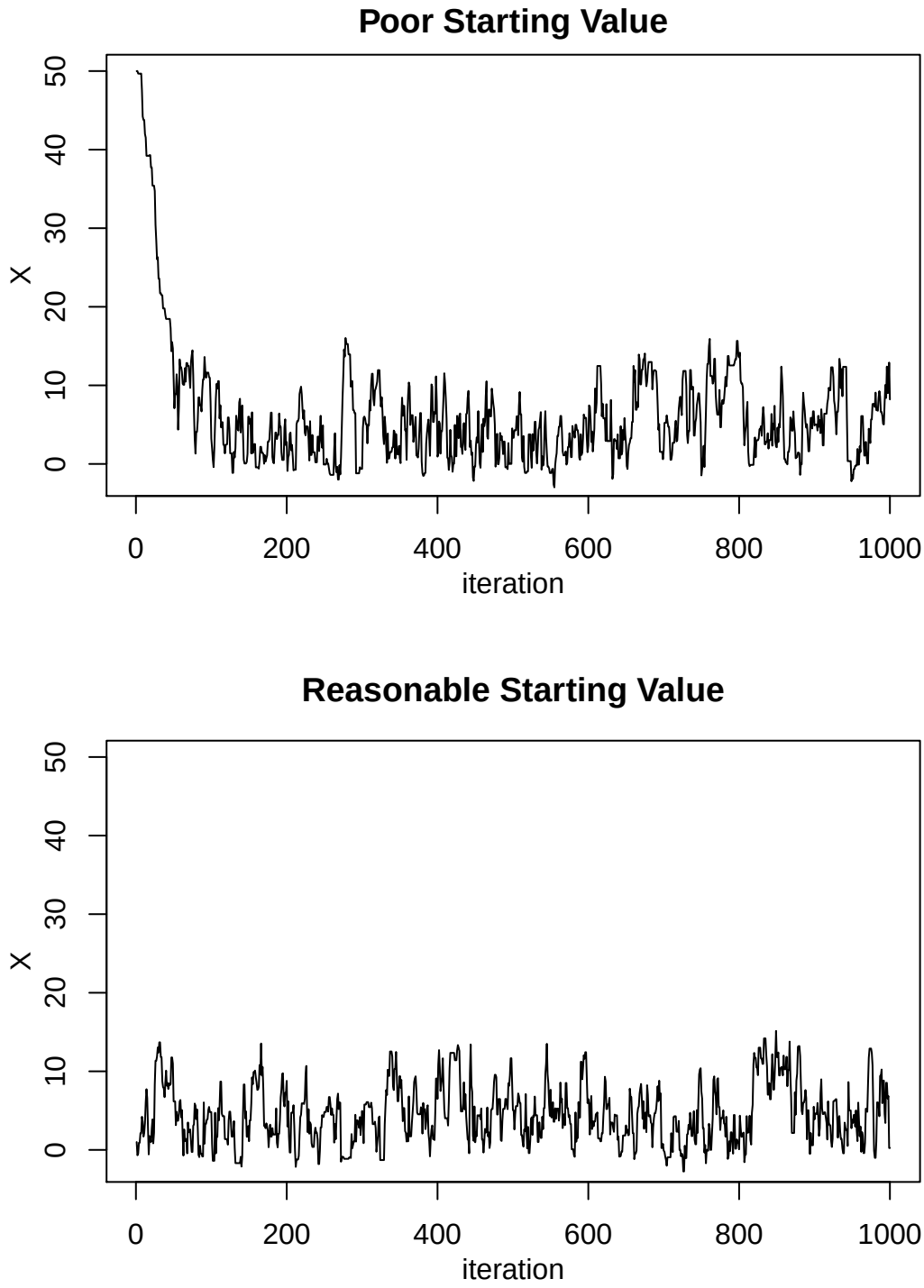

Figure 3.1: Trace plots for the first 1,000 iterations of a MH random walk sampler with reasonable and poor starting values

linchpin sampler (Acosta et al., 2014) may be used to generate an initial sample achieving a stationary process, however this is not required for the existence of a CLT and SLLN we require. In some cases it may be more effective to run a preliminary chain to identify areas of higher density.

3.1.2 Mixing

Another concern with any sampler are the mixing properties of the Markov chain. These indicate how quickly and well the Markov chain explores the target distribution. The theoretical mixing properties are further discussed in Section 2.1. Mixing properties may also be examined visually using trace plots as in the top row of Figure 3.2. The nature of the random walk provides two conflicting goals in wanting to accept a reasonable number of samples while also wanting the samples to explore the state space. Large jumps in the chain are good for exploring the state space but often have low acceptance probabilities leading to the chain getting stuck for long periods of time, as is the case in the poor mixing in Figure 3.2. To get a large acceptance probability the steps become very small leading to heavy correlation in the chain. The moderate mixing example provides a case where the steps are fairly small but still provide a reasonable exploration of the state space. The good mixing example balances jump size with acceptance probability to provide quicker exploration of the state space. In our good mixing example the acceptance probability is very close to the theoretical optimal value for a one dimensional chain of .44 (Roberts and Rosenthal, 2001). A discussion on optimal acceptance probabilities for MH random walk samplers may be found in Rosenthal's chapter of Brooks et al. (2010). The bottom row of Figure 3.2 provides the density estimates based on each chain with the actual density

superimposed in red. The results provide another visual of the quality of mixing with the poor mixing not attaining an adequate view of the target distribution while the moderate mixing provides a reasonable fit.

Another tool to visualize mixing is through an auto-correlation plot as seen in the middle row of Figure 3.2. Auto-correlation for lag k is the correlation between X_j and X_{j+k}, and the auto-correlation plot plots the estimated auto-correlation for increasing lag values. Since mixing is measure of how quickly the dependence fades between values early and late in the chain, a chain with good mixing should exhibit rapidly decaying auto-correlation and approach 0. Here we see the auto-correlation decreasing most rapidly for our good mixing example while our poor mixing example has the slowest decay.

3.1.3 Burn-In

The practice of discarding early iterations of the Markov chain is known as burn-in. This practice lacks a theoretical justification as it is just a method of specifying an initial distribution. Burn-in does not provide stationarity of the Markov chain and should not be viewed as such. While the theory says it is unnecessary, Figure 3.1 provides an example of when burn-in may be helpful. The estimate based on the poor starting value is closer to truth when omitting the first 50 samples. However, discarding the initial 50 samples may have been avoided by picking a better starting value such as in the reasonable starting value plot. In practice, burn-in may be reasonable to consider when it is more efficient to run the chain for extra iterations than to perform a quick starting value search.

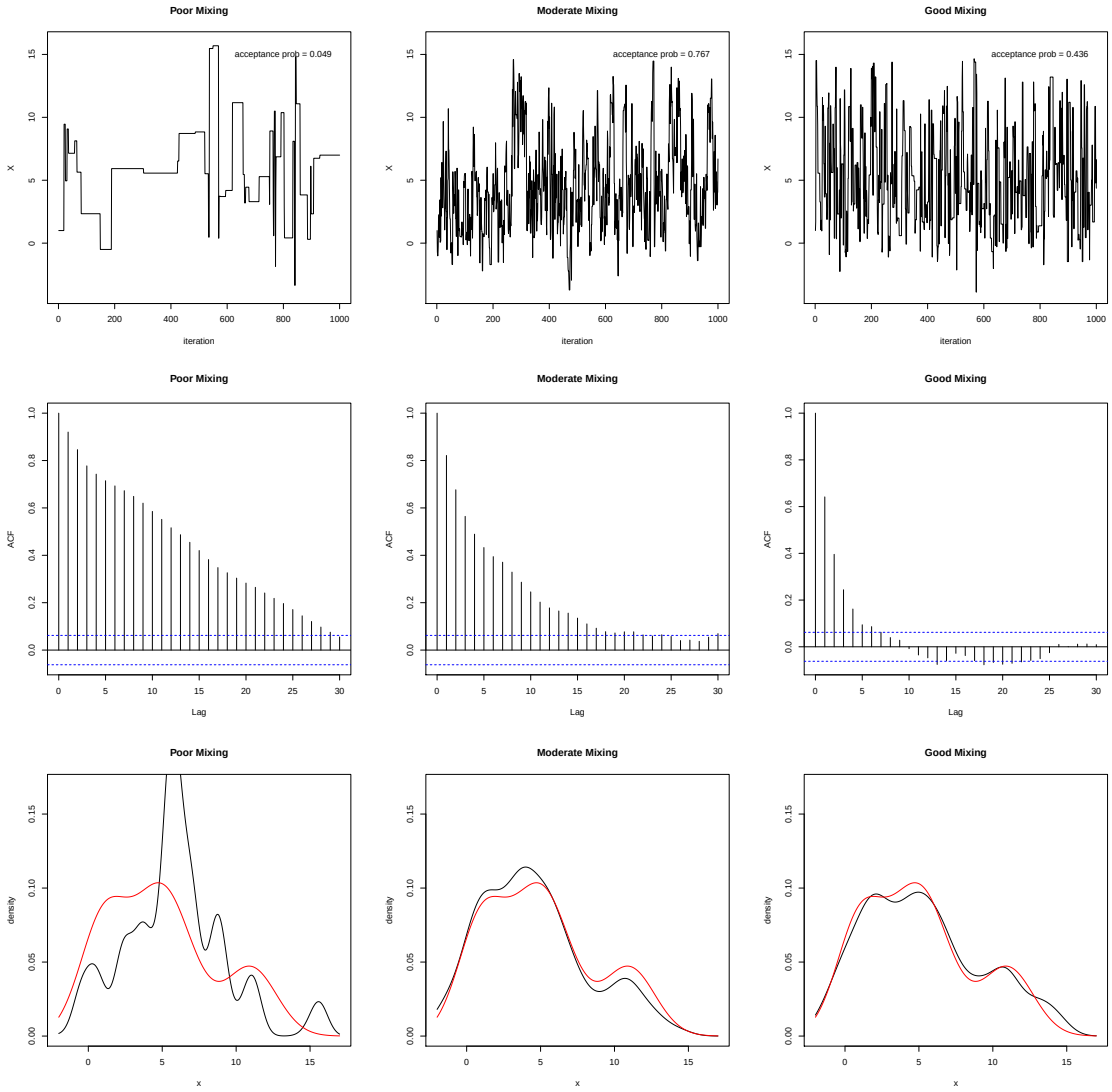

Figure 3.2: Top: trace plots, Middle: auto-correlation plots, Bottom: density estimates for the first 1,000 iterations of a MH random walk for various proposal distributions

3.2 Sampling Termination

To address the question of when have enough samples been collected, we turn to sequential stopping rules and effective sample sizes. Vats et al. (2019) established the connection between termination based on a fixed-volume relative standard deviation sequential stopping rule and termination once the effective sample size has reached some prespecified threshold. Each of these quantities is dependent upon the standard errors of the estimate.

We set $\epsilon = .05$, see (2.20), and a 95% confidence level, which for a two parameter estimation and a relative standard deviation corresponds to an effective sample size (2.15) of 7,529. Our procedure is as follows. Take 10,000 initial samples and calculate the stopping criteria. If the stopping criteria is not met, collect 10,000 additional samples and recalculate the stopping criteria. This is repeated until the stopping criteria is met at which time the total number of samples taken is considered the stopping time. Figure 3.3 depicts the confidence region calculated after each 10,000 samples for various samplers. Each ellipse is colored corresponding to how close the number of samples is to the stopping time with green representing the first check. The colors gradual shift to blue the closer to the stopping time. The ellipse corresponding to the stopping time is drawn in red. The lone black point marks the true parameter values being estimated. The chains with good and moderate mixing stop relatively quickly while the poor mixing chain requires substantially more simulation effort. The starting values also demonstrate an effect on the confidence region progression with very large confidence regions initially. Figure 3.4 places all the confidence regions at stopping time in one frame. We see that the confidence regions are all of a similar size with

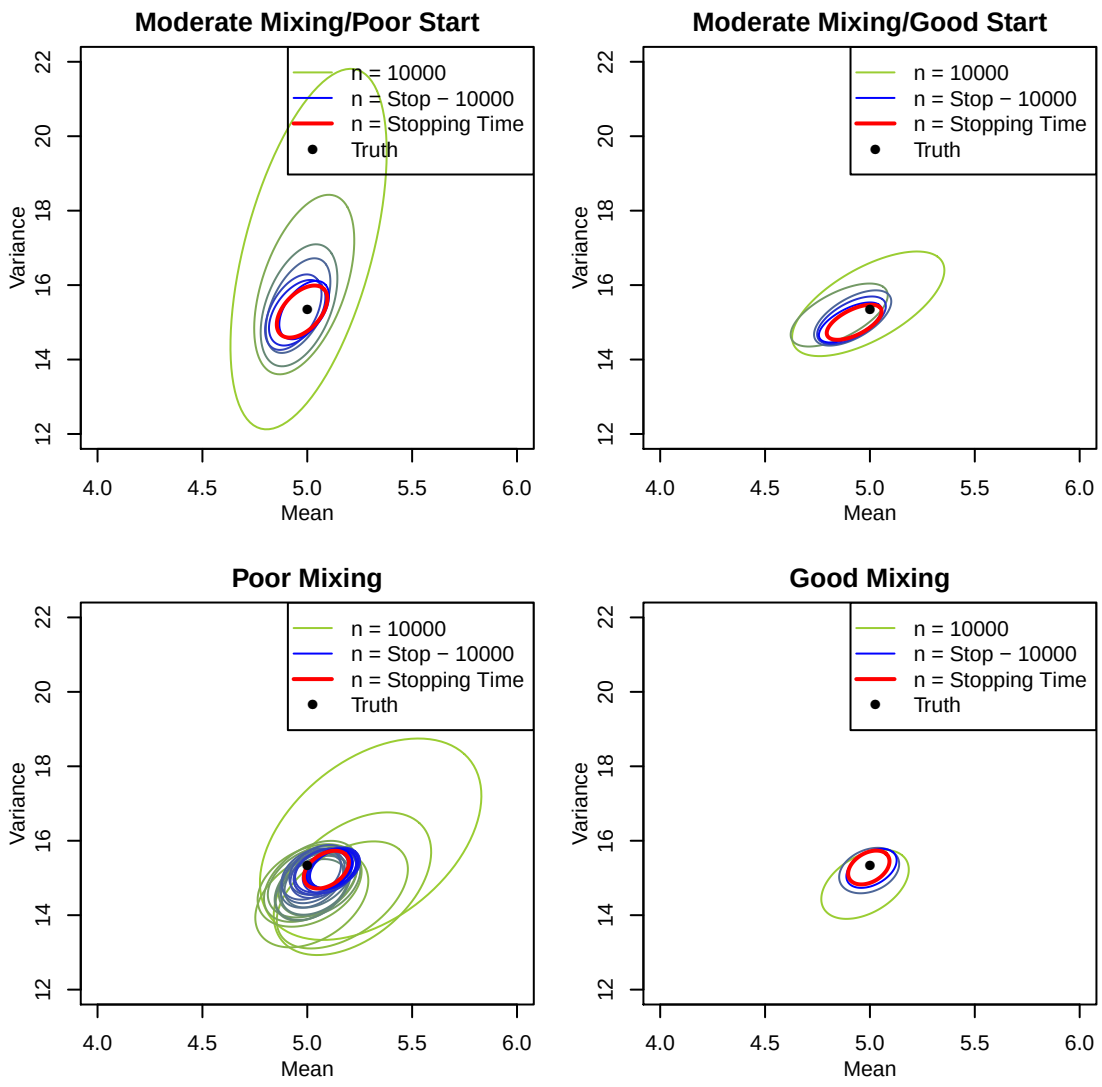

Figure 3.3: Confidence regions over increasing sample size for various samplers

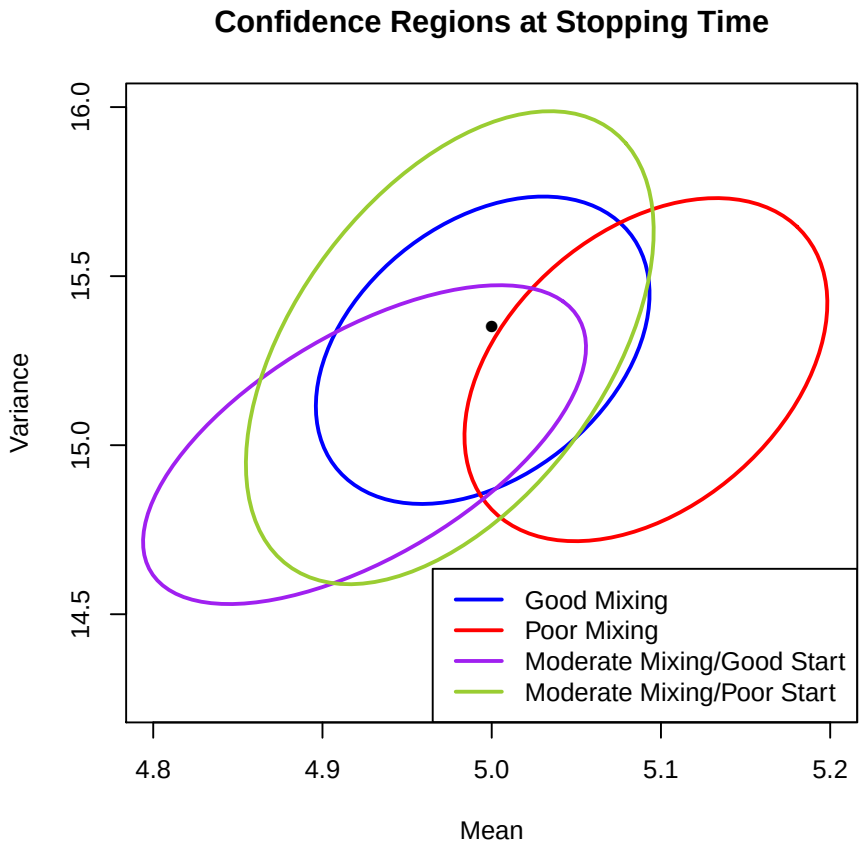

Figure 3.4: Stopping time confidence regions for various samplers

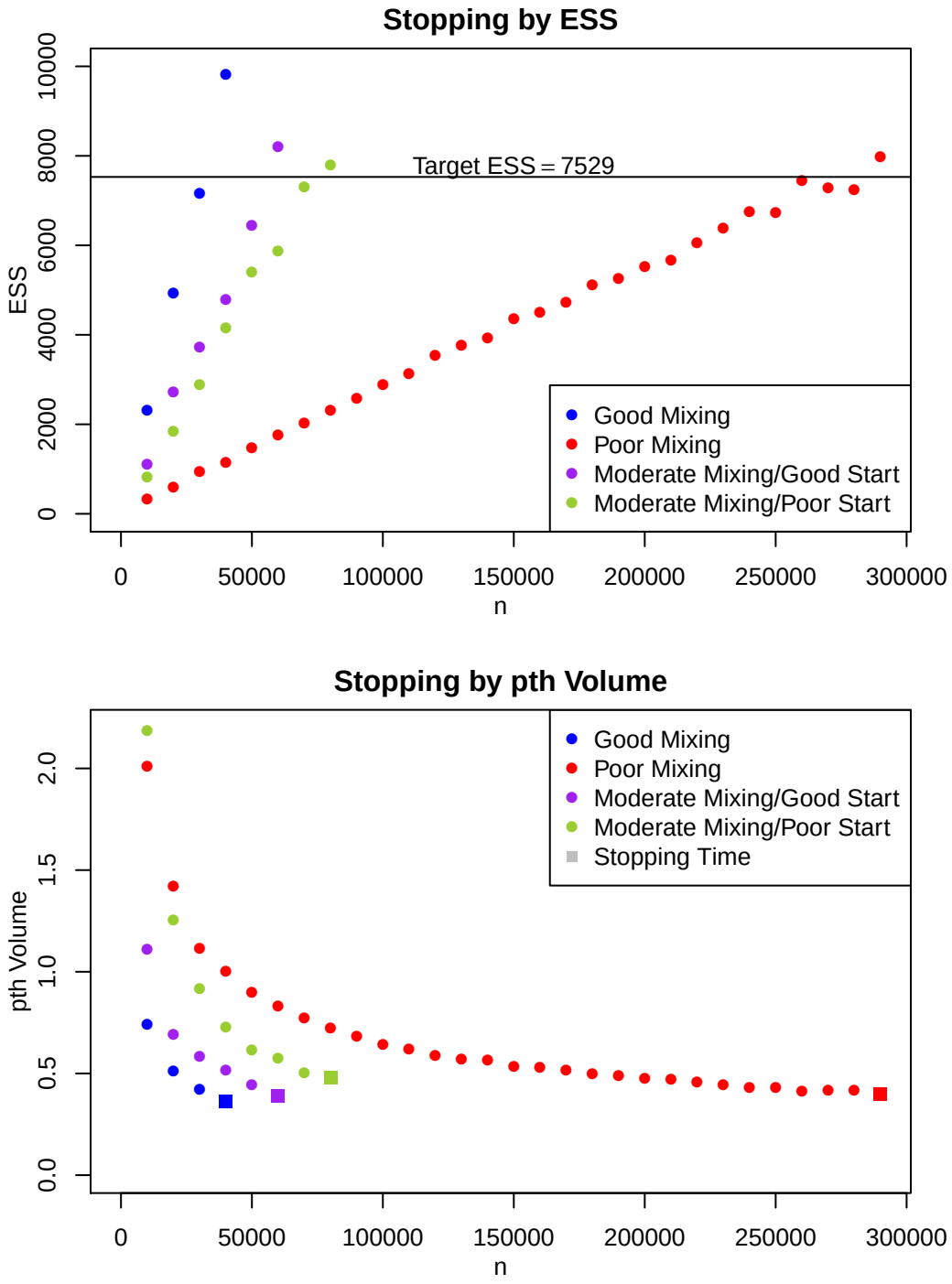

Figure 3.5: Stopping criteria over increasing sample size for various samplers

387	182	244	600	627	332	418	300
798	584	660	39	274	174	50	34
1895	158	974	345	1755	1752	473	81
954	1407	230	464	380	131	1205	

Table 3.1: LCD Projector lifetimes in projection hours

variations arising for two main reasons. The first reason is each chain has to estimate the relative standard deviation metric which takes the form of the $2p^{th}$ root of the determinate of Σ_g. The second reason is due to the number of samples collected between stopping time checks. This is most visible in the good mixing case which is close to the stopping criteria by 30,000 but is next checked at 40,000 when it might have stopped around 32,000 with smaller increments. The stopping time criteria progression in terms of volume and ESS is given in Figure 3.5. Figure 3.6 tracks the estimation of each parameter over time. Unsurprisingly, the poorly mixing chain stabilizes the slowest and still maintains a comparatively large error in its mean estimate. This is also noticeable in the confidence region plot in Figure 3.3 where the truth is not captured in the confidence region.

3.3 Reliability Example

We consider fitting a Weibull model of LCD projector lifetimes as discussed in Hamada et al. (2008) with data presented in Table 3.1. In particular, we consider the mean time to failure and the reliability after 1,500 hours of use.

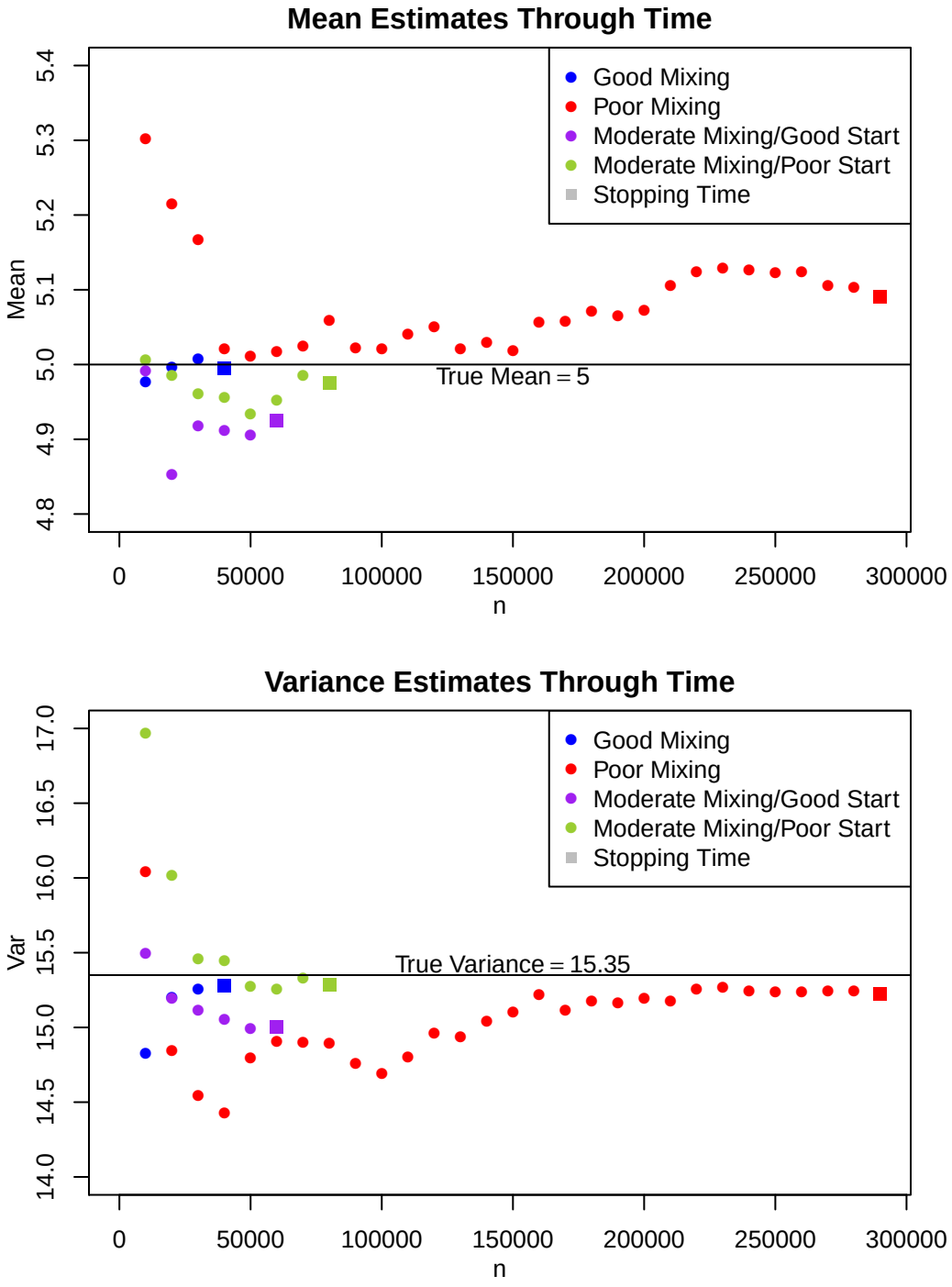

Figure 3.6: Parameter estimates over increasing sample size until stopping time for various samplers

Let T_i be the failure time of the i^{th} subject in hours. We consider the model

$$T_i|\lambda,\beta \sim Weibull(\lambda,\beta) \tag{3.2}$$

$$\lambda \sim Gamma(\alpha_\lambda = 2.5, \theta_\lambda = 2350) \tag{3.3}$$

$$\beta \sim Gamma(\alpha_\beta = 1, \theta_\beta = 1) \tag{3.4}$$

with densities

$$f(t_i|\lambda,\beta) = \lambda\beta(t)^{\beta-1}\exp\{-\lambda(t)^\beta\} \tag{3.5}$$

$$f(\lambda) = \frac{\theta_\lambda^{\alpha_\lambda}}{\Gamma(\alpha_\lambda)}\lambda^{\alpha_\lambda-1}\exp\{-\theta_\lambda\lambda\} \tag{3.6}$$

$$f(\beta) = \frac{\theta_\beta^{\alpha_\beta}}{\Gamma(\alpha_\beta)}\beta^{\alpha_\beta-1}\exp\{-\theta_\beta\beta\}. \tag{3.7}$$

The Weibull lifetimes yield a closed form expression for the mean time to failure (MTTF)

$$\text{MTTF} = \lambda^{-1/\beta}\Gamma\left(\frac{\beta+1}{\beta}\right), \tag{3.8}$$

and reliability function

$$R_T(t) = 1 - F_T(t) = \exp\{-\lambda t^\beta\}. \tag{3.9}$$

To estimate these quantities, we choose the function $g : \mathbb{R}^+ \to \mathbb{R}^+ \times \mathbb{R}^+$ such that

$$g(T) = \begin{pmatrix} \lambda^{-1/\beta}\Gamma\left(\frac{\beta+1}{\beta}\right) \\ \exp\{-\lambda 1500^\beta\} \end{pmatrix}. \tag{3.10}$$

3.3.1 Sampling

In order to sample from our approximate posterior we take a Gibbs Sampler with a Metropolis within Gibbs update for β. We then update λ by drawing from the conditional distribution

$$\lambda|\beta, t \sim Gamma(\alpha_\lambda + n, \theta_\lambda + \sum t_i^\beta). \tag{3.11}$$

To simulate β, we consider a Metropolis-Hastings random walk with increment distribution $N(0, .005)$. This increment distribution was choosen as it lead to an acceptance probability $\approx .4$, which is near the optimal rate for a one-dimensional MH random walk. As we are using a Gibbs sampler which samples from conditional distributions, we only need to specify a starting value for β. We may then sample λ as it is independent of the previous λ value conditional on the current β value. We set an initial sample size of 10,000 and a starting value $\beta = 1$.

We start by examining the trace plots of λ and β in Figure 3.7. The chain appears to be mixing fairly well as the samples appear to traverse the distribution in relatively few samples. Additionally, the starting value chosen for β appears to be well within an area of high density and seems reasonable. We next consider the auto-correlation within the chain in Figure 3.8. There is a large amount of lag correlation in each parameter. The correlation fades somewhat steadily indicating that the estimates should be reasonable provided the sample size is appropriately large. We are however interested in functions of these parameters and may be interested in the auto-correlation for the transformed samples. Here the auto-correlation decays quickly as illustrated in Figure 3.9. Also of

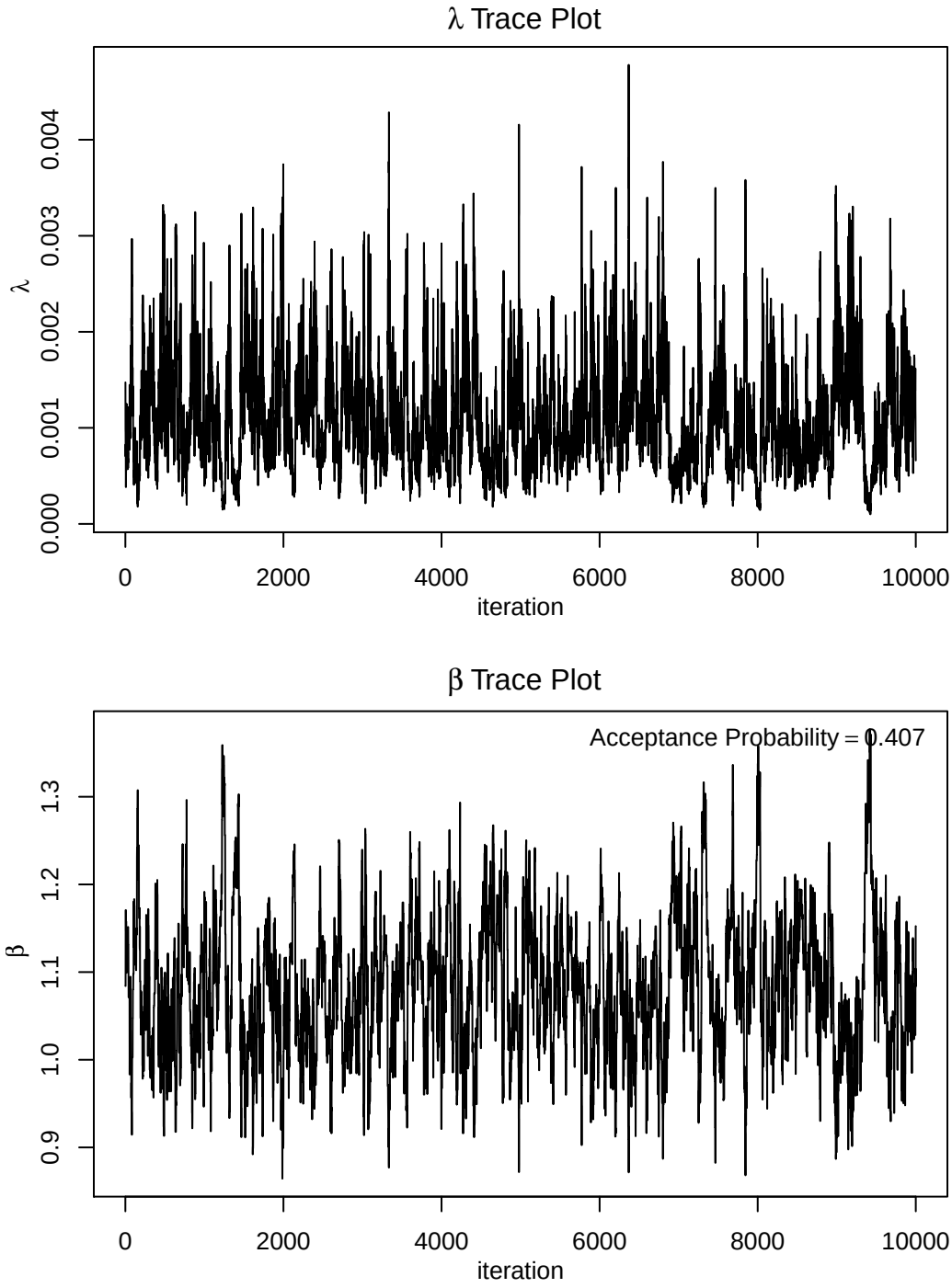

Figure 3.7: Trace plots of 10,000 MCMC samples for λ and β

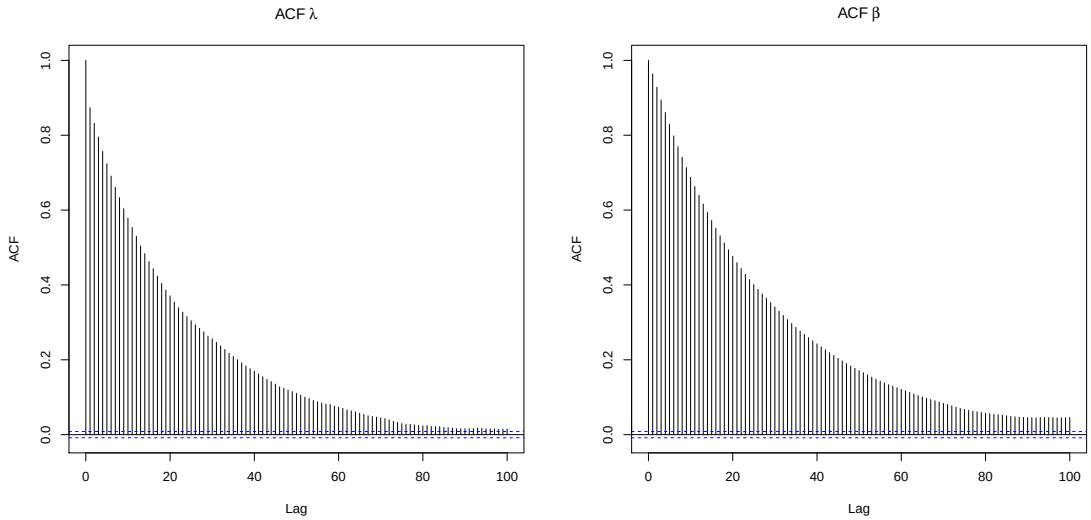

Figure 3.8: ACF plots for λ and β

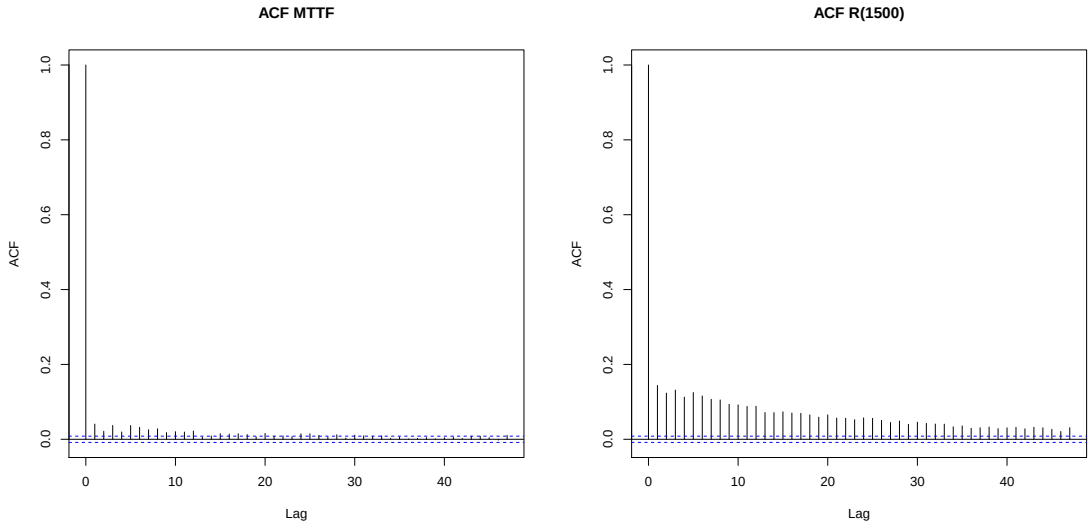

Figure 3.9: ACF plots for MTTF and $R(1500)$

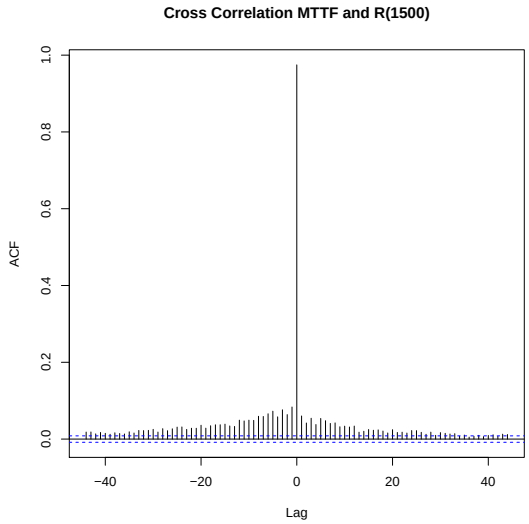

Figure 3.10: Cross Correlation plot of MTTF and $R(1500)$

interest to us is the cross-correlation between MTTF and $R_T(1500)$ in Figure 3.10. Cross-correlation measures the lag correlation between two sequences of random variables, in this case MTTF and $R(1500)$. The two quantities exhibit small amounts of lag correlation but extremely high correlation between paired terms, or $g(X_i) = (g_1(X_i), g_2(X_i))$. This indicates information about MTTF is highly informative to $R(1500)$ and that the elliptical confidence region orientation will be nearly 45°.

3.3.2 Stopping and Estimation

We now run the sampler for batches of 5,000 samples at a time until the estimated ESS surpasses 7529 which corresponds to a 95% confidence level and $\epsilon = .05$. This occurs after 55,000 samples and yields the estimates MTTF = 597.658 with standard error .654 and $R_T(1500) = .0737$ with standard error .00042. The final estimated effective sample size

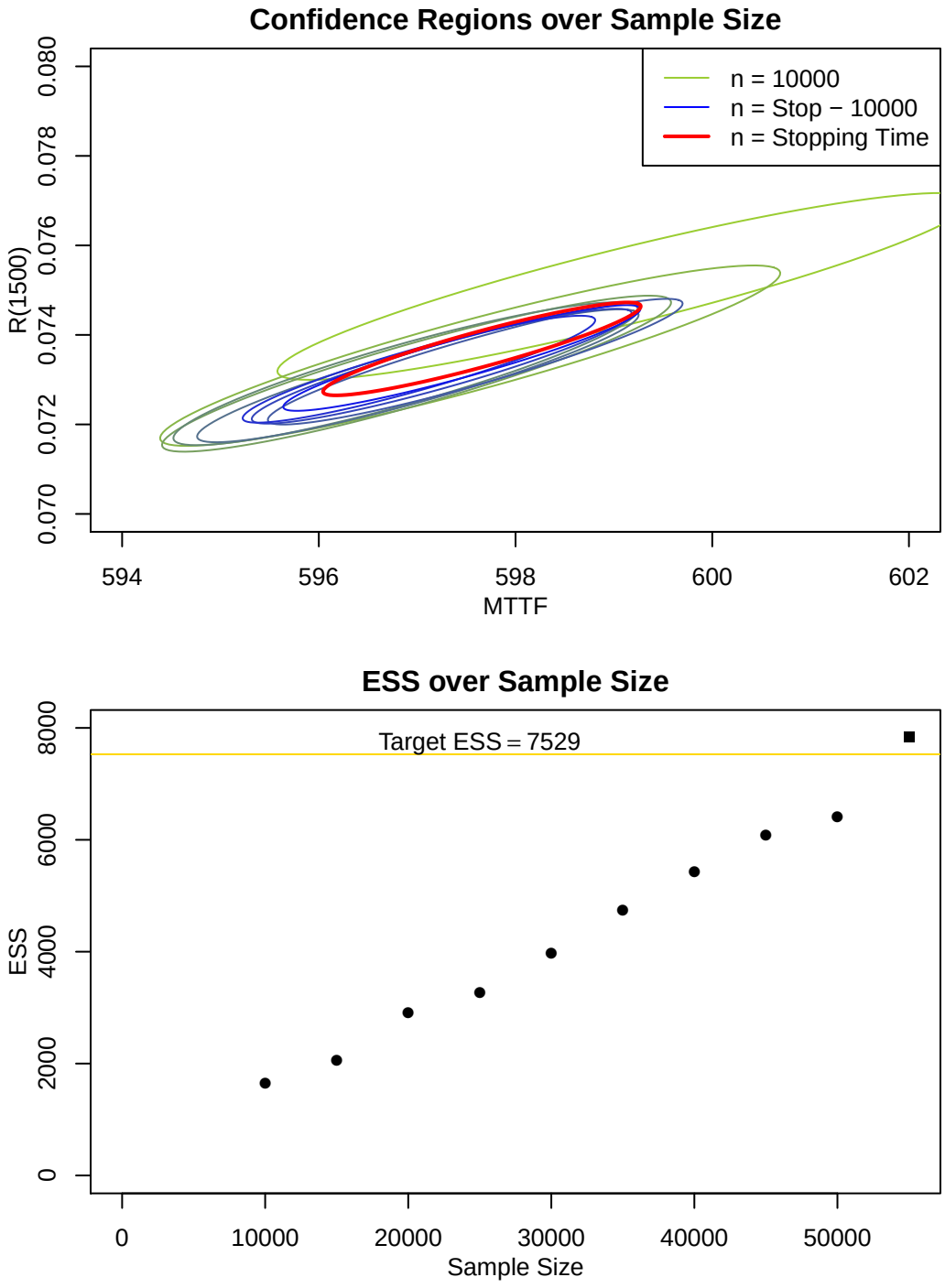

Figure 3.11: Confidence Regions and ESS for various sample sizes

is 7840. The confidence region and ESS progression are pictured in Figure 3.11. We may conclude that the mean time to failure for a LCD projector is just short of 600 and that it is unlikely for it to last at least 1,500 hours.

Chapter 4

Monte Carlo Visualizations

The following is from the submitted paper "New Visualizations for Monte Carlo Simulations" Robertson et al. (2019).

4.1 Introduction

The analysis of output obtained from a Monte Carlo simulation is an essential part of ensuring reliable simulation studies. We propose new visualization tools based on simultaneous confidence intervals with a desired confidence level, which are narrower than conservative approaches (e.g. Bonferonni). Our focus is on Monte Carlo settings including sampling independent and identically distributed (i.i.d.) random variables and correlated sampling of random variables arising from strongly mixing sequences or Markov chain Monte Carlo (MCMC). However, the conditions we require allows our approach to be used more broadly.

Suppose possibly correlated Monte Carlo samples are used to estimate features of a distribution. This scenario frequently arises in physical and mathematical problems when other approaches are intractable. Typically, the interest is in estimating several parameters simultaneously, which may include both expectations and quantiles. Further, there is often dependence between the parameters. For example, simulation studies often compare statistical techniques or models using i.i.d. replications to estimate multiple quantities simultaneously (e.g. estimation error and prediction error). Alternatively, a Bayesian may be interested in estimating multiple posterior means along with credible intervals simultaneously.

Due to variability in repeated simulations, it is imperative to include estimated simulation uncertainty with feature estimates to give a sense of the simulation's quality and reliability. Reporting simulation uncertainty usually amounts to reporting Monte Carlo standard errors or confidence intervals. We note reporting only the Monte Carlo sample size (or effective sample size) does not, in general, provide the necessary indication of simulation uncertainty; see Flegal et al. (2008) and Koehler et al. (2009) for additional discussion.

When standard errors are reported, they are almost always univariate, ignoring multiplicity and dependence among parameters. One can address this by providing multivariate confidence regions when estimating expectations Vats et al. (2018); Vats et al. (2019), but this does not address the estimation of quantiles or the simultaneous estimation of means and quantiles. More importantly, visualization and interpretation of confidence regions is problematic for multi-dimensional quantities. Most current visualizations ignore Monte Carlo uncertainty altogether and report quantities from empirical marginal distributions, e.g., by using sample boxplots. Software such as `Stan` (Stan Development Team,

2018) and `tidybayes` (Kay, 2018) suffer from the same drawbacks when reporting credible intervals or prediction intervals. We provide a flexible and novel class of visualizations for assessing the quality of simultaneous estimation of means and quantiles from Monte Carlo simulations.

Consider a motivating example estimating the mean and $(.10, .90)$-quantiles for a three component mixture of normal densities. We simulate draws using a Metropolis-Hastings (MH) random walk to estimate the 3-dimensional quantity of interest and its corresponding 3×3 asymptotic covariance matrix. Figure 4.1 shows 90% simultaneous confidence intervals superimposed on a plot containing an empirical density estimate. Figure 4.1 indicates substantial uncertainty around the estimates when only 1,000 samples are drawn and that 50,000 samples provide far more certainty. A closer examination reveals that the confidence regions displayed around each quantity of interest have different lengths. Figure 4.1 enables practitioners to visualize simultaneous simulation uncertainty. It clearly illustrates both the variability of the distribution, π, and the uncertainty in estimation from using a Monte Carlo simulation without overemphasizing point estimates and hence follows suggested practice (see e.g. Shubin, 2015). Our visualization tools apply much more broadly than the motivating example in Figure 4.1. For example, one can consider more than one parameter, additional means and quantiles, or even boxplots. We illustrate these in a series of examples, but many other uses of our methodology are possible.

The techniques proposed here provide simultaneous intervals for any combination of quantiles and expectations. That is, the parameters of interest will be in their respective confidence intervals simultaneously with the desired level of significance, say $1 - \alpha$. It

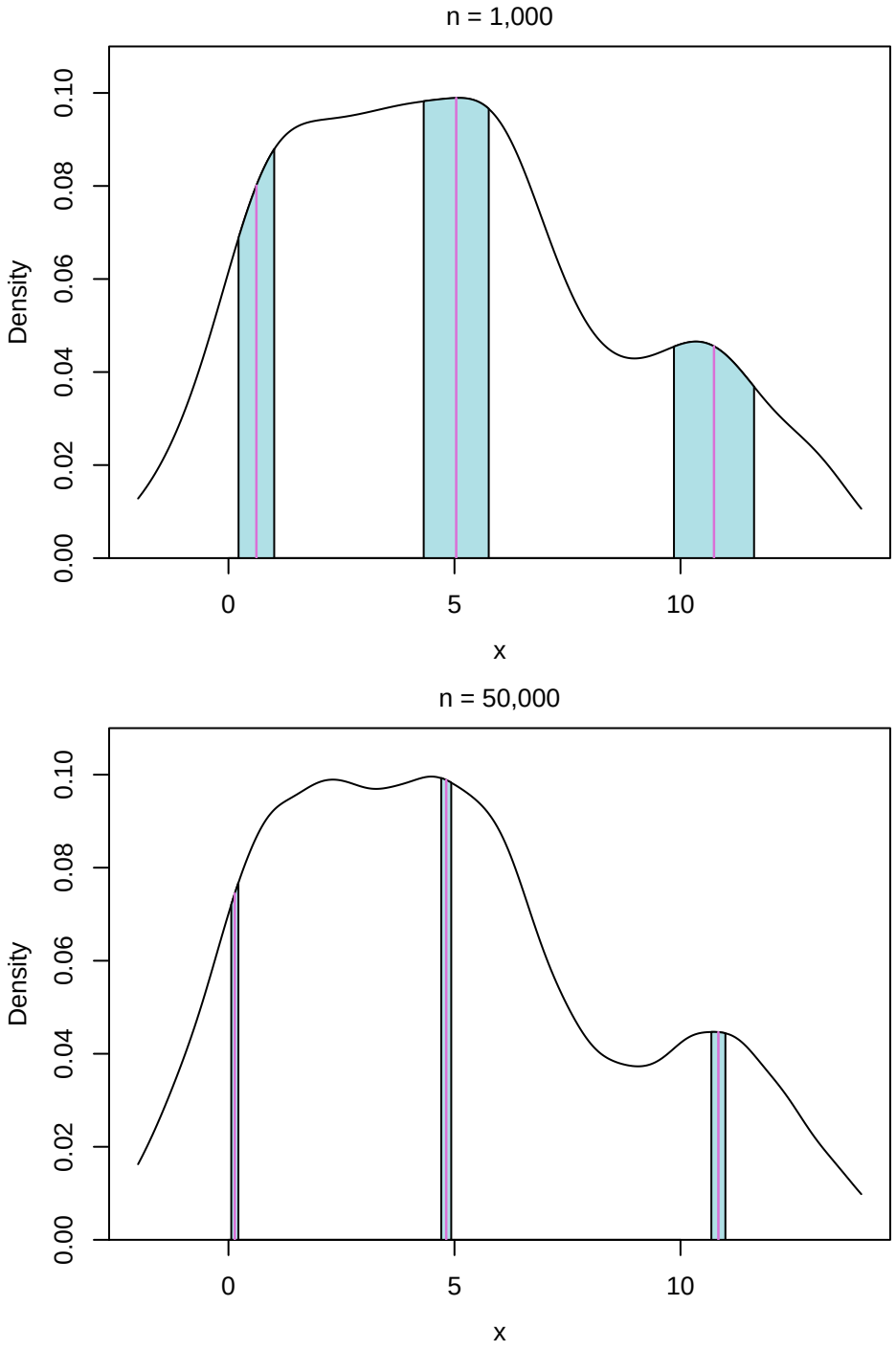

Figure 4.1: Visualization of simultaneous uncertainty bounds for the mean and $(.10, .90)$-quantiles.

is important to note that our results are neither pointwise intervals each having coverage probability $1-\alpha$ nor conservative simultaneous intervals (e.g. Bonferonni) where the overall coverage probability often greatly exceeds $1-\alpha$.

To construct simultaneous confidence intervals with the desired level of significance, we propose a parametric approach that utilizes joint asymptotic normality for the Monte Carlo error of expectations and quantiles. To this end, our first result establishes asymptotic normality of sample means and quantiles where we also provide estimators for the covariance matrix of the asymptotic normal distribution. Our result holds for settings where a strong law, a central limit theorem (CLT) for sample means, and a Bahadur (1966) quantile representation hold. We illustrate sufficient conditions for i.i.d. sampling, strongly mixing processes, and Markov chain sampling. Therefore, techniques described here are widely applicable.

Given joint asymptotic normality, one could construct marginal confidence intervals, conservative simultaneous intervals, or confidence regions with $1-\alpha$ coverage based on an ellipsoid. However, each of these suffer from drawbacks mentioned previously. Instead, we construct simultaneous intervals using joint asymptotic normality to obtain a $1-\alpha$ confidence hyper-rectangle. As we illustrate later, our algorithm reduces to a univariate optimization problem. These simultaneous intervals form the building blocks for the proposed visualization techniques.

We implement our methodology and visualizations in three examples. Our first example continues the three component mixture of normals. Since the truth is known in this example, we assess finite sample performance of simultaneous confidence intervals by comparing coverage probabilities with other univariate and multivariate methods. This

Monte Carlo simulation study illustrates the utility of our visualization tools. Our second example considers another Monte Carlo simulation study comparing estimation error for different regression methods, where we equip the standard side-by-side boxplots with simultaneous confidence intervals. The resulting plot is especially useful in indicating whether sufficient Monte Carlo replications have been run. Our third example illustrates a couple of different marginal-friendly visualizations for Bayesian inference on two sets of features from a posterior distribution.

The remainder is organized as follows. In Section 4.2, we present the joint asymptotic distribution of sample means and quantiles for both independent and dependent sequences. Section 4.3 presents a univariate optimization method for obtaining simultaneous confidence intervals. Section 4.4 presents example visualizations for the three component mixture of normals, two Monte Carlo simulations studies, and a Bayesian analysis. We conclude with a discussion in Section 4.5.

4.2 Joint asymptotic distribution

Consider a probability distribution, π, with support $\mathsf{X} \subseteq \mathbb{R}^d, d \geq 1$. We develop a joint asymptotic distribution for the estimators of p_1 expectations and p_2 quantiles associated with π using a process $X = \{X_1, X_2, \dots\}$. We will be more specific below about what must be assumed about X.

First, suppose $g : \mathsf{X} \to \mathbb{R}^{p_1}$ and consider estimating p_1 expectations

$$\theta_g = E_\pi \left[g(X) \right] = \int_{\mathsf{X}} g(x) \pi(dx) \; .$$

We assume that with probability one, as $n \to \infty$,

$$\bar{g}_n = \frac{1}{n} \sum_{j=1}^{n} g(X_j) \to \theta_g \tag{4.1}$$

and that the sampling distribution for the Monte Carlo error, $\bar{g}_n - \theta_g$, can be obtained via a CLT. That is, there exists a $p_1 \times p_1$ positive definite matrix Σ_g such that, as $n \to \infty$,

$$\sqrt{n}(\bar{g}_n - \theta_g) \xrightarrow{d} N_p(0, \Sigma_g)\,. \tag{4.2}$$

The formulation and estimation of Σ_g will be discussed in detail later.

Now consider estimating p_2 quantiles associated with π. These quantiles could be with respect to any functional of X and not merely the components of X or $g(X)$. Unfortunately, this level of generality leads to somewhat cumbersome notation. Define a function $h : \mathsf{X} \to \mathbb{R}^{p_2}$ such that $h(X) = (h_1(X), \ldots, h_{p_2}(X))'$ where each $h_i(X)$ represents some functional of interest. Further, define $Q = (q_1, \ldots, q_{p_2})'$ where q_i is the desired quantile from $h_i(X)$. We note that two functionals $h_i(X)$ and $h_j(X)$ may be the same, as it is often the case that multiple quantiles of the same functional are being estimated. For $V = h(X)$, the p_2 quantiles of interest are associated with marginal distribution functions of V, say $F_{h_i}(v)$, which we assume are absolutely continuous with continuous densities $f_{h_i}(v)$. Finally, define the q_i-quantile associated with F_{h_i} as

$$\xi_{q_i} = F_{h_i}^{-1}(q_i) = \inf\{v : F_{h_i}(v) \geq q_i\},$$

where our interest is in estimating the vector of p_2 quantiles denoted

$$\phi = \left(\xi_{q_1}, \ldots, \xi_{q_{p_2}}\right)'.$$

Estimation is straightforward using marginal order statistics from $h(X)$. That is, let $\hat{\xi}_{q_i} = h_i(X)_{\lceil nq_i \rceil:n}$ be the $\lceil nq_i \rceil^{th}$ order statistic of $h_i(X)$ and denote the vector of p_2 estimated quantiles as

$$\hat{\phi}_n = \left(\hat{\xi}_{q_1}, \ldots, \hat{\xi}_{q_{p_2}}\right)'.$$

A strong law for estimators of $F_{h_i}(v)$ is enough to ensure $\hat{\phi}_n \to \phi$ as $n \to \infty$ with probability one (see e.g. Doss et al., 2014; Serfling, 1981).

The joint asymptotic distribution for p_1 expectations and p_2 quantiles can be established using the Bahadur quantile representation. To this end, define empirical distributions for F_{h_i} as

$$\bar{F}_{h_i}(v) = \frac{1}{n} \sum_{j=1}^{n} I(h_i(X_j) \leq v),$$

with vectorized representation

$$\bar{F}_h(v) = \left(\bar{F}_{h_1}(v_1), \ldots, \bar{F}_{h_{p_2}}(v_{p_2})\right)'.$$

Since probabilities can be expressed as expectations of indicator functions, the strong law ensures that, as $n \to \infty$, $\bar{F}_h(\phi) \to Q$ with probability one. Further, the CLT at (4.2) can

be re-expressed as

$$\sqrt{n}\left(\begin{pmatrix}\bar{g}_n\\ 1-\bar{F}_h(\phi)\end{pmatrix}-\begin{pmatrix}\theta_g\\ 1-Q\end{pmatrix}\right)\xrightarrow{d}N_{p_1+p_2}\left(0,\Sigma=\begin{pmatrix}\Sigma_g & \Sigma_{gh}\\ \Sigma_{hg} & \Sigma_h\end{pmatrix}\right) \tag{4.3}$$

as $n\to\infty$, where Σ and Σ_h are positive definite covariance matrices, and $\Sigma_{gh}=\Sigma_{hg}'$ are the $p_1\times p_2$ cross-covariance matrices. Next, consider the Bahadur (1966) quantile representation

$$\hat{\xi}_{q_i}=\xi_{q_i}+\frac{\left(1-\bar{F}_{h_i}(\xi_{q_i})\right)-(1-q_i)}{f_{h_i}(\xi_{q_i})}+r_{n,q_i}, \tag{4.4}$$

where r_{n,q_i} is $o_p(n^{-1/2})$. Then the joint distribution for estimation of p_1 expectations and p_2 quantiles is established in the following theorem.

Theorem 2 *Suppose a strong law, CLT, and Bahadur quantile representation hold as at* (4.1), (4.3), *and* (4.4), *respectively. Let A_h be a $p_2\times p_2$ diagonal matrix with i^{th} diagonal elements $f_{h_i}(\xi_{q_i})$. If*

$$\Lambda=\begin{pmatrix}I_{p_1} & 0_{p_1\times p_2}\\ 0_{p_2\times p_1} & A_h\end{pmatrix},$$

then, as $n\to\infty$,

$$\sqrt{n}\begin{pmatrix}\bar{g}_n-\theta_g\\ \hat{\phi}_n-\phi\end{pmatrix}\xrightarrow{d}N\left(0,\Lambda^{-1}\Sigma\Lambda^{-1}\right). \tag{4.5}$$

Proof. Let $R_n=\left(r_{n,q_1},\ldots,r_{n,q_{p_2}}\right)'$, then by (4.4),

$$\left(1-\bar{F}_h(\phi)\right)-(1-Q)=A_h\left(\hat{\phi}_n-\phi\right)+A_hR_n\xrightarrow{P}A_h\left(\hat{\phi}_n-\phi\right). \tag{4.6}$$

Combining (4.3) and (4.6) we have

$$\sqrt{n}\left(\begin{pmatrix}\bar{g}_n \\ 1-\bar{F}_h(\phi)\end{pmatrix}-\begin{pmatrix}\theta_g \\ 1-Q\end{pmatrix}\right)=\sqrt{n}\begin{pmatrix}\bar{g}_n-\theta_g \\ A_h\left(\hat{\phi}_n-\phi\right)\end{pmatrix}+o_p(1)$$

$$=\sqrt{n}\Lambda\begin{pmatrix}\bar{g}_n-\theta_g \\ \hat{\phi}_n-\phi\end{pmatrix}+o_p(1).$$

■

Using Theorem 2 requires estimation of Λ^{-1} and Σ. Since Λ is a diagonal matrix with non-zero diagonals, its inverse Λ^{-1} is readily available. Then kernel density estimators with a Gaussian kernel can be used to estimate $f_{h_i}(\hat{\xi}_{q_i})$, and hence estimate Λ. The matrix Σ requires more specific attention since i.i.d. and dependent sampling schemes yield different structures of Σ. We discuss these differences in the next two sections, both of which are implemented in `R` as part of the supplementary material.

4.2.1 Independent sequences

Suppose $X = \{X_1, X_2, \dots\}$ are i.i.d. realizations from π. The strong law and CLT hold for estimating θ_g provided, $E_\pi\|g\| < \infty$ and $E_\pi\|g\|^2 < \infty$, respectively. Since $|\bar{F}_h(\cdot)| \leq 1$, these conditions also ensure the joint distribution at (4.3). The Bahadur quantile representation at (4.6) requires $0 < f_{h_i}(\xi_{q_i}) < \infty$ and continuity of f_{h_i} in a neighborhood of ξ_{q_i} for all i (Ghosh, 1971). This is weaker than our prior assumption that $F_{h_i}(v)$ is absolutely continuous with continuous density $f_{h_i}(v)$, which is required for more general processes.

Joint asymptotic distributions for i.i.d. sampling have received substantial attention. Laplace found the joint asymptotic distribution of the sample mean and the sample median (Stigler, 1973). Ferguson (1998) provides a nice derivation for a sample mean and an arbitrary quantile along with an expression of the covariance (also see Lin et al., 1980). This result can be generalized to sample means and arbitrary quantiles associated with distinct marginal random variables. Babu and Rao (1988) provide an expression of the covariance between two quantiles. These results yield an exact, albeit complicated, expression for $\Lambda^{-1}\Sigma\Lambda^{-1}$ from (4.5). Suppose $Y_j = \left(g(X_j), I(h(X_j) > \hat{\phi}_n)\right)'$, then our `R` implementation estimates Σ by the sample covariance of $\{Y_1, Y_2, \ldots, Y_n\}$.

4.2.2 Dependent sequences

This section provides conditions for the strong law, CLT, and Bahadur quantile representation when $X = \{X_1, X_2, \ldots\}$ is a dependent sequence. Specifically, we consider a stationary strongly mixing (or α-mixing) setting and its connection to MCMC simulations. The discussion here is not exhaustive and does not provide minimal known conditions; see Bradley (1986, 2005) and Jones (2004) for more information.

The strong law holds for estimating θ_g, provided $E_\pi\|g\| < \infty$ and X is strongly mixing (Blum and Hanson, 1960). Since Harris ergodic Markov chains are strongly mixing, the strong law holds under the same moment conditions (Jones, 2004; Meyn and Tweedie, 2009). Ibragimov (1962) provides a CLT if there exists a $\delta > 0$ such that $E_\pi\|g\|^{2+\delta} < \infty$ and X mixes sufficiently fast. Corollary 2 of Jones (2004) (using additional results from Ibragimov and Linnik, 1971) provides a CLT for geometrically and polynomial ergodic Markov chains. Yoshihara (1995) provides the Bahadur quantile representation at (4.6)

when $F_{h_i}(v)$ is absolutely continuous with continuous density $f_{h_i}(v)$ such that $0 < f_{h_i}(\xi_{q_i}) < \infty$ and X mixes sufficiently fast. Such a mixing condition holds for polynomial ergodic Markov chains (Jones, 2004). Wang et al. (2011) weakens these mixing conditions, but their Bahadur quantile representation is not applicable for MH algorithms.

Recall that $Y_j = \left(g(X_j), I(h(X_j) > \hat{\phi})\right)'$. For a stationary strongly mixing sequence, an expression for Σ is

$$\Sigma = \mathrm{Cov}(Y_j, Y_j) + \sum_{i=1}^{\infty} \left[\mathrm{Cov}(Y_j, Y_{j+i}) + \mathrm{Cov}(Y_j, Y_{j+i})'\right] . \tag{4.7}$$

Estimation of Σ at (4.7) is a well studied problem and may be accomplished using batch means (Chen and Seila, 1987; Vats et al., 2019), weighted batch means (Liu and Flegal, 2018b), spectral variance (Andrews, 1991; Priestley, 1981; Vats et al., 2018), initial sequence (Dai and Jones, 2017), recursive (Chan and Yau, 2017) or regenerative sampling estimators (Hobert et al., 2002; Seila, 1982).

Due to computational simplicity we restrict our attention to batch means estimators with batch size equal to $\lfloor \sqrt{n} \rfloor$. Let $n = ab$ where a is the number of batches and b is the batch size. The mean for the k^{th} batch is $\bar{Y}_k(b) = b^{-1} \sum_{t=1}^{b} Y_{kb+t}$ and the overall mean is $\bar{Y} = a^{-1} \sum_{k=1}^{a} \bar{Y}_k(b)$. Then the batch means estimator with batch size b is

$$\hat{\Sigma} = \frac{b}{a-1} \sum_{k=0}^{a-1} (\bar{Y}_k(b) - \bar{Y})(\bar{Y}_k(b) - \bar{Y})' .$$

4.3 Simultaneous confidence intervals

We now develop simultaneous confidence intervals for the p-dimensional vector $(\theta_g, \phi)'$ having a $1-\alpha$ coverage level. The general procedure will use lower and upper bounds on the confidence region and search over possible values between.

Confidence regions in multivariate settings often take one of two approaches, a region of minimum volume or intervals based on marginal distributions. A minimum volume region takes an elliptical form for a limiting multivariate normal distribution. The location of this ellipsoid is determined by the mean vector while the shape is determined by the eigenvalues and eigenvectors of the covariance matrix. Unfortunately, visualizing an ellipsoid is challenging when $p \geq 4$, hence they are rarely presented in Monte Carlo output analysis. It is more common to report confidence intervals based on the marginal distributions creating a hyper-rectangular confidence region. The location of the hyper-rectangle is determined by the mean vector while the diagonals of the covariance matrix, i.e. the marginal variances, determine the length of each side. The popularity of hyper-rectangles stems from the fact they can be easily reported and visualized. We improve hyper-rectangular confidence regions by incorporating the full covariance information to determine appropriate marginal confidence levels yielding a simultaneous confidence level of $1-\alpha$.

First, consider a hyper-rectangular confidence region based on marginal intervals each with confidence level $1-\alpha$, i.e. intervals not adjusted for multiplicity. Denote such a region as C_α^{LB}. If the random variables, $(\bar{g}_n, \hat{\phi}_n)$, are perfectly correlated, this will yield the correct overall coverage level while yielding undercoverage in any other case. The region C_α^{LB} will act as the lower bound for our coverage level. We consider Bonferroni corrected

simultaneous intervals as an upper bound, denoted C_α^{UB}. If all components of $(\bar{g}_n, \hat{\phi}_n)$ are uncorrelated, C_α^{UB} yields the approximately correct overall coverage level. However, overcoverage occurs in any other case. Assuming a fixed estimate of $\Lambda^{-1}\Sigma\Lambda^{-1}$, intervals of these forms maintain a constant aspect ratio in the axes. That is, the ratio of the lengths of the intervals in C_α^{LB} and C_α^{UB} is the same for all components. This property allows searching for a hyper-rectangular confidence region between C_α^{LB} and C_α^{UB} with the correct confidence level $1-\alpha$ to reduce to a one-dimensional line search. A sketch of two-dimensional confidence regions and the appropriate line search is pictured in Figure 4.2.

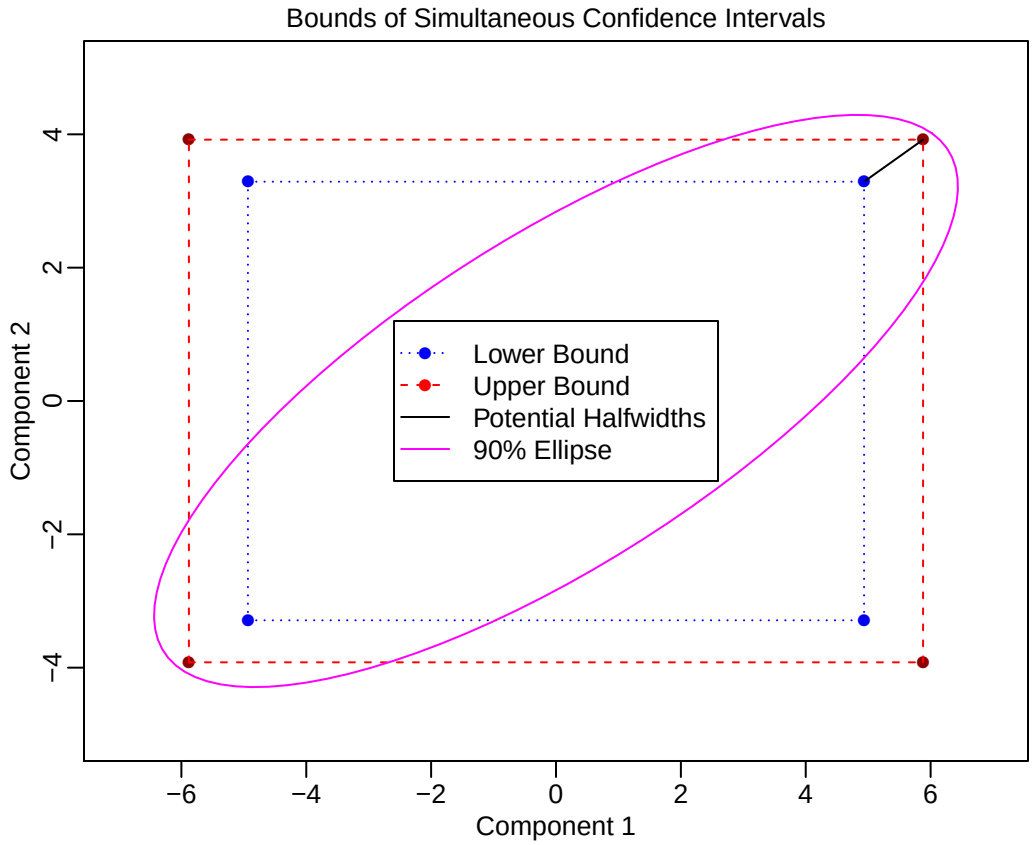

Figure 4.2: Plot of C_α^{LB} (blue) and C_α^{UB} (red) from a 90% confidence region for a bivariate normal distribution with component variances 9 and 4. The black line in the first quadrant indicates the potential search values to achieve the desired overall coverage level.

Consider a hyper-rectangular confidence region $\mathcal{C}_\alpha^{SI}(z)$ with critical value z constructed using estimators of $\Lambda^{-1}\Sigma\Lambda^{-1}$ described in Section 4.2. We are intersted in $P(X \in \mathcal{C}_\alpha^{SI}(z))$ where X is approximately from a multivariate normal distribution with covariance matrix $\hat{\Lambda}^{-1}\hat{\Sigma}\hat{\Lambda}^{-1}$. For $C_\alpha^{LB} = \mathcal{C}_\alpha^{SI}(z_{1-\alpha/2})$, we have $P(X \in C_\alpha^{LB}) \leq (1-\alpha)$ and for $C_\alpha^{UB} = \mathcal{C}_\alpha^{SI}(z_{1-\alpha/2p})$, we have $P(X \in C_\alpha^{UB}) \geq (1-\alpha)$. Since $P(X \in \mathcal{C}_\alpha^{SI}(z))$ is strictly increasing as z increases, we can use the bisection method between $z_{1-\alpha/2}$ and $z_{1-\alpha/2p}$ to find z^* such that $P(X \in \mathcal{C}_\alpha^{SI}(z^*)) \approx (1-\alpha)$.

The solution using the bisection method will be up to some error tolerance $\epsilon = P(X \in \mathcal{C}_\alpha^{SI}(z^*)) - (1-\alpha)$. The choice of ϵ deserves some care depending on how $P(X \in \mathcal{C}_\alpha^{SI}(z))$ is calculated. We rely on the function `pmvnorm` from the `R` package `mvtnorm` (Genz et al., 2018) which provides an error bound on the probability calculated. We conducted several simulations varying covariance, dimension, and probability values. The largest error recorded was approximately .002 with most errors less than .001. Results from this study are available upon request. We recommend setting $\epsilon = .001$ or $\epsilon = .002$.

4.4 Example visualizations

This section demonstrates the flexibility of our class of visualizations in various Monte Carlo simulation settings. In each example, we identify a combination of means and quantiles of interest, obtain $1-\alpha$ level simultaneous confidence intervals, and integrate the intervals within a standard plot. The full `R` code is available as part of the supplementary material to ensure reproducibility of simulations and plots presented below.

4.4.1 Mixture of normal distributions

This section provides details for the mixture of normal distributions example from the introduction. In the subsequent section we will demonstrate the accuracy of our procedure by estimating coverage probabilities via a Monte Carlo simulation. Let X be a random variable distributed according to a mixture of 3 normal distributions, with density

$$f_X(x) = .3f_1(x; 1, 2.5) + .5f_2(x; 5, 4) + .2f_3(x; 11, 3). \tag{4.8}$$

where $f_j(x; \mu_j, \sigma_j^2)$ is the density of the j^{th} mixture component with mean μ_j and variance σ_j^2. We consider simultaneous estimation of the mean, .10 quantile, and .90 quantile denoted μ, $\xi_{.10}$, and $\xi_{.90}$, respectively. Specifically, we have $(\theta_g, \phi)' = (\mu, \xi_{.10}, \xi_{.90})$ from (4.5). If f_X is a posterior density, then $(\xi_{.10}, \xi_{.90})$ would be characterized as an 80% credible interval.

We collect i.i.d. samples from (4.8) and estimate $(\theta_g, \phi)'$ and $\Lambda^{-1}\Sigma\Lambda^{-1}$ as described in Section 4.2. We then calculate $C_\alpha^{SI}(z^*)$ at the 90% confidence level and estimate the density to create our visualization in the top row of Figure 4.3. The estimates of $(\theta_g, \phi)'$ are represented by purple lines with the blue region around each estimate representing the simultaneous simulation uncertainty. As the number of samples increases, simulation uncertainty decreases. Another point of interest is in the different amount of simulation uncertainty surrounding $\xi_{.10}$ and $\xi_{.90}$. The shape of the density is asymmetric, thus the two quantiles occur at different density values. This affects the value of Λ^{-1} and contributes to the different lengths of the error regions around each estimate.

To illustrate our methods for a dependent sampling case we use a random walk MH sampler with proposal distribution $N(0, 9)$. We estimate $(\theta_g, \phi)'$ and $\Lambda^{-1}\Sigma\Lambda^{-1}$ as described in Section 4.2. Simultaneous confidence intervals are presented in the bottom row of Figure 4.3 where the MCMC plots contain notably more simulation uncertainty than the i.i.d. case. This occurs due to within chain correlation captured by the infinite sum at (4.7). One measure of this is effective sample size (ESS), which estimates how many i.i.d. samples a correlated sample is equivalent to. Our MH sampler had an ESS (Vats et al., 2019) to n ratio of about .2, hence it is unsurprising the i.i.d. $n = 10,000$ and MCMC $n = 50,000$ plots in Figure 4.3 illustrate similar levels of simulation uncertainty.

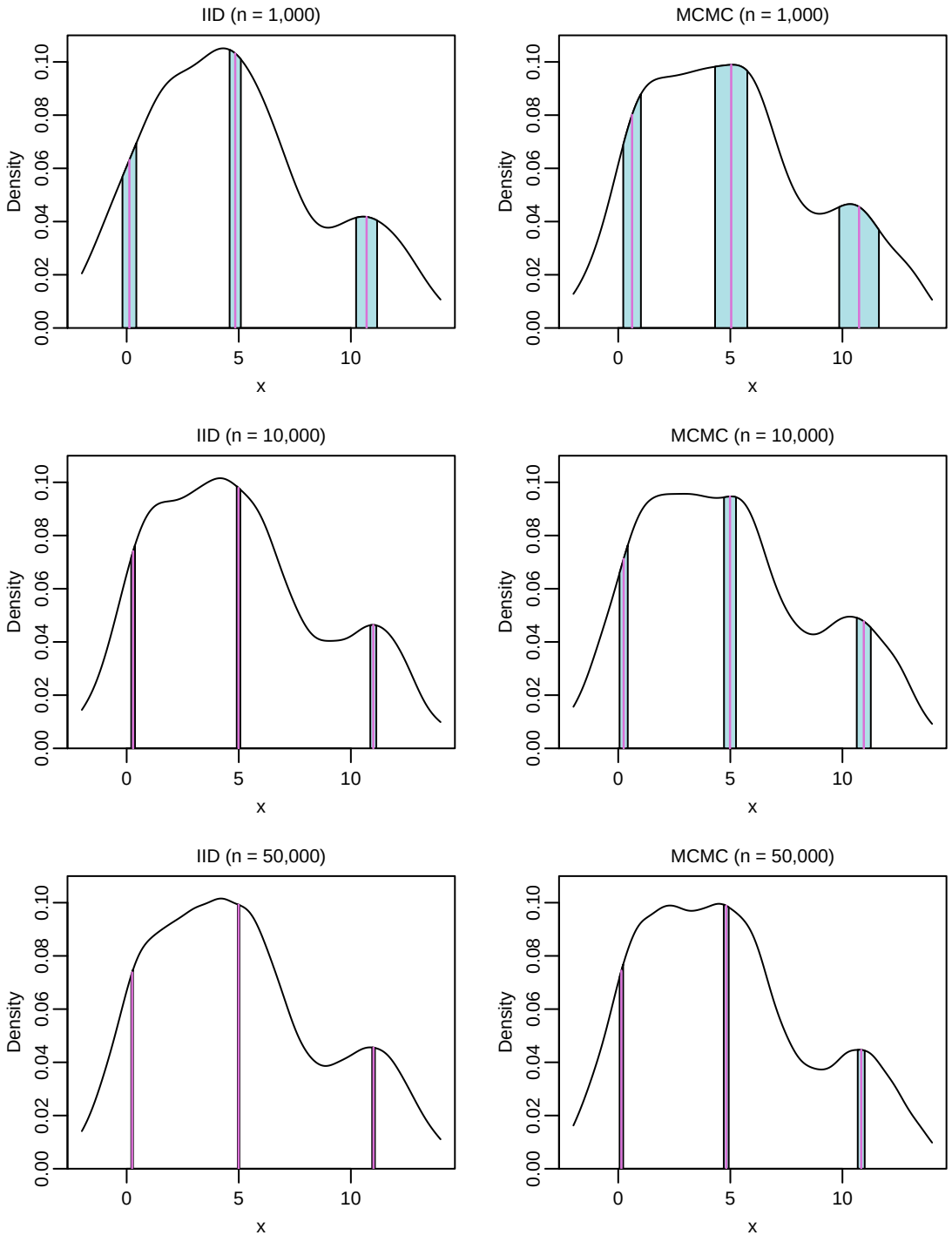

Figure 4.3: Simultaneous 90% confidence intervals of the mean, .10 quantile, and .90 quantile from a mixture of normal distributions.

4.4.2 Coverage probabilities

This section continues the mixture normal distributions where we illustrate two points. First, we demonstrate simultaneous confidence intervals $\mathcal{C}_\alpha^{SI}(z^*)$ have the correct coverage probability in finite samples via a Monte Carlo simulation. Second, we illustrate the utility of our visualization tools for this Monte Carlo simulation.

To examine the coverage properties of our estimating procedure, n i.i.d. samples are again collected from (4.8) to estimate $(\theta_g, \phi)'$. Using the same simulated data, simultaneous confidence intervals $\mathcal{C}_\alpha^{SI}(z^*)$, uncorrected marginal intervals C_α^{LB}, and simultaneous Bonferonni intervals C_α^{UB} are calculated at the 80% and 90% confidence levels for which we record whether each region contains the true value. The true value of $\theta_g = \mu$ is expressible as the sum of each mixture mean multiplied by the mixture probability yielding $\mu = 5$. To calculate $\phi = (\xi_{.10}, \xi_{.90})$, a numerical optimization technique may be used to integrate $\int_{-\infty}^{y} f_X(x)dx$ over values of y until the desired probability is found. We use the integrate function in `R` and found $\xi_{.10} = .2544116$ and $\xi_{.90} = 11.0143117$ with absolute error less than 2.5e-6. Thus, we have six binary outcomes, one for each region and confidence level combination, which are naturally correlated since we are using the same simulated data. We replicate this sampling scheme 2,000 times to create a Monte Carlo sample of six Bernoulli estimates based on the n samples. We then calculate simultaneous intervals, using Theorem 2, with overall 95% confidence level and plot the results in the top row of Figure 4.4. This procedure was repeated for $n = 500$, 1000, 5000, and 10000.

Within each plotting window, Figure 4.4 shows observed coverage probabilities for the uncorrected marginal intervals C_α^{LB}, simultaneous confidence intervals $\mathcal{C}_\alpha^{SI}(z^*)$, and

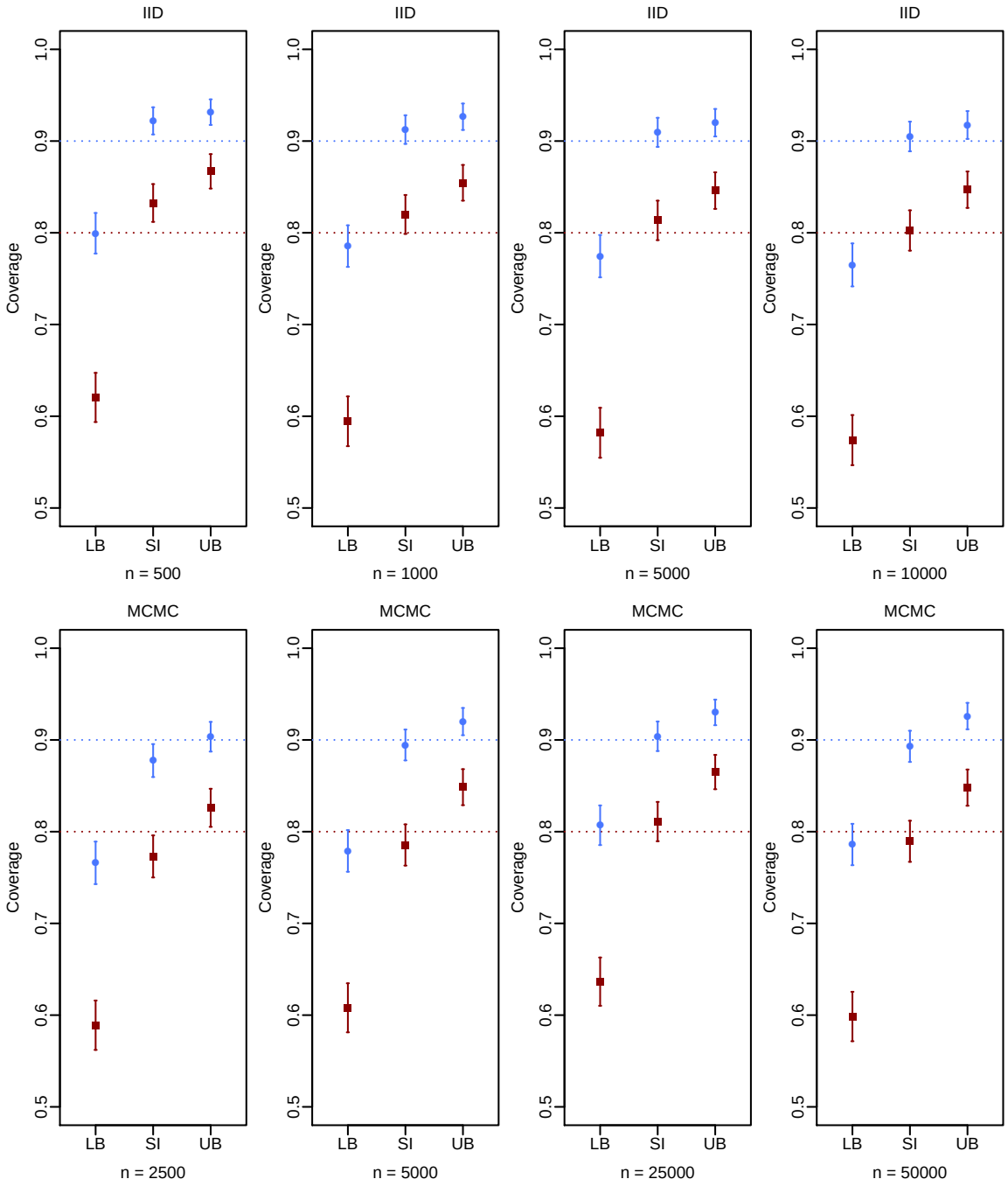

Figure 4.4: Simultaneous 95% confidence intervals for coverage probabilities based on 2,000 replications comparing uncorrected marginal intervals C_α^{LB}, simultaneous confidence intervals $\mathcal{C}_\alpha^{SI}(z^*)$, and simultaneous Bonferonni intervals C_α^{UB}. Blue intervals with a circle and dark red intervals with a square correspond to .9 and .8 nominal levels, respectively.

simultaneous Bonferonni intervals C_α^{UB} from left to right. Blue intervals around a circle and drak red intervals around a square correspond to .9 and .8 nominal levels, respectively. Figure 4.4 also contains dashed lines for the .9 and .8 target nominal levels. Clearly, C_α^{LB} yields significant undercoverage while failing to ever capture the nominal coverage probability within any of its interval estimates. For $C_\alpha^{SI}(z^*)$, the confidence intervals contain the nominal level as the sample size increases illustrating simultaneous intervals yield coverage close to the nominal level. Bonferonni intervals, C_α^{UB}, approach a value which overestimates the nominal level. This overcoverage is relatively small since the adjustment is based on a small number of quantities. However, estimation procedures of higher dimensionality will correspond to more conservative estimates for the upper bound. Usually in a Monte Carlo simulation such as this, only point estimates would be provided in a table. Then the difference between simultaneous and Bonferonni intervals would be difficult to observe and virtually impossible to argue its significance.

Now consider the dependent sampling case using our random walk MH sampler with proposal distribution $N(0, 9)$. All simulation settings remain the same except sample size which we take to be five times larger based on our previous discussion on ESS. That is, we consider $n = 2500$, 5000, 25000, and 50000. The bottom row of Figure 4.4 provides results from the MCMC simulation, which are consistent with the i.i.d. results.

4.4.3 Side-by-side boxplots

Performance of statistical methodologies is often illustrated by loss function comparisons with existing methods over repeated simulations. Visualization of such Monte Carlo studies is often done using side-by-side boxplots. Our visualization tools can be used

to illustrate the variability in the estimation of the quantiles in the boxplots. More importantly, the visualization provides a tool to assess whether sufficient replications have been used.

Consider a comparison of lasso (Tibshirani, 1996), ridge (Hoerl and Kennard, 1970), and ordinary least squares (OLS) regressions. Let $y \in \mathbb{R}^{100}$ be the observed response vector, X be a 100×21 dimensional matrix of covariates, and $\beta^* \in \mathbb{R}^{21}$ be the true regression coefficient vector. For $\epsilon \sim N_{100}(0, I_{100})$, our data generating model is

$$y = X\beta^* + \epsilon\,.$$

We set β^* to be such that the first 11 elements are zero, and the last 10 are random draws from a normal distribution with mean 0 and variance 2. The matrix X is constructed such that the first column is all 1s, and the rows of X_{-1}, the matrix X with the first column removed, are drawn from $N_{20}(0, \Omega)$, where the ijth entry of Ω is $.90^{|i-j|}$. Over repeated simulations, we fit a lasso, ridge, and OLS regressions to estimate the vector of coefficients. Lasso and ridge estimates are obtained using the `glmnet` package (Friedman et al., 2009) with tuning parameters chosen using cross-validation. In each replication, we note the squared estimation error of the estimated coefficient, $\hat{\beta}$, that is, $\|\hat{\beta} - \beta^*\|^2$. We repeat the simulation for 100, 500, and 2000 Monte Carlo replications. Figure 4.5 presents the resulting boxplots with and without simultaneous confidence intervals.

Recall a box in the boxplot has 25%, 50%, and 75% quantiles. To construct simultaneous confidence intervals, we appeal to the 9-dimensional joint asymptotic distribution for i.i.d. sequences for these quantiles and line search algorithm of Section 4.3. The simulta-

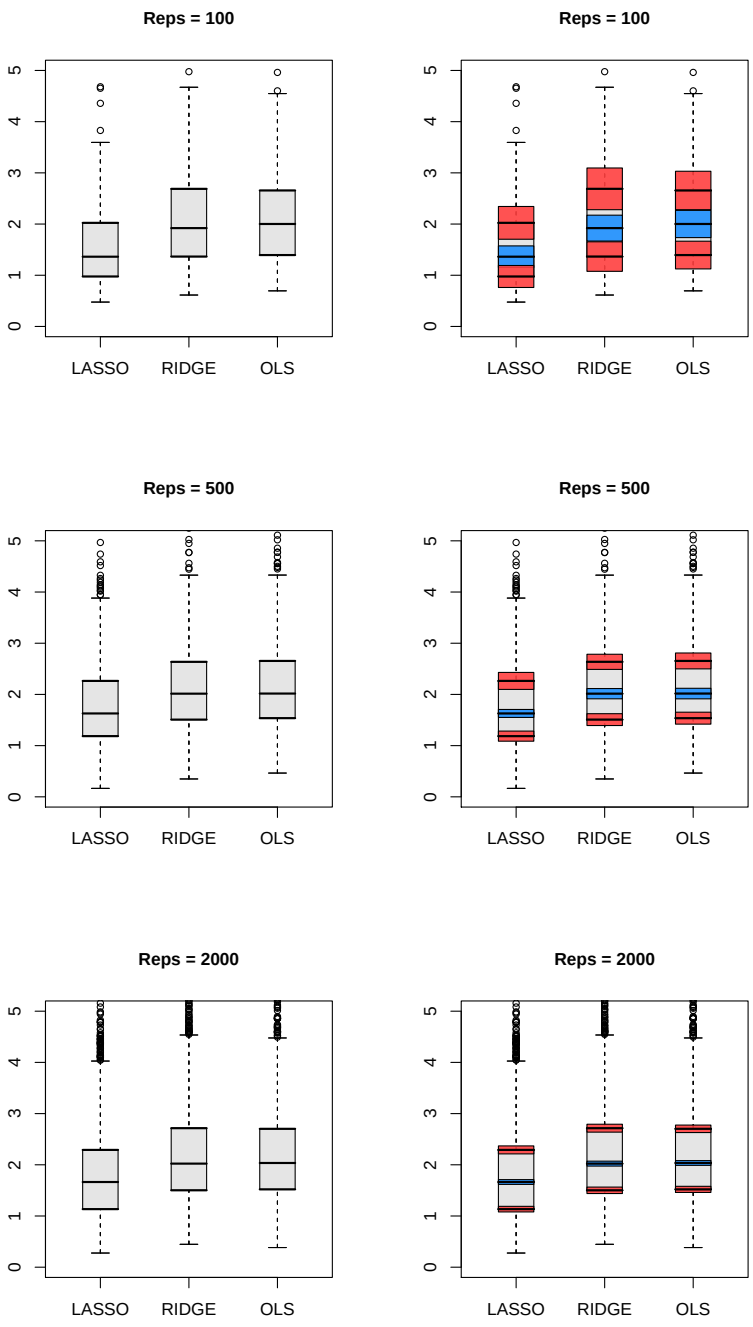

Figure 4.5: Boxplots of squared estimation error for lasso, ridge, and OLS with and without simultaneous confidence intervals. Monte Carlo sample size increases from left to right.

neous confidence intervals immediately indicate that with only 100 Monte Carlo replications, all quantile estimates have large variation. This variability is significantly improved with 2000 Monte Carlo replications. Such an analysis is impossible with the upper plots.

4.4.4 Visualizations for Bayesian analysis

Our final example integrates simultaneous confidence intervals within standard density plots and boxplots for Bayesian analysis. The specific example we consider is a Gibbs sampler targeting the posterior distribution for a hierarchical normal model, analyzing the school data of Gelman et al. (2004).

Consider, for $j = 1, \ldots, J$, the hierarchical model

$$Y_j \mid \theta_j \sim N(\theta_j, \sigma_j^2)$$

$$\theta_j \sim N(\mu, \tau^2)\,,$$

with σ_j^2 known and priors $f(\mu) \propto 1$ and $f(\tau) \propto 1/\tau$. The school data are comprised of estimated effects on student performance on verbal SAT scores after undergoing a coaching program. There are 8 schools in the sample for which each y_j is an estimated effect and σ_j is a known standard error for school j. We are interested in estimating features of the posterior distributions of the θ_j's, the coaching effect for each school. To estimate this we simulate draws from the joint posterior $\theta, \mu, \tau | y$ with a deterministic scan Gibbs sampler

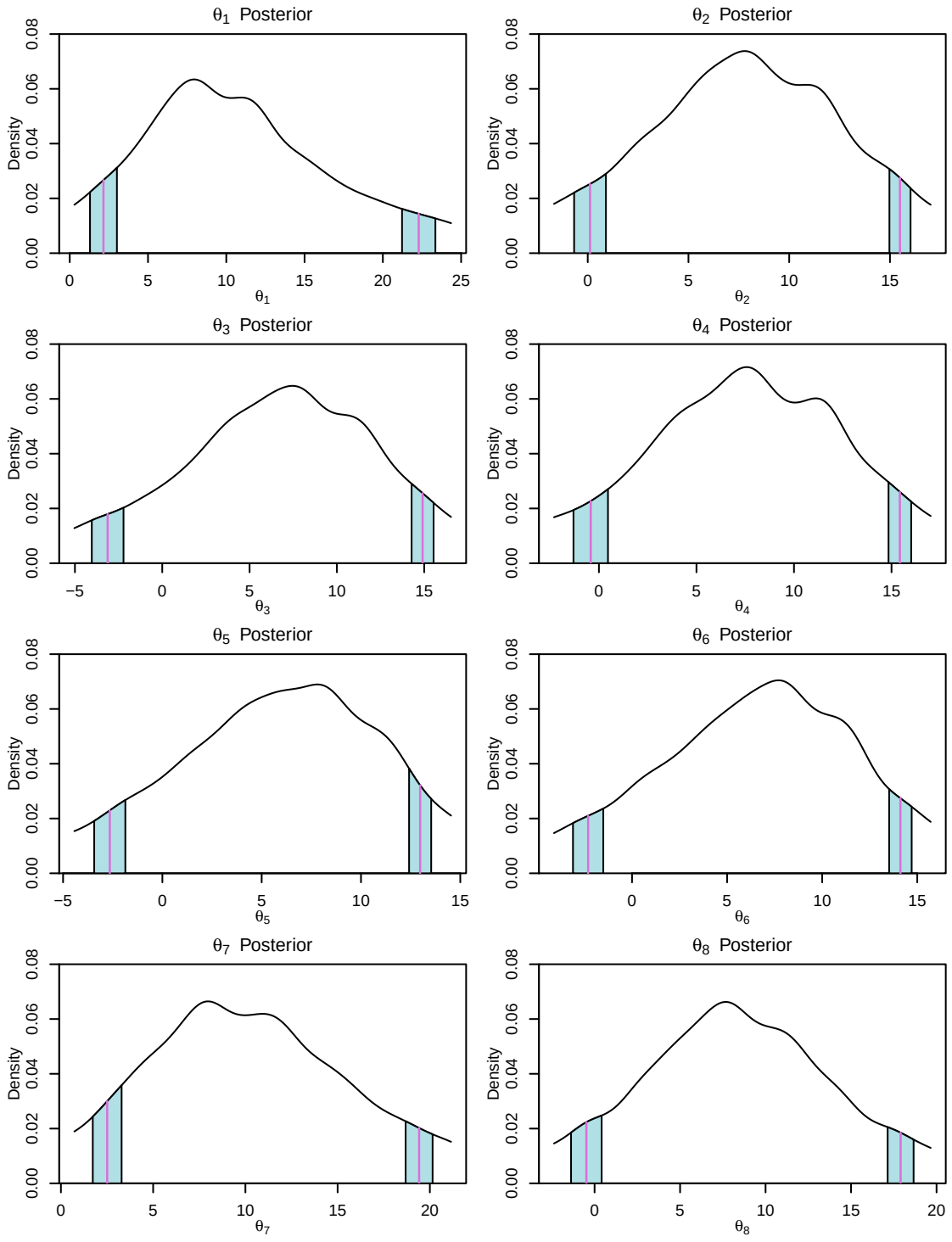

Figure 4.6: Plot of the estimates of an 80% credible interval for each θ with simultaneous 90% confidence intervals for 10,000 samples.

using the following full conditional distributions

$$\theta_j|\mu,\tau,y \sim N\left(\frac{y_j\tau^2+\mu\sigma_j^2}{\tau^2+\sigma_j^2}, \frac{1}{\frac{1}{\sigma_j^2}+\frac{1}{\tau^2}}\right), \quad \mu|\theta,\tau,y \sim N\left(\bar{\theta}, \frac{\tau^2}{J}\right), \text{ and}$$

$$\tau^2|\theta,\mu,y \sim \text{Inv}-\chi^2\left(J-1, \frac{1}{J-1}\sum_{j=1}^{J}(\theta_j-\mu)^2\right),$$

where $\bar{\theta}$ is the average of the θ_j's.

One might only be interested in credible intervals for each parameter. In Figure 4.6 we plot the simultaneous intervals in the form of marginal plots from 10,000 samples, while recalling that the simulation uncertainty estimates incorporate the full covariance structure. Many of the plots have similar shapes but different scales, thus some care should be taken in interpreting the length of each error region in each plot. Figure 4.7 presents the same analysis from a sample of 100,000, for which the simulation uncertainty of each quantity is substantially smaller. Figures 4.6 and 4.7 illustrate one reason accounting for simulation uncertainty can be important. Consider the left endpoints of the credible regions in the plots for θ_2, θ_4, and θ_8. Notice that in Figure 4.6 the left endpoints are indistinguishable from zero when we account for the Monte Carlo error, but this is no longer an issue with Figure 4.7.

Our procedure allows us to display something akin to the output of a `Stan` plot. Figure 4.8 includes a mean estimate and both an 80% and 95% credible interval for each θ in a boxplot inspired design. To make this plot, we estimated the resulting 40-dimensional $(\theta_g, \phi)'$ vector and covariance matrix. This approach makes it perhaps easier to compare the size of error regions around each estimate but discards the information gained by examining

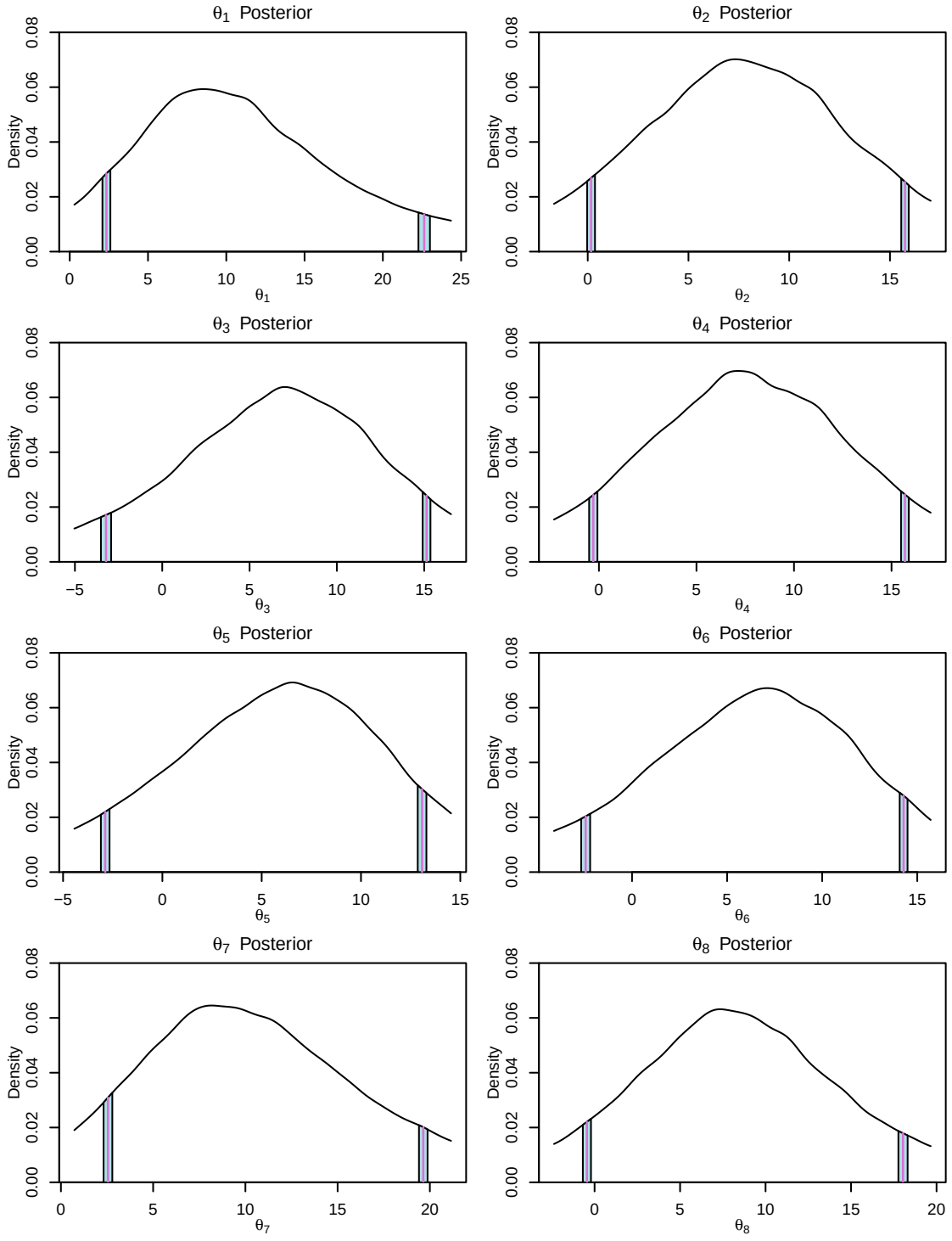

Figure 4.7: Plot of the estimates of an 80% credible interval for each θ with simultaneous 90% confidence intervals for 100,000 samples.

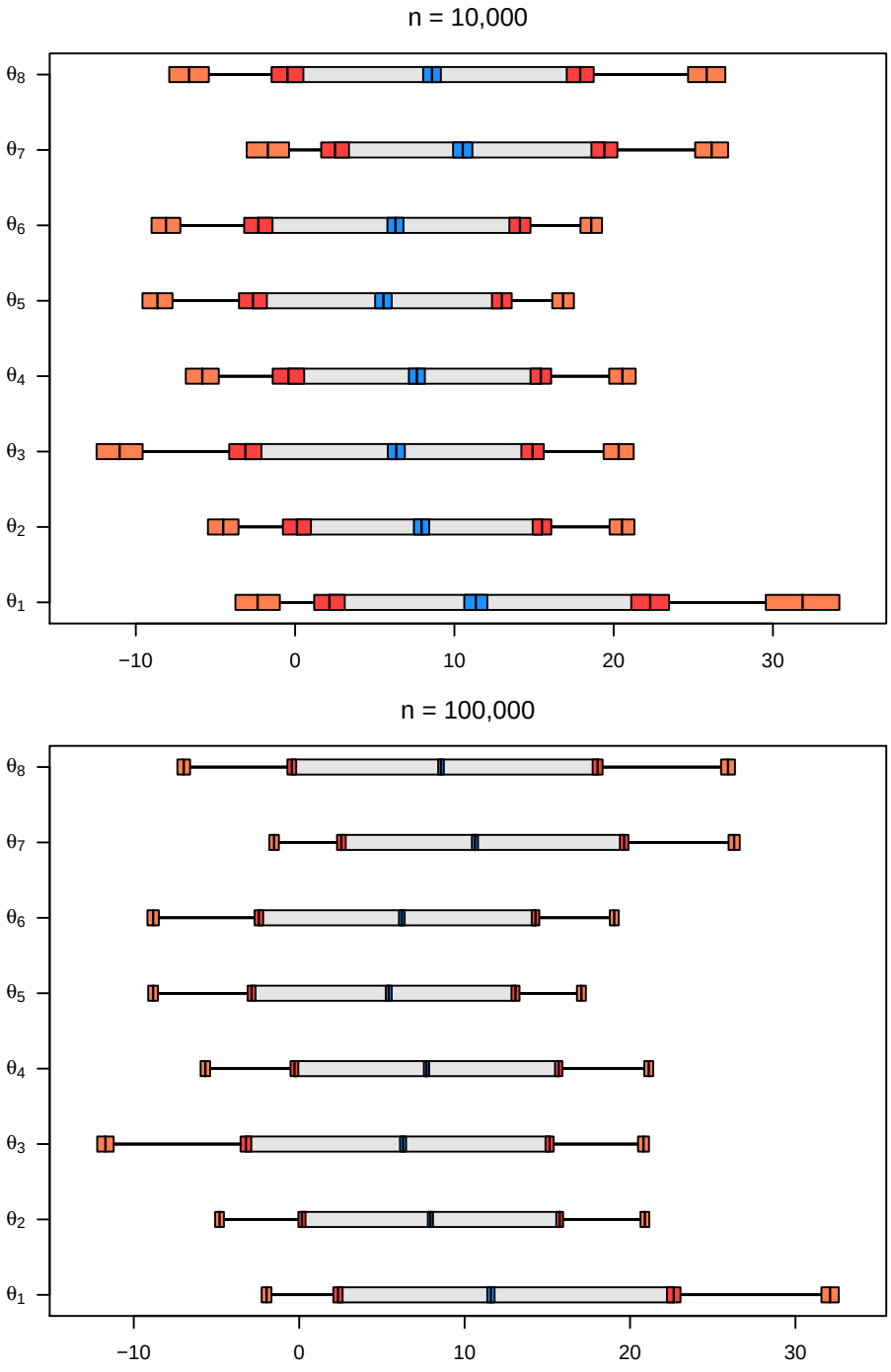

Figure 4.8: Boxplot inspired design where blue, red, and orange boxes correspond to a simultaneous 90% confidence level uncertainty of the posterior mean, 80% and 95% credible intervals, respectively.

the marginal densities. Additionally, this sort of visualization requires each θ to be on a similar scale. Once again the sample size determines the size of the confidence regions, with $n = 100,000$ having substantially less simulation uncertainty.

4.5 Discussion

We provide a novel flexible class of visualizations for assessing the quality of estimation for Monte Carlo sampling. These visualizations are particularly helpful addressing concerns of Monte Carlo sample sizes and may be applied to a wide variety of problems including Monte Carlo estimation of expectations and quantiles, simulation studies, and visualization of a Bayesian analysis. The marginal-friendly interpretation retains more information than previously available methods. The line search algorithm that yields the $1-\alpha$ simultaneous confidence intervals can be more widely applied to any statistic with an approximately multivariate normal sampling distribution (e.g. maximum likelihood estimators).

One issue not addressed is the case of a large number of quantities of interest, particularly in MCMC sampling. In these cases, downward bias exhibited by batch means may lead to noticeable undercoverage of the simultaneous intervals. Other variance estimators such as weighted batch means or a lugsail window function Vats and Flegal (2018) may be used to induce upward bias to combat undercoverage. Additionally, our methods are only as good as the sampling method allows. A poor sampler may not provide representative samples from the target yielding misleading results or could require an enormous number

of samples to be meaningful. We have offered no sampling guidance, but note this is a fundamental challenge in MCMC simulations (see e.g. Brooks et al., 2010; Fishman, 1996).

Chapter 5

Sequential Stopping Rule for Joint Expectations and Quantiles

The previous chapter developed a joint limiting distribution for combinations of expectations and quantiles of Monte Carlo simulations and developed a class of plots to visualize the associated uncertainty. We concluded that it is desirable to have a sample size for which the uncertainty regions are small enough that they become difficult to distinguish from the estimate. We formalize this procedure through the creation of a sequential stopping rule that handles more general quantities than those of Glynn and Whitt (1992), Jones et al. (2006), and Vats et al. (2019).

5.1 Sequential Stopping Rule

We consider using Theorem 2 to create a new sequential stopping rule for any confidence region satisfying $\sqrt{n}(\text{Vol}(C_\alpha(n)))^{1/p} \xrightarrow{w.p.1} c > 0$, where c is a constant. This

result largely mirrors Vats et al. (2019) but places the assumptions on quantities of interest and the form of the confidence region rather than the underlying Markov chain. Once the general results is available we may consider the restrictions on the Markov chains and forms of the confidence region which satisfy our assumptions.

Consider the stopping rule

$$\tau(\epsilon) = \inf\{n > 0 : \text{Vol}(C_\alpha(n))^{1/p} + s(n) < \epsilon \hat{K}_n\}, \tag{5.1}$$

where $\hat{K}_n$ is an estimator of an estimation metric K and $s(n)$ is a positive decreasing function of n such that $s(n) = o(n^{-1/2})$. Common choices for K_n include 1, the magnitude of the estimate, and the determinant of the posterior covariance matrix. These metrics yeild fixed-volume, relative magnitude, and relative covariance stopping rules as discussed in Vats et al. (2019). To simplify notation, let $\hat{\Theta}_n = \begin{pmatrix} \bar{g}_n \\ \hat{\phi} \end{pmatrix}$ and $\Theta = \begin{pmatrix} \theta \\ \phi \end{pmatrix}$.

Theorem 3 *Suppose* $\sqrt{n}\hat{\Sigma}_n^{-1/2}\hat{\Lambda}_n(\hat{\Theta}_n - \Theta)$ *satisfies a functional CLT,* $\hat{\Lambda}_n \to \Lambda$ *w.p. 1,* $\hat{\Sigma}_n \to \Sigma$ *w.p. 1,* $\hat{K}_n \to K$ *w.p. 1, and* $\sqrt{n}(\text{Vol}(C_\alpha(n)))^{1/p} \xrightarrow{w.p.1} c > 0$. *Then as* $\epsilon \to 0$, $\tau(\epsilon) \to \infty$ *and* $Pr\{\Theta \in C_\alpha(\tau(\epsilon))\} \to 1 - \alpha$.

Proof. Let

$$V(n) = (\text{Vol}(C_\alpha(n)))^{1/p} + s(n),$$

and simplifying the notation to

$$\tau(\epsilon) = \inf\{n > 0 : V(n) < \epsilon \hat{K}_n\}.$$

Since $V(n) > s(n)$,

$$T(\epsilon) = \inf\{n > 0 : s(n) > \epsilon \hat{K}_n\} < \tau(\epsilon).$$

Since $T(\epsilon) \to \infty$ as $\epsilon \to 0$, $T(\epsilon) < \tau(\epsilon) \to \infty$.

By assumption $\sqrt{n}\text{Vol}(C_\alpha(n))^{1/p} \to c$ and $s(n) = o(n^{-1/2})$ implies $\sqrt{n}s(n) \to 0$. Therefore

$$\sqrt{n}V(n) = \sqrt{n}[\text{Vol}(C_\alpha(n))^{1/p} + s(n)] = \sqrt{n}\text{Vol}(C_\alpha(n))^{1/p} + \sqrt{n}s(n) \xrightarrow{w.p.1} c > 0$$

as $n \to \infty$.

By definition of $\tau(\epsilon)$,

$$V(\tau(\epsilon) - 1) > \epsilon \hat{K}_{\tau(\epsilon)-1} \text{ and } \frac{V(\tau(\epsilon) - 1)}{\hat{K}_{\tau(\epsilon)-1}} > \epsilon.$$

Therefore,

$$\limsup_{\epsilon \to 0} \epsilon[\tau(\epsilon)]^{1/2} \leq \limsup_{\epsilon \to 0} \left(\frac{V(\tau(\epsilon) - 1)}{\hat{K}_{\tau(\epsilon)-1}} \right) [\tau(\epsilon)]^{1/2} = \frac{c}{K}. \tag{5.2}$$

Also by the definition of $\tau(\epsilon)$, there exists $u(\epsilon)$ be a positive random variable on $[0, 1]$ such that

$$V(\tau(\epsilon) + u(\epsilon)) < \epsilon \hat{K}_{\tau(\epsilon)+u(\epsilon)} \text{ and } \frac{V(\tau(\epsilon) + u(\epsilon)}{\hat{K}_{\tau(\epsilon)+u(\epsilon)}} > \epsilon.$$

Therefore,

$$\liminf_{\epsilon \to 0} \epsilon[\tau(\epsilon)]^{1/2} \geq \liminf_{\epsilon \to 0} \left(\frac{V(\tau(\epsilon) + u(\epsilon))}{\hat{K}_{\tau(\epsilon)+u(\epsilon)}} \right) [\tau(\epsilon)]^{1/2} = \frac{c}{K}. \tag{5.3}$$

Then, by (5.2) and (5.3)

$$\lim_{\epsilon \to 0} \epsilon[\tau(\epsilon)]^{1/2} = \frac{c}{K}. \tag{5.4}$$

We may now apply a standard random time change argument as discussed in Billingsley (1968), yielding

$$\sqrt{\tau(\epsilon)}\hat{\Sigma}_{\tau(\epsilon)}^{-1/2}\hat{\Lambda}_{\tau(\epsilon)}(\hat{\Theta}_{\tau(\epsilon)} - \Theta) \to N_p(0, I_p).$$

Then

$$Pr\{\Theta \in C_\alpha(\tau(\epsilon))\} \xrightarrow{w.p.1} 1 - \alpha$$

as $\epsilon \to 0$. ■

We now address conditions on the Markov chain $\{X_j\}$, for these assumptions.

5.1.1 FCLT

We build upon Theorem 2,

$$\sqrt{n}(\hat{\Theta}_n - \Theta) \to N(0, \Lambda^{-1}\Sigma\Lambda^{-1}),$$

which implies

$$\sqrt{n}\Sigma^{-1/2}\Lambda(\hat{\Theta}_n - \Theta) \to N_p(0, I_p).$$

When $\hat{\Lambda}_n \to \Lambda$ w.p.1 and $\hat{\Sigma}_n \to \Sigma$ w.p.1, then $\hat{\Sigma}_n^{-1/2}\hat{\Lambda}_n \to \Sigma^{-1/2}\Lambda$ w.p.1 and

$$\hat{\Sigma}_n^{-1/2}\hat{\Lambda}_n(\Sigma^{-1/2}\Lambda)^{-1} \xrightarrow{w.p.1} I_p. \tag{5.5}$$

Then,

$$\hat{\Sigma}_n^{-1/2}\hat{\Lambda}_n(\Sigma^{-1/2}\Lambda)^{-1}\sqrt{n}\Sigma^{-1/2}\Lambda(\hat{\Theta}_n-\Theta)=\sqrt{n}\hat{\Sigma}_n^{-1/2}\hat{\Lambda}_n(\hat{\Theta}_n-\Theta)\to I_pN_p(0,I_p).$$

Finally, we find the normal limiting distribution stated in terms of our estimators,

$$\sqrt{n}\hat{\Sigma}_n^{-1/2}\hat{\Lambda}_n(\hat{\Theta}_n-\Theta)\to N_p(0,I_p). \tag{5.6}$$

The conditions to establish a Markov chain CLT are often sufficient to establish a FCLT. In particular, Vats et al. (2018) establish a strong invariance principle under the conditions $\{X_j\}$ is polynomial ergodic of order $k>(1+\epsilon_1)(1+2/\delta)$ and $E_F||g||^{2+\delta}<\infty$. This strong invariance principle implies the existence of a FCLT. These assumptions and the techniques of Doss et al. (2014) appear enough to satisfy a FCLT for (5.6) however a formal proof is left for future work.

Thus we establish our FCLT with the conditions of Theorem 2 and strongly consistent estimators. This requires the assumptions for Bahadur's representation of a quantile, i.e. $\{X_j\}$ is polynomial ergodic, there exists a $\delta>0$ such that $E||g||^{2+\delta}<\infty$, and for each $i=p_1+1,\ldots,p_1+p_2$ F_{h_i} is absolutely continuous with continuous density f_{h_i} such that $0<f_{h_i}(\xi_{q_i})$.

5.1.2 Strongly Consistent Estimator of Λ

Recall

$$\Lambda=\begin{pmatrix} I_{d_1} & 0 \\ 0 & A \end{pmatrix},$$

where

$$[A]_{ij} = \begin{cases} f_{h_i}(\xi_{q_i}) & i = j \\ 0 & \text{otherwise} \end{cases} \tag{5.7}$$

Thus, estimating Λ reduces to estimating the densities f_{h_i} at a single point. Masry and Györfi (1987) provide a strongly consistent estimator for densities of stationary processes that are asymptotically uncorrelated. This estimation deserves its own section and will be discussed in Section 5.2. We will establish that a geometrically ergodic Markov chain satisfying detailed balance is a sufficient condition on the Markov chain to estimate the associated densities with probability one.

5.1.3 Strongly Consistent Estimators for Σ and K

These quantities are associated with the estimation of only expectations and thus the restrictions on the Markov chain are identical to those for the case of expectations. We restrict our discussion to the batch means estimator for Σ and turn to Theorem 2 from Vats et al. (2019) for which we need the following conditions.

Condition 4 *Let $||\cdot||$ be the Euclidean norm and $\{B(t), t \geq 0\}$ be a $p-$dimensional multivariate Brownian motion. There exists a $p \times p$ lower triangular matrix L, a non-negative increasing function γ on the positive integers, a finite random variable D, and a sufficiently rich probability space such that, with probability 1,*

$$||n(\hat{\theta}_n - \theta) - LB(n)|| < D\gamma(n) \text{ as } n \to \infty. \tag{5.8}$$

Condition 5 *The batch size b_n satisfies the following conditions,*

1. *the batch size* b_n *is an integer sequence such that* $b_n \to \infty$ *and* $n/b_n \to \infty$ *as* $n \to \infty$ *where,* b_n *and* n/b_n *are monotonically increasing,*

2. *there exists a constant* $c \geq 1$ *such that* $\sum_n (b_n n^{-1})^c < \infty$.

Theorem 6 *Let* g *be such that* $E_F||g||^{\delta+2} < \infty$ *for some* $\delta > 0$. *Let* X *be an* $F-$*invariant polynomial ergodic Markov chain of order* $k > (1+\epsilon_1)(1+2/\delta)$ *for some* $\epsilon_1 > 0$. *Then* (5.8) *holds with* $\gamma(n) = n^{1/2-\lambda}$ *for some* $\lambda > 0$. *If Condition 5 holds and* $b_n^{-1/2}(\log n)^{1/2-\lambda} \to 0$ *as* $n \to \infty$, *then* $\hat{\Sigma}_n \to \Sigma$, *with probability 1, as* $n \to \infty$.

We satisfy the requirements for Theorem 6 with $b_n = \lfloor \sqrt{n} \rfloor$ and require $\{X_j\}$ be polynomial ergodic of order at least $m = (1+\epsilon_1)(1+2/\delta)$ and $E|g|^{2+\delta} < \infty$.

5.1.4 Convergence of Confidence Region Volume

We consider two forms of confidence regions.

$$C_\alpha^E(n) = \left\{ \hat{\Theta}_n : n(\hat{\Theta}_n - \Theta)^T \hat{\Lambda}_n \hat{\Sigma}_n^{-1} \hat{\Lambda}_n (\hat{\Theta}_n - \Theta) \leq \chi^2_{1-\alpha,p} \right\}, \tag{5.9}$$

and

$$C_\alpha^{SI}(n) = \bigcap_{i=1}^{p} \left\{ \hat{\Theta}_n : \sqrt{n} \left([\hat{\Theta}_n]_i - [\Theta]_i \right) \leq z_\alpha^* \frac{\sqrt{[\hat{\Sigma}]_{ii}}}{[\hat{\lambda}]_{ii}} \right\}. \tag{5.10}$$

$C_\alpha^E(n)$ is the classic ellipsoidal confidence region for a multivariate normal distribution while $C_\alpha^{SI}(n)$ is the hyperrectangular confidence region discussed in Section 4.3. We first prove convergence of the volume of $C_\alpha^E(n)$.

Theorem 7 *Assume there exists a CLT of the form* (4.5). *Then as* $n \to \infty$ $\sqrt{n}\, Vol(C_\alpha^E(n))$ *converges to a positive finite constant.*

Proof. The volume of the ellipsoidal confidence region is

$$\mathrm{Vol}(C_\alpha^E(n)) = \frac{2\pi^{p/2}}{p\Gamma(p/2)} \left(\frac{\chi^2_{1-\alpha,p}}{n} \right)^{p/2} |\hat{\Lambda}_n^{-1} \hat{\Sigma}_n \hat{\Lambda}_n^{-1}|^{1/2}. \tag{5.11}$$

Since

$$|\hat{\Lambda}_n^{-1} \hat{\Sigma}_n \hat{\Lambda}_n^{-1}|^{1/2} \xrightarrow{w.p.1} |\Lambda^{-1} \Sigma \Lambda^{-1}|^{1/2},$$

and all the other terms are constant,

$$\mathrm{Vol}(C_\alpha^E(n)) \xrightarrow{w.p.1} n^{-p/2} \frac{2\pi^{p/2}}{p\Gamma(p/2)} \left(\chi^2_{1-\alpha,p}\right)^{p/2} |\Lambda^{-1} \Sigma \Lambda^{-1}|^{1/2}.$$

Thus

$$\begin{aligned} \sqrt{n} \left\{ \mathrm{Vol}(C_\alpha^E(n)) \right\}^{1/p} &\xrightarrow{w.p.1} \sqrt{n} \left\{ n^{-p/2} \frac{2\pi^{p/2}}{p\Gamma(p/2)} \left(\chi^2_{1-\alpha,p}\right)^{p/2} |\Lambda^{-1} \Sigma \Lambda^{-1}|^{1/2} \right\}^{1/p} \\ &= \left\{ \frac{2\pi^{p/2}}{p\Gamma(p/2)} \left(\chi^2_{1-\alpha,p}\right)^{p/2} |\Lambda^{-1} \Sigma \Lambda^{-1}|^{1/2} \right\}^{1/p} = c^{(E)} > 0. \end{aligned}$$

■

The result for $C_\alpha^{SI}(n)$ is similar.

Theorem 8 *Assume there exists a CLT of the form* (4.5)*. Then as* $n \to \infty$ $\sqrt{n}\,Vol(C_\alpha^{SI}(n))$ *converges to a positive finite constant.*

Proof. The simultaneous intervals confidence region has volume

$$\begin{aligned}
\text{Vol}(C_\alpha^{SI}(n)) &= \prod_{i=1}^{p} \left(\left\{ [\hat{\Theta}_n]_i + z_\alpha^* \frac{\sqrt{[\hat{\Sigma_n}]_{ii}}}{\sqrt{n}[\hat{\Lambda_n}]_{ii}} \right\} - \left\{ [\hat{\Theta}_n]_i + z_\alpha^* \frac{\sqrt{[\hat{\Sigma}_n]_{ii}}}{\sqrt{n}[\hat{\Lambda}_n]_{ii}} \right\} \right) \\
&= \prod_{i=1}^{p} \left(2z_\alpha^* \frac{\sqrt{[\hat{\Sigma_n}]_{ii}}}{\sqrt{n}[\hat{\Lambda_n}]_{ii}} \right) \\
&= (2z_\alpha^*)^p n^{-p/2} \prod_{i=1}^{p} \left(\frac{\sqrt{[\hat{\Sigma}_n]_{ii}}}{[\hat{\Lambda}_n]_{ii}} \right).
\end{aligned}$$

Since

$$\frac{\sqrt{[\hat{\Sigma}_n]_{ii}}}{[\hat{\Lambda}_n]_{ii}} \xrightarrow{w.p.1} \frac{\sqrt{[\Sigma]_{ii}}}{[\Lambda]_{ii}},$$

for each $i = 1, \ldots, p$, and all other terms are constant

$$\text{Vol}(C_\alpha^{SI}(n)) \xrightarrow{w.p.1} n^{-p/2}(2z_\alpha^*)^p \prod_{i=1}^{p} \left(\frac{\sqrt{[\Sigma]_{ii}}}{[\Lambda]_{ii}} \right).$$

Thus,

$$\begin{aligned}
\sqrt{n} \left\{ \text{Vol}(C_\alpha^{SI}(n)) \right\}^{1/p} &\xrightarrow{w.p.1} \sqrt{n} \left\{ n^{-p/2}(2z_\alpha^*)^p \prod_{i=1}^{p} \left(\frac{\sqrt{[\Sigma]_{ii}}}{[\Lambda]_{ii}} \right) \right\}^{1/p} \\
&= 2z_\alpha^* \left\{ \prod_{i=1}^{p} \left(\frac{\sqrt{[\Sigma]_{ii}}}{[\Lambda]_{ii}} \right) \right\}^{1/p} = c^{SI} > 0.
\end{aligned}$$

■

Hence both $C_\alpha^E(n)$ and $C_\alpha^{SI}(n)$ satisfy our condition as $n \to \infty$,

$$\sqrt{n}(\text{Vol}(C_\alpha(n)))^{1/p} \xrightarrow{w.p.1} c > 0.$$

5.2 Density Estimation

Consider the kernel density estimator

$$\hat{f}_n(y) = \frac{1}{n}\sum_{j=1}^{n} \frac{1}{b_j} K\left(\frac{y - Y_j}{b_j}\right), \tag{5.12}$$

with bandwidth $b_n \to 0$ and kernel K which satisfies the following two conditions.

1. $K \in L_1, \int_{\mathbb{R}} K(u)du = 1$

2. K is bounded and $K(y) = O(|y|^{-1-\epsilon})$ for some $\epsilon > 0$

Masry and Györfi (1987) establish conditions on the kernel density and process which yield a strongly consistent estimator for $f(y)$. Assumption 1 is concerned with the sequence $\{Y_j\}$ while Assumption 2 is concerned with the kernel and bandwidth choice.

Assumption 1 *The process $\{Y_j\}$ is asymptotically uncorrelated with*

$$\sum_{n=1}^{\infty} (\log n)(\log_2 n) u_n^2 \rho^2(n) < \infty$$

and $u_n = n^{-v}$ for some $0 \leq v \leq 1/2$.

Assumption 2 *Let K be a kernel and b_n be a bandwidth parameter satisfying*

- $\int_{\mathbb{R}} uK(u)du = 0$
- $\int_{\mathbb{R}} u^2 K(u)du < \infty$
- $\sum_{i=1}^{\infty} b_i^2 = \infty$.

We now present a simplified univariate version of Masry and Györfi's (1987) Theorem 2.3

Theorem 9 *Let $\hat{f}_n(y)$ be defined as in* (5.12).

(a) For almost all y we have

$$\lim_{n\to\infty} E[\hat{f}_n(y)] = f(y).$$

(b) Assume that f is twice differentiable and its second derivative is and bounded and continuous on $\mathbb{R}$. The kernel K is assumed to satisfy Assumption 2. Then

$$\lim_{n\to\infty} \left(\frac{1}{n}\sum_{i=1}^{n} b_i^2\right)^{-1} (E[\hat{f}_n(y)] - f(y)) = \frac{1}{2}\int_{\mathbb{R}} u^2 f''(y) K(u) du.$$

(c) If in addition the process $\{Y_j\}$ satisfies Assumption 1 and b_n is chosen as

$$b_n \sim n^{-(1-2v)/5},$$

then

$$\left\{\frac{n^{4(1-2v)(5)}}{(\log n)(\log_2 n)^{1+\delta}}\right\}^{1/2} (\hat{f}_n(y) - f(y)) \to 0 \tag{5.13}$$

with probability 1 as $n \to \infty$ for every $\delta > 0$.

The conditions placed on the kernel density, $K(\cdot)$, are independent of $\{Y_j\}$, and the bandwidth, b_n, is determined by the value v. Therefore, the only condition placed on the Markov chain is that $\{Y_j\}$ is asymptotically uncorrelated with the specified rate of convergence. Jones (2004) provides us with the following corollary relating asymptotically uncorrelated sequences to ergodic Markov chains.

Corollary 10 *A Harris ergodic Markov chain $\{Y_j\}$ is asymptotically uncorrelated if Y is geometrically ergodic and satisfies detailed balance. Additionally, there exists a $\theta > 0$ such that $\rho(n) = O(e^{-\theta n})$.*

We may then relate Corollary 10 to Assumption 1 through the following Lemma.

Lemma 11 *Let $\{Y_j\}$ be a geometrically ergodic Markov chain satisfying detailed balance. Then*

$$\sum_{j=1}^{\infty}(\log(n))(\log_2(n))u_n^2\rho^2(n) < \infty$$

with $u_n = n^{-v}$, for any $0 \leq v \leq 1/2$.

Proof. We begin by setting $u_n = n^{-v}$ and rearranging terms.

$$\begin{aligned}\sum_{n=1}^{\infty}(\log(n))(\log_2(n))u_n^2\rho^2(n) &= \sum_{n=1}^{\infty}(\log(n))(\log_2(n))(n^{-v})^2\rho^2(n)\\ &= \sum_{n=1}^{\infty}(\log(n))\left(\frac{\log(n)}{\log(2)}\right)n^{-2v}\rho^2(n)\\ &= \frac{1}{\log(2)}\sum_{n=1}^{\infty}(\log(n))^2n^{-2v}\rho^2(n)\end{aligned}$$

Since $\rho(n) = O(e^{-\theta n})$, there exist an $l > 0$ and n^* such that for $n \geq n^*$,

$$|\rho(n)| \leq le^{-\theta n}.$$

Hence,

$$\rho^2(n) \leq (le^{-\theta n})^2 = l^2e^{-(2\theta)n}.$$

We let $M = l^2$ and define a new $\theta := 2\theta$ yielding

$$\rho^2(n) \leq Me^{-\theta n}. \tag{5.14}$$

We split the sum into two parts between $n^\star - 1$ and $n^\star$. Then we substitute the upper bound (5.14) for the sum containing $n^\star$.

$$\begin{aligned}
&\frac{1}{\log(2)} \sum_{n=1}^{\infty} (\log(n))^2 n^{-2v} \rho^2(n) \\
&\qquad = \frac{1}{\log(2)} \sum_{n=1}^{n^*-1} (\log(n))^2 n^{-2v} \rho^2(n) + \frac{1}{\log(2)} \sum_{n=n^*}^{\infty} (\log(n))^2 n^{-2v} \rho^2(n) \\
&\qquad \leq \frac{1}{\log(2)} \sum_{n=1}^{n^*-1} (\log(n))^2 n^{-2v} \rho^2(n) + \frac{1}{\log(2)} \sum_{n=n^*}^{\infty} (\log(n))^2 n^{-2v} M e^{-\theta n}.
\end{aligned}$$

Notice $0 \leq \rho^2(n) \leq 1$, thus $(\log(n))^2 n^{-2v} \rho^2(n)$ is finite for all n. Hence,

$$\frac{1}{\log(2)} \sum_{n=1}^{n^*-1} (\log(n))^2 n^{-2v} \rho^2(n) = C_{n^*} < \infty.$$

Additionally, note that $(\log(n))^2 n^{-2v} M e^{-\theta n}$ is positive and monotone decreasing for n such that $\log(n)\{\frac{2v}{n} + \theta\} > \frac{2}{n}$, since

$$\begin{aligned}
&\frac{d}{dn}\left([\log(n)]^2 n^{-2v} M e^{-\theta n}\right) \\
&\quad = \frac{2\log(n)}{n} n^{-2v} M e^{-\theta n} + [\log(n)]^2 \frac{d}{dn}\left[n^{-2v} M e^{-\theta n}\right] \\
&\quad = \frac{2\log(n)}{n} n^{-2v} M e^{-\theta n} + [\log(n)]^2 M\left[-2v n^{-2v-1} e^{-\theta n} + n^{-2v} e^{-\theta n}(-\theta)\right] \\
&\quad = \frac{2\log(n)}{n} n^{-2v} M e^{-\theta n} + [\log(n)]^2 M(-n^{-2v} e^{-\theta n})[2\frac{v}{n} + \theta] \\
&\quad = \log(n) n^{-2v} M e^{-\theta n}\left[\frac{2}{n} - \log(n)\left\{2\frac{v}{n} + \theta\right\}\right].
\end{aligned}$$

We choose n^* to be large enough to satisfy $\log(n)\{\frac{2v}{n} + \theta\} > \frac{2}{n}$. Then

$$\begin{aligned}
&\frac{1}{\log(2)} \sum_{n=1}^{n^*-1} (\log(n))^2 n^{-2v} \rho^2(n) + \frac{1}{\log(2)} \sum_{n=n^*}^{\infty} (\log(n))^2 n^{-2v} M e^{-\theta n} \\
&\qquad\qquad = C_{n^*} + \frac{M}{\log(2)} \sum_{n=n^*}^{\infty} (\log(n))^2 n^{-2v} e^{-\theta n},
\end{aligned}$$

and

$$\sum_{n=n^*}^{\infty} (\log(n))^2 n^{-2v} e^{-\theta n} < \infty \iff \int_{n^*}^{\infty} (\log(n))^2 n^{-2v} e^{-\theta n} dn < \infty.$$

To evaluate $\int_{n^*}^{\infty} (\log(n))^2 n^{-2v} e^{-\theta n} dn$ we note n^{-2v} increases as v decreases to 0. Thus,

$$\int_{n^*}^{\infty} (\log(n))^2 n^{-2v} e^{-\theta n} dn < \int_{n^*}^{\infty} (\log(n))^2 n^{-2(0)} e^{-\theta n} dn = \int_{n^*}^{\infty} (\log(n))^2 e^{-\theta n} dn. \qquad (5.15)$$

Additionally, $n^2 > [\log(n)]^2$ for all n, thus

$$\int_{n^*}^{\infty} (\log(n))^2 e^{-\theta n} dn < \int_{n^*}^{\infty} (n)^2 e^{-\theta n} dn.$$

We can now solve this integral using the property $\int udv = uv - \int vdu$.

$$\begin{aligned}
\int_{n^*}^{\infty} n^2 e^{-\theta n} dn &= \left[-\frac{n^2}{\theta} e^{-\theta n} \right]_{n^*}^{\infty} + \frac{2}{\theta} \int_{n^*}^{\infty} n e^{-\theta n} dn \\
&= \left[-\frac{n^2}{\theta} e^{-\theta n} \right]_{n^*}^{\infty} + \frac{2}{\theta} \left\{ \left[-\frac{n}{\theta} e^{-\theta n} \right]_{n^*}^{\infty} + \frac{1}{\theta} \int_{n^*}^{\infty} e^{-\theta n} dn \right\} \\
&= -\frac{1}{\theta} \left[n^2 e^{-\theta n} \right]_{n^*}^{\infty} - \frac{2}{\theta^2} \left[n e^{-\theta n} \right]_{n^*}^{\infty} - \frac{2}{\theta^3} \left[e^{-\theta n} \right]_{n^*}^{\infty} \\
&= -\frac{1}{\theta} \left[0 - (n^*)^2 e^{-\theta n^*} \right] - \frac{2}{\theta^2} \left[0 - n^* e^{-\theta n^*} \right] - \frac{2}{\theta^3} \left[0 - e^{-\theta n^*} \right] \\
&= \frac{(n^*)^2 e^{-\theta n^*}}{\theta} + \frac{2 n^* e^{-\theta n^*}}{\theta^2} + \frac{2 e^{-\theta n^*}}{\theta^3} < \infty.
\end{aligned}$$

Therefore,

$$\rho(n) = O(e^{-\theta n}) \Longrightarrow \sum_{j=1}^{\infty} (\log(n))(\log_2(n)) u_n^2 \rho^2(n) < \infty. \tag{5.16}$$

Notice in (5.15) choosing $v = 0$ provides that any $v > 0$ will hold. Thus we may choose any $0 \leq v \leq 1/2$. ■

We summarize the main result of this section with the following corollary to Theorem 9.

Corollary 12 *Let $\{Y_j\}$ be a geometrically ergodic Markov chain satisfying detailed balance, $\hat{f}_n(y)$ be defined as in (5.12), the bandwidth $b_n = n^{-(1-2v)/5}$ for some $0 \leq v \leq 1/2$, and the kernel K satisfy Assumption 2. Then $\hat{f}_n(y) \to f(y)$ with probability 1 as $n \to \infty$.*

5.3 MCMC Sequential Stopping Rule

We may now state a more specific version of Theorem 3 with the assumptions on the Markov chain for confidence regions (5.9) and (5.10).

Theorem 13 *Let* $\{Xn\}$ *be a geometric ergodic Markov chain satisfying detailed balance, there exists a* $\delta > 0$ *such that* $E||g||^{2+\delta} < \infty$*, for each* $i = 1, \ldots, p_2$ F_{h_i} *is absolutely continuous with continuous density* f_{h_i} *such that* $0 < f_{h_i}(\xi_{q_i})$*,* $\hat{\Sigma}_n$ *is a batch means estimator with batch size* $\lfloor \sqrt{n} \rfloor$*,* $\hat{f}$ *is a kernel density estimator with bandwidth* $\sim n^{-1/5}$ *and a Gaussian kernel,* $\hat{K}_n \to K$ *w.p.1., and stopping rule* $\tau_\epsilon(C_\alpha^E(n))$ *or* $\tau_\epsilon(C_\alpha^{SI}(n))$*. Then as* $\epsilon \to, 0$ $\tau_\epsilon \to 0$ *w.p.1 and* $Pr\{\Theta \in C_\alpha\} = 1 - \alpha$*.*

The proof of this theorem follows from Theorem 3, a strong invariance principle for polynomial ergodic Markov chains, Theorem 6, Theorem 7, Theorem 8, and Corollary 12.

5.4 Simulation Study

To test the stopping rule in Theorem 13, we preform repeated simulation from a probability distribution of interest. We conduct replications of an MCMC sampler which terminates based on our stopping rule and calculate the proportion of replicates whose confidence region contains the truth. This forms a Monte Carlo sample from which we estimate the coverage probabilities for each confidence region. We consider the ellipsoidal and simultaneous interval confidence regions as well as hyper-rectangular regions based on the lower and upper bound intervals. The lower and upper bound confidence regions also satisfy the criteria for Theorem 3, albeit for an unknown confidence level. Each simulation will be conducted with various choices of ϵ, confidence level, and relative metric. The first

example returns to our mixture of normals described in Section 4.4.1. The second example considers a logistic regression on the Anguilla eel data included in the *R* package *dismo* (Hijmans et al., 2017).

5.4.1 Mixture of Normals

We consider a mixture of three normal distribution for which we estimate the mean and a 90% credible interval which takes the form of the .1 and .9 quantiles. We conduct 2,000 replications for each combination of relative metric, ϵ, and confidence level. Taking only three quantities of interest, we expect the upper bound confidence region to provide a conservative coverage estimate. The amount of overcoverage may be rather small due to the low dimension of the problem. Similarly, we expect the lower bound confidence regions to provide undercoverage that may be less exaggerated than examples of higher dimension.

		$\epsilon = .1$			$\epsilon = .05$			$\epsilon = .02$		
Conf		C_α	$\bar{n}$	σ_n	C_α	$\bar{n}$	σ_n	C_α	$\bar{n}$	σ_n
.90	E	0.89	10008	194	0.89	33590	2513	0.89	199390	7703
	SI	0.90	18333	2410	0.88	66463	4322	0.88	405028	15655
	LB	0.79	13323	2367	0.78	45693	3573	0.80	275183	12766
	UB	0.93	20313	1985	0.91	75235	4996	0.92	458808	18026
.95	E	0.94	11325	2207	0.94	41520	2814	0.93	248570	8899
	SI	0.95	23330	2478	0.94	86905	5446	0.95	530833	18977
	LB	0.88	17602	2504	0.87	64198	4613	0.88	388873	15877
	UB	0.96	25070	2382	0.95	94648	6017	0.96	578928	21233
.99	E	0.99	15778	1812	0.99	59210	3396	0.99	359600	11369
	SI	0.99	34998	2986	0.99	133215	7158	0.98	819585	27898
	LB	0.97	28878	2790	0.97	108993	6549	0.97	670203	24193
	UB	0.99	36848	3131	0.99	140858	7737	0.99	869628	29130

Table 5.1: Mixnormal coverage probabilities for relative standard deviation stopping rule

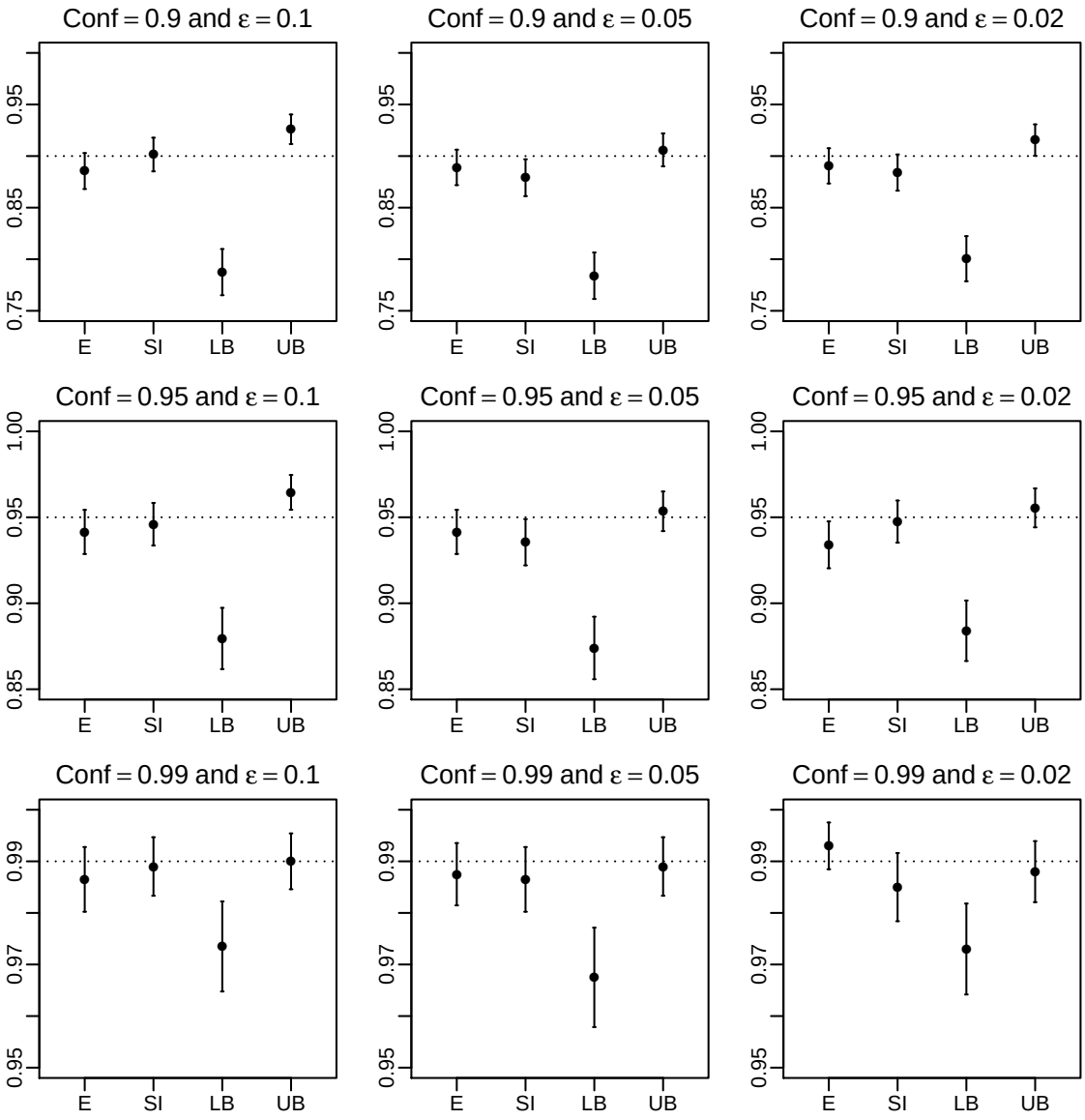

Figure 5.1: Mixnormal coverage probabilities for relative standard deviation stopping rule

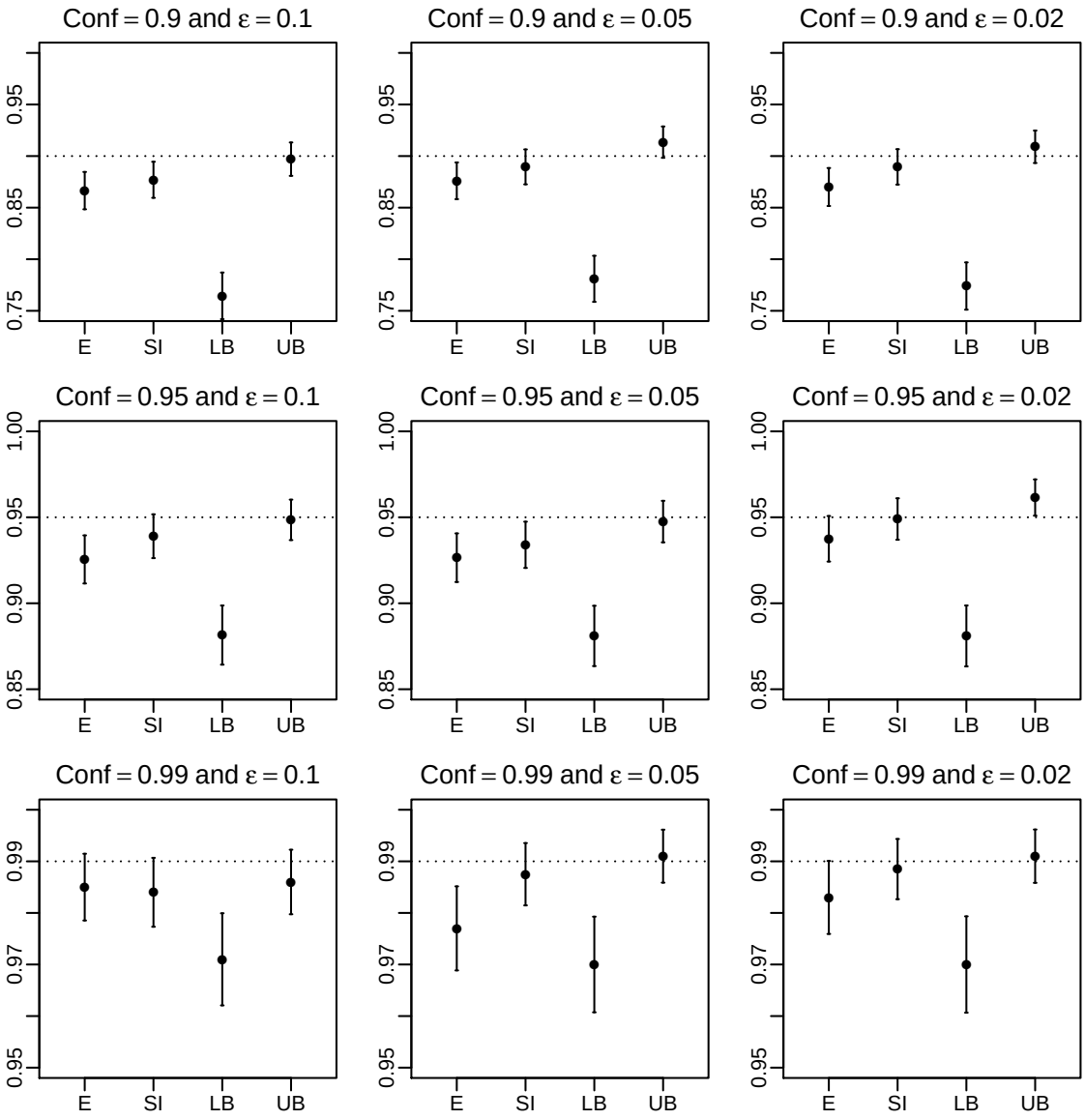

Figure 5.2: Mixnormal coverage probabilities for relative magnitude stopping rule

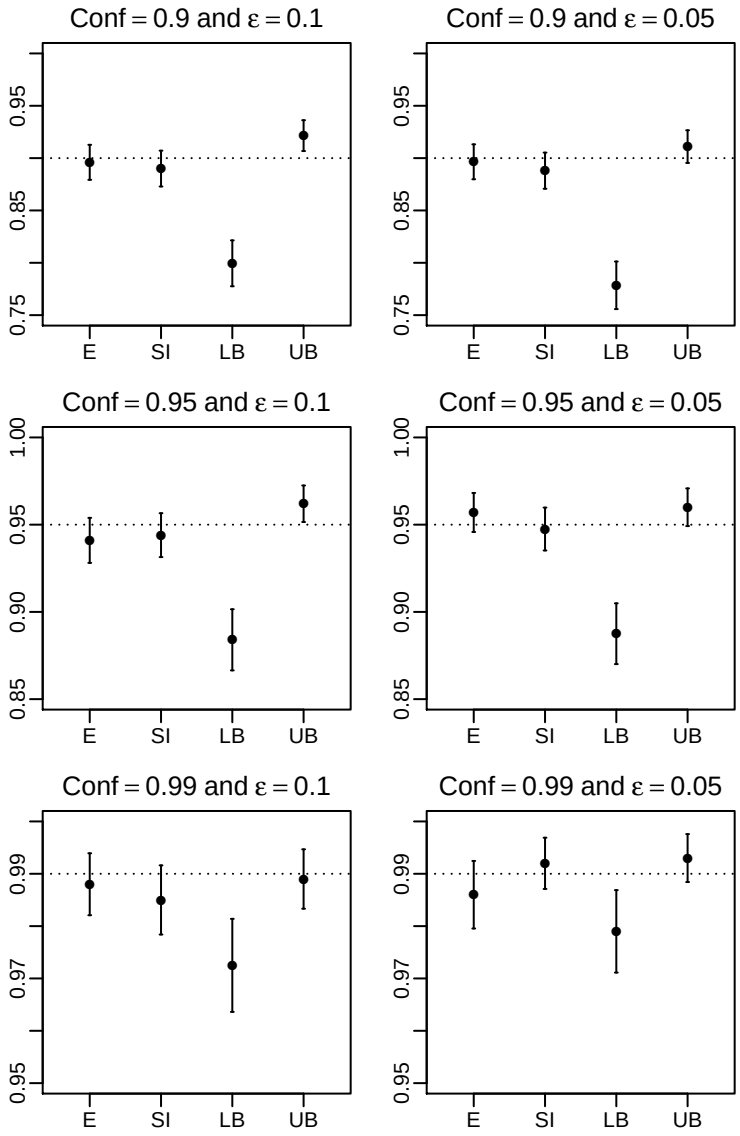

Figure 5.3: Mixnormal coverage probabilities for fixed volume stopping rule

		$\epsilon = .1$			$\epsilon = .05$			$\epsilon = .02$		
Conf		C_α	$\bar{n}$	σ_n	C_α	$\bar{n}$	σ_n	C_α	$\bar{n}$	σ_n
.90	E	0.87	5000	0	0.88	5000	0	0.87	14910	1169
	SI	0.88	5000	0	0.89	5357	1289	0.89	27655	2972
	LB	0.76	5000	0	0.78	5000	0	0.77	19588	2296
	UB	0.90	5000	0	0.91	6138	2097	0.91	30968	3136
.95	E	0.93	5000	0	0.93	5000	0	0.94	18033	2448
	SI	0.94	5000	0	0.93	7858	2475	0.95	35705	3269
	LB	0.88	5000	0	0.88	5183	938	0.88	26680	2865
	UB	0.95	5000	0	0.95	8810	2130	0.96	38620	3515
.99	E	0.98	5000	0	0.98	5035	417	0.98	24858	2190
	SI	0.98	5000	0	0.99	10203	998	0.99	53830	4268
	LB	0.97	5000	0	0.97	9765	1082	0.97	44363	3753
	UB	0.99	5000	0	0.99	10558	1582	0.99	56908	4662

Table 5.2: Mixnormal coverage probabilities for relative magnitude stopping rule

The simulation results for the relative standard deviation, relative magnitude, and fixed-volume metrics are presented in Tables 5.1, 5.2, and 5.3 respectively. Additionally, we include coverage plots inspired by Figure 4.4 in Figures 5.1, 5.2, and 5.3. All simulation settings yield more accurate coverage probabilities for smaller ϵ values with Ellipse and SI approaching the nominal level, LB approaching some value less than the nominal level, and UB approaching some value greater than the nominal level. Particularly interesting is that Ellipse consistently has the smallest average number of samples $\bar{n}$ even though it approaches correct coverage levels and LB does not. Of the interval based confidence regions we find the number of samples to be smallest for LB and largest for UB as expected. Results for the coverage probabilities at the 99% confidence level do not appear to be precise enough to demonstrate the differences in coverage. However, the average sample size at termination seems to follow a pattern consistent with the other settings. To get a sense for the spread in termination values, we report σ_n, the standard deviation of the number of samples per replicate. The relative magnitude results contain some settings for which $\sigma_n = 0$ and

		$\epsilon = .1$			$\epsilon = .05$		
Conf		C_α	$\bar{n}$	σ_n	C_α	$\bar{n}$	σ_n
.90	E	0.90	132625	6134	0.90	521280	15590
	SI	0.89	267373	11919	0.89	1058365	31881
	LB	0.80	182468	9572	0.78	718435	25106
	UB	0.92	303338	13627	0.91	1199133	37806
.95	E	0.94	165160	7005	0.96	649592	18361
	SI	0.94	350948	14757	0.95	1388540	38354
	LB	0.88	257780	12555	0.89	1016008	33821
	UB	0.96	382230	16531	0.96	1515395	44462
.99	E	0.99	238588	8922	0.99	941145	23788
	SI	0.99	540615	21181	0.99	2142218	56093
	LB	0.97	442778	18100	0.98	1754028	49018
	UB	0.99	573163	21802	0.99	2275185	60001

Table 5.3: Mixnormal coverage probabilities for fixed volume stopping rule

$\bar{n} = 5,000$. These correspond to cases where the stopping time is identified as occurring during the check at the minimum specified sample size. This acts as an indication that either the minimum sample size is high or the particular value of ϵ may not appropriately small enough for estimating these quantities.

5.4.2 Bayesian Logistic Regression

The Anguilla eel data examines 1,000 sites in New Zealand and recorded whether the presence of *Anguilla australis* was observed. Each site contains a variety of observations about the environment and how the observations were reported which may be used to predict the presence of eels. We consider fitting the following Bayesian logistic regression

model

$$
\begin{aligned}
Y_i &\sim \text{Bernoulli}(\mu_i) \\
\mu_i &= \frac{\exp\{x_i^T\beta\}}{1+\exp\{x_i^T\beta\}} \\
\beta &\sim \text{Normal}(0, 100I_d),
\end{aligned}
$$

where x_i^T is the i^{th} row of the $d \times d$ design matrix X. We aim to fit a model using the variables *DSDist*, *USNative*, *DSMaxSlope*, and *Method*. *Method* is a categorical variable will four levels yielding a total of $d = 8$ β values and $p = 24$ quantities of interest when estimating an 80% credible interval for each β. We use the function *MCMClogit* from the *R* package *MCMCpack* (Martin et al., 2011). As the true value of each quantity of interest is unknown, we take 20,000,000 samples from one long run of the Markov chain and treat the resulting estimate as "truth." Relying on an estimated value of truth is expected to affect our estimated coverage probabilities with undercoverage.

We limit ourselves to 500 replications and the fixed-volume stopping rule for the simultaneous confidence interval and ellipsoid confidence regions. A relative metric provides a way to weight ϵ to yield an appropriately small threshold for a particular sampler. We forgo a relative metric and demonstrate a small choice for ϵ is problem dependent. We start by considering the logistic regression model fit without including the categorical variable *Method*, reducing the dimension of the sampler to $d = 4$ and estimation to $p = 12$. We consider $\epsilon = .01, .005$, and .001 at a 90% confidence level with results in Table 5.4. The coverage probabilities for the ellipsoid confidence regions are very far from the nominal level

		$\epsilon = .01$			$\epsilon = .005$			$\epsilon = .001$		
Conf		C_α	$\bar{n}$	σ_n	C_α	$\bar{n}$	σ_n	C_α	$\bar{n}$	σ_n
.90	E	0.48	10000	0	0.61	19220	2684	0.83	364100	12251
	SI	0.88	14720	4997	0.89	45480	5062	0.89	997380	30914

Table 5.4: Coverage probabilities for logistic regression example without categorical variable included ($p = 12$)

		$\epsilon = .01$			$\epsilon = .005$		
Conf		C_α	$\bar{n}$	σ_n	C_α	$\bar{n}$	σ_n
.90	E	0.61	90260	4167	0.78	364260	8934
	SI	0.86	285480	8862	0.89	1118020	21708

Table 5.5: Coverage probabilities for logistic regression example with categorical variable included ($p = 24$)

for $\epsilon = .01$ and .005. We note that the largest ϵ we considered in this problem is smaller than the smallest value we considered in our mixture normal example, demonstrating that ϵ is problem specific. The ellipsoid confidence region particularly suffers when $\epsilon = .01$ as every replicate terminated at the minimum sample size 10,000. As expected, the coverage probabilities improve as ϵ decreases with $\epsilon = .001$, yielding an estimated coverage probability that may be reasonable considering the "truth" is an estimated quantity. The simultaneous intervals confidence region fairs better with close coverage probabilities, even in the $\epsilon = .01$ case. The simultaneous interval procedure includes a line search component which may be providing some self-correction in the choice of z^* causing the improved accuracy in coverage probabilities.

We now consider the full logistic regression model with $d = 8$ and $p = 24$ in Table 5.5. The additional terms in the model appear to help alleviate the termination at minimum sample size when $\epsilon = .01$. The results are similar to those in Table 5.4 in

that coverage probabilities improve as ϵ decreases, but the sample size at termination is significantly larger.

Chapter 6

Future Work

The visualizations developed in Chapter 4 and stopping rule proposed in Chapter 5 provide useful tools for analyzing MCMC output. Future work will focus on four main tasks: relaxing conditions on the Markov chain in Theorem 13, improving Theorem 2 by incorporating the distribution of $\hat{\Lambda}$, extending our results to other samplers such as Adaptive MCMC (Rosenthal, 2011), and output analysis based on parallel MCMC simulations.

Our current conditions on the Markov chain in Theorem 13 are largely determined by the need for a strongly consistent density estimator as the other assumptions are satisfied with merely polynomial ergodicity. Results from Roussas (1988) look promising, however it is not obvious that all the assumptions made are satisfied without stronger assumptions aside from the weakened ergodicity. With the newly relaxed conditions a new limiting distribution may be found for Theorem 2. This new limiting distribution should converge as $n \to \infty$ to a multivariate normal distribution and mirror the result as in the case with just expectations where a batch means estimator yields a Hoetelling's T^2 distribution. This new

limiting distribution should improve coverage probabilities and result in smaller stopping times.

Adaptive MCMC samplers are an MCMC-like algorithm in which the covariance of the proposal distribution updates every n° steps. The updating mechanism is non-Markovian but still may satisfy a CLT with restrictions placed on the adaptation. In particular, cases where the covariance converges to a constant or changes very little, known as diminishing adaptation, may emit a CLT. This CLT may then be used to create visualizations such as those we have proposed. Extensions of the sequential stopping rules are more challenging as the set of assumptions are stricter and not obviously satisfied.

Parallel computing provides another arena for which our tools may prove useful. We concern ourselves with a recent parallel approach to sequential stopping rules for stochastic simulations proposed by Dong and Glynn (2016). By using independent parallel simulations they avoid the assumption of a strongly consistent estimate of the covariance of the CLT for the Monte Carlo error. Their methods focus on general stochastic simulations, however no assumptions are violated in the MCMC machinery and thus we focus on the MCMC setting. The general idea is to simulate m independent replications of a chain (sectioning) and compute a parameter estimate $\tilde{\theta}_m$ and a sample variance estimator $S_m^2(n)$ using all m chains. A major advantage to this strategy is that $S_m^2(n)$ does not need to be a strongly consistent estimator of σ^2. In addition, each chain is independent and thus parallel computing is employable.

6.1 Sectioning Method

For a fixed number of chains m, as $n \to \infty$

$$\frac{\sqrt{m}(\tilde{\theta}_m - \theta)}{S_m^2(n)} \xrightarrow{D} t_{m-1}, \tag{6.1}$$

where t_{m-1} is the Student-t distribution with $m-1$ degrees of freedom. This yields the fixed sample size confidence interval

$$C_1[n] = \left[\tilde{\theta}_m - t_{\delta/2,m-1}\frac{S_m(n)}{\sqrt{m}}, \tilde{\theta}_m - t_{\delta/2,m-1}\frac{S_m(n)}{\sqrt{m}}\right] \tag{6.2}$$

with $t_{\delta/2,m-1}$ chosen such that $P(-t_{\delta/2,m-1} < t_{m-1} < t_{\delta/2,m-1})$. If we want a sequential stopping rule, then the simulation stops when $u_{\delta/2,m}S_m(n)/\sqrt{m} \leq \epsilon$ where $u_{\delta/2,m}$ is the upper $\delta/2$ quantile of a distribution, u presented in Table 6.1, which will be discussed in detail later.

We let $\theta = \psi(F)$ where ψ is a real-valued functional on the space of probability distributions and F is the stationary distribution of X. We will primarily concern ourselves with expectations such as $\psi(F) = \int xF(dx)$, however other functionals such as quantiles will satisfy the following constraints. Let $\hat{F}(n,\cdot)$ denote the empirical distribution based on the chain $\{X(s) : 0 \leq s \leq n\}$, that is

$$\hat{F}(n,x) = \frac{1}{n}\sum_{k=1}^{n} 1\{X(k) \leq x\}. \tag{6.3}$$

We are also interested in an empirical distribution based on all m chains

$$\hat{F}_0^m(n,x) = \frac{1}{nm}\sum_{i=1}^{m}\sum_{k=1}^{n} 1\{X(k) \leq x\} = \frac{1}{m}\sum_{i=1}^{m}\hat{F}_i(n,x). \tag{6.4}$$

Let $\tilde{\theta}_m = \psi(\hat{F}_0^m(n))$ be an estimator for θ based on the m chains. The following assumptions will be required to establish the proper probability coverage.

Assumption 3 *$\tilde{\theta}_m$ satisfies a FCLT.*

Assumption 4 *As $\epsilon \to 0$,*

$$\frac{n}{\epsilon}\left(\psi\left(\hat{F}_0^m(n/\epsilon^2)\right) - \frac{1}{m}\sum_{i=1}^{m}\psi\left(\hat{F}_i(n/\epsilon^2)\right)\right) \Rightarrow 0. \tag{6.5}$$

Based on these assumptions we may estimate θ with either $\frac{1}{m}\sum_{i=1}^{m}\psi(\hat{F}_i(n))$ or $\psi(\hat{F}_0^m(n))$. Typically $\psi(\hat{F}_0^m(n))$ is preferred as it has less bias for finite sample sizes (Dong and Glynn, 2016).

The error size may be estimated with

$$\Gamma(n,\hat{F}) = \sqrt{\frac{1}{m(m-1)}\sum_{i=1}^{m}\left(\psi\left(\hat{F}_i(n)\right) - \psi\left(\hat{F}_0^m(n)\right)\right)^2}. \tag{6.6}$$

This $\Gamma(n,\hat{F})$ depends on the function ψ. We will suppress this in our notation for clarity.

Proposition 14 *(Dong and Glynn, 2016) Under Assumption 3 and 4*

$$\frac{\psi\left(\hat{F}_0^m(n/\epsilon^2)\right) - \theta}{\Gamma(n,\hat{F})} \Rightarrow \frac{\bar{B}(n)}{\sqrt{\frac{1}{m(m-1)}\sum_{i=1}^{m}\left(B(n) - \bar{B}(n)\right)^2}} \tag{6.7}$$

as $\epsilon \to 0$, *where* B_i *for* $i = 1, ..., m$ *are independent Brownian motion and* $\bar{B}(n) = \sum_{i=1}^{m} B_i(n)$. *For a fixed* n *this follows a* t_{m-1} *distribution.*

This then allows the construction of $100(1-\delta)\%$ confidence intervals with

$$\left[\psi\left(\hat{F}_0^m(n)\right) - t_{\delta/2,m-1}\Gamma\left(n,\hat{F}\right), \psi\left(\hat{F}_0^m(n)\right) + t_{\delta/2,m-1}\Gamma\left(n,\hat{F}\right)\right]. \tag{6.8}$$

6.1.1 Sectioning method sequential stopping rules

We now consider sequential stopping rules for the sectioning method. To prevent early termination for badly behaved $\Gamma\left(n,\hat{F}\right)$, let $a(n)$ be a strictly positive function monotonically decreasing to 0, satisfying $a(n) = O(n^{-\gamma})$ for $\gamma > 1/2$, to ensure $a(n)$ decays at a faster rate than $\Gamma\left(n,\hat{F}\right)$. We define our stopping rule

$$\kappa(\epsilon) = \inf\left\{n > 0 : \Gamma(n,\hat{F}) + a(n) < \epsilon\right\}. \tag{6.9}$$

This again depends on the function ψ which we will suppress in our notation. To determine the limiting distribution of $\kappa(\epsilon)$ we first define

$$K(\sigma) = \inf\left\{n > 0 : \sigma\sqrt{\frac{1}{m(m-1)n^2}\sum_{i=1}^{m}(B_i(n) - \bar{B}(n))^2} < 1\right\}. \tag{6.10}$$

The distribution of $K(\sigma)$ is determined by noting that $\frac{K(\sigma)}{\sigma^2} = K(1)$ and the distribution of $K(1)$ is provided by the following lemma.

Lemma 15 *(Dong and Glynn, 2016)*

(1) When $m = 2, K(1) = 0$.

(2) When $m = 3, K(1) = 0$.

(3) When $m \geq 4, K(1)$ *follows a Gamma distribution with shape parameter* $\gamma = (m-3)/2$ *and rate parameter* $\lambda = m(m-1)/2$.

Theorem 16 *(Dong and Glynn, 2016) Under Assumption 3 and 4, for* $m \geq 4$,

$$\frac{\psi\left(\hat{F}_o^m(\kappa(\epsilon))\right) - \theta}{\Gamma(\kappa(\epsilon))} \Rightarrow \frac{\sigma \bar{B}(K(\sigma))}{K(\sigma)} \stackrel{d}{=} \frac{Z}{\sqrt{mK(1)}}. \tag{6.11}$$

as $\epsilon \to 0$, *where* $Z \sim N(0,1)$ *is independent of* $K(1)$.

Let u be the appropriate quantile from the distribution $Z/\sqrt{mK(1)}$. This lets us ease notation by defining a stopping rule based on the limiting distribution u,

$$\kappa_1(\epsilon) = \inf\left\{n > 0 : u\left(\Gamma(n, \hat{F}) + a(n)\right) < \epsilon\right\}. \tag{6.12}$$

Now we may construct $100(1-\delta)\%$ confidence intervals with the following

$$C_m[\kappa_1(\epsilon)] = \left[\psi\left(\hat{F}_0^m(\kappa_1(\epsilon))\right) \pm u_{\delta/2}\left(\Gamma(\kappa_1(\epsilon), \hat{F}) + n(\kappa_1(\epsilon))\right)\right]. \tag{6.13}$$

Some previously computed quantiles of the distribution $Z/\sqrt{mK(1)}$ were calculated by Dong and Glynn (2016) and are reported in table 6.1. $Z/\sqrt{mK(1)}$ has larger tails than the corresponding t distribution used in non-sequential procedures. Plots of $Z/\sqrt{mK(1)}$ for $m = 10$ and $m = 20$ overlaid with the corresponding t distributions for a non-sequential are presented in Figure 6.1.

	.9	.925	.95	.975	.995
m = 10	1.603±.001	1.833±.001	2.148±.002	2.680±.002	3.963±.005
m = 15	1.465±.001	1.661±.001	1.925±.002	2.352±.002	3.298±.003
m = 20	1.409±.001	1.593±.001	1.840±.001	2.230±.002	3.064±.003
m = 25	1.380±.001	1.558±.001	1.792±.001	2.164±.002	2.945±.003
m = 30	1.361±.001	1.536±.001	1.765±.001	2.126±.002	2.872±.002

Table 6.1: (Dong and Glynn, 2016) Quantiles of $Z/\sqrt{mK(1)}$ based on 10^6 i.i.d. samples

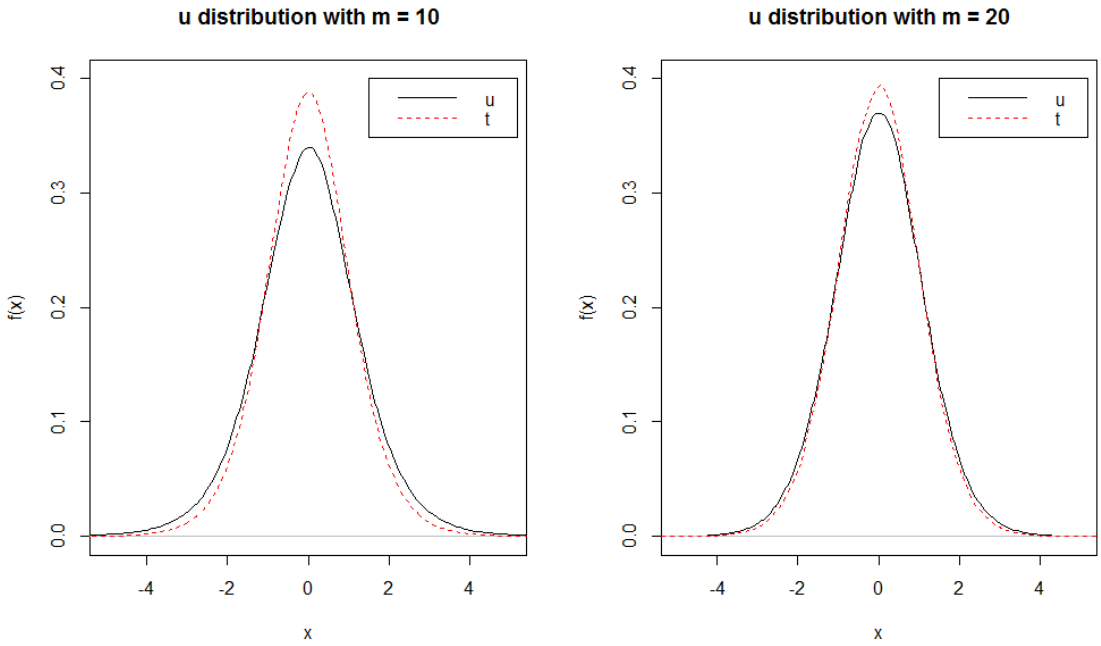

Figure 6.1: Density approximation of $u = \frac{Z}{\sqrt{mK(1)}}$ for $m = 10$ and $m = 20$ compared to the t distribution with 9 and 19 degrees of freedom respectively.

We may now consider a relative metric ν_θ with estimator $\hat{\nu}_m$ based on the m sections. The stopping rule may now be defined as

$$\kappa^*(\epsilon) = \inf\left\{n > 0 : u\left(\Gamma(n, \hat{F}) + a(n)\right) < \epsilon|\hat{\nu}_m|\right\}. \tag{6.14}$$

The asymptotic validity is then given as a generalization to Dong and Glynn's (2016) Theorem 4.

Proposition 17 *Under Assumption 3 and 4, for* $m \geq 4$,

$$\frac{\psi\left(\hat{F}_o^m(\kappa^*(\epsilon))\right) - \theta}{\Gamma(\kappa^*(\epsilon))} \Rightarrow \frac{\sigma\bar{B}(K(\sigma/|\nu_n|))}{K(\sigma)} \stackrel{d}{=} \frac{Z}{\sqrt{mK(1)}}, \tag{6.15}$$

as $\epsilon \to 0$, *where* $Z \sim N(0,1)$ *is independent of* $K(1)$.

A relative error sequential stopping rule then is defined as

$$\kappa_2(\epsilon) = \inf\left\{n > 0 : u\left(\Gamma(n, \hat{F}) + a(n)\right) < \epsilon\left|\psi\left(\hat{F}_0^m(n)\right)\right|\right\}. \tag{6.16}$$

The following theorem characterizes the distribution yielding the scaling parameter of interest.

Theorem 18 *(Dong and Glynn, 2016) Under Assumption 3 and 4, for* $m \geq 4$,

$$\frac{\psi\left(\hat{F}_o^m(\kappa_2(\epsilon))\right) - \theta}{\Gamma(\kappa_2(\epsilon))} \Rightarrow \frac{\sigma\bar{B}(K(\sigma/|\theta|))}{K(\sigma)} \stackrel{d}{=} \frac{Z}{\sqrt{mK(1)}}. \tag{6.17}$$

as $\epsilon \to 0$, *where* $Z \sim N(0,1)$ *is independent of* $K(1)$.

The resulting $100(1-\delta/2)\%$ confidence interval is

$$C_m[\kappa_2(\epsilon)] = \left[\psi\left(\hat{F}_0^m(\kappa_2(\epsilon))\right) \pm u_{\delta/2}\left(\Gamma(\kappa_2(\epsilon), \hat{F}) + a(\kappa_2(\epsilon))\right)\right]. \tag{6.18}$$

We now consider our Bayesian setting and define

$$\kappa_3(\epsilon) = \inf\left\{n > 0 : u\left(\Gamma(n, \hat{F}) + a(n)\right) < \epsilon\hat{\lambda}_m\right\}, \tag{6.19}$$

as the posterior standard deviation stopping rule where $\hat{\lambda}$ is the estimated posterior standard deviation. The following theorem characterizes the distribution yielding the scaling parameter of interest.

Theorem 19 *Under Assumption 3 and 4, for $m \geq 4$,*

$$\frac{\psi\left(\hat{F}_o^m(\kappa_3(\epsilon))\right) - \theta}{\Gamma(\kappa_3(\epsilon))} \Rightarrow \frac{\sigma\bar{B}(K(\sigma/\lambda))}{K(\sigma)} \stackrel{d}{=} \frac{Z}{\sqrt{mK(1)}}. \tag{6.20}$$

as $\epsilon \to 0$, where $Z \sim N(0,1)$ is independent of $K(1)$.

The resulting $100(1-\delta/2)\%$ confidence interval is

$$C_m[\kappa_3(\epsilon)] = \left[\psi\left(\hat{F}_0^m(\kappa_3(\epsilon))\right) \pm u_{\delta/2}\left(\Gamma(\kappa_3(\epsilon), \hat{F}) + a(\kappa_3(\epsilon))\right)\right]. \tag{6.21}$$

6.1.2 Example

We consider a Metropolis-Hastings random walk with a Normal($\mu = 2, \sigma^2 = 2$) target distribution and a Normal($0, \sigma^2 = 1/2$) increment distribution. Thus we are trying

to make an inference about $\theta = \mu$ and use the following sampler. We start with an initial value x_0 and let $f(x)$ be the pdf of the target distribution.

1. Draw $\epsilon \sim N(0, 1/2)$
2. Set $x^* = x_{t-1} + \epsilon$
3. Set $a = \min\left\{1, \frac{f(x^*)}{f(x_{t-1})}\right\}$
4. Let $x_t = \begin{cases} x^* & \text{w.p. } a \\ x_{t-1} & \text{w.p. } 1-a \end{cases}$
5. repeat steps 1-4.

The results are contained in Table 6.2. We once again see our fixed-width rules yielding the largest stopping times for each value of ϵ. Every simulation attained a coverage probability of at least $1-\delta/2$, however there are some issues of overcoverage. This stems from not using a CLT based procedure as we are basing the distribution u on the exact stopping time and do not have u as a limiting distribution for $n \to \infty$. As such we obtain overcoverage when we sample more than $\kappa_i(\epsilon)$ which occurs when $\kappa_i(\epsilon)$ is greater than the minimum simulation effort. In these cases the proper coverage distribution should be between $Z/\sqrt{mK(1)}$ and t. It is important to note that the recorded values of n are for each independent chain and thus for each simulation the total number of samples is $m \times n$. One advantage gained here is that since each chain is independent parallel computing may be employed and may in certain situations be faster overall than CLT based methods.

ϵ	stopping rule	mean θ estimate	mean n	std. dev. n	coverage probability
.1	$\kappa_1(\epsilon)$	1.994	3380	1610	.900
.1	$\kappa_2(\epsilon)$	1.987	1160	276	.929
.1	$\kappa_3(\epsilon)$	1.989	1820	749	.917
.05	$\kappa_1(\epsilon)$	1.998	12700	6270	.910
.05	$\kappa_2(\epsilon)$	1.996	3400	1570	.914
.05	$\kappa_3(\epsilon)$	1.998	6640	3310	.912
.02	$\kappa_1(\epsilon)$	2.000	73700	37600	.915
.02	$\kappa_2(\epsilon)$	2.000	19300	9280	.911
.02	$\kappa_3(\epsilon)$	2.000	37800	18500	.903

Table 6.2: Results of 1000 replications of Metropolis-Hastings random walk sampler for MCMC sectioning stopping rules based on 90% confidence intervals and $m = 10$ independent chains.

6.2 Multivariate Section Methods

Sectioning methods present their own set of challenges in developing extensions to multivariate settings. A naive approach estimates each dimension separately and terminates once each component has met its stopping time for a Bonferroni adjusted confidence level. More precisely, for $\Theta \in \mathbb{R}^p$, let $\kappa^{(i)}(\epsilon)$ be the stopping time with confidence level $1 - \alpha/(2p)$ based on estimating Θ_i with the marginal process $\{X_{ij}\}_{j=1}^{\infty}$. We define the stopping time

$$\kappa^{\text{comp}}(\epsilon) = \sup\{\kappa^{(1)}(\epsilon), ..., \kappa^{(p)}(\epsilon)\}. \tag{6.22}$$

The resulting coverage probability for $\kappa^{\text{comp}}(\epsilon)$ will approach a limit greater than $1 - \alpha$ as $\epsilon \to 0$. Generally overcoverage is not a major concern in terms of accuracy, but the increased computational burden is problematic. Further, no information about the dependencies between quantities of interest is maintained. We turn our attention to developing results analogous to Vats et al.'s (2019) extension of univariate procedures in Jones et al. (2006).

Consider the univariate setting where for a fixed sample size Proposition 14 gives

$$\frac{\psi\left(\hat{F}_0^m(n/\epsilon^2)\right) - \tilde{\theta}_m}{\Gamma(n,\hat{F})} \Rightarrow \frac{\bar{B}(n)}{\sqrt{\frac{1}{m(m-1)}\sum_{i=1}^m \left(B(n) - \bar{B}(n)\right)^2}} \sim t_{m-1}.$$

Squaring the distribution allows us to write this in matrix notation. Let

$$\Phi = \left(\frac{1}{m(m-1)}\right)\left[\sum_{i=1}^m \left(\psi(\hat{F}_i(n)) - \psi(\hat{F}_0^m(n))\right)\left(\psi(\hat{F}_i(n)) - \psi(\hat{F}_0^m(n))\right)^T\right],$$

then

$$\left[\frac{\psi\left(\hat{F}_0^m(n/\epsilon^2)\right) - \tilde{\theta}_m}{\Gamma(n,\hat{F})}\right]^2 = \left[\psi\left(\hat{F}_0^m(n/\epsilon^2)\right) - \tilde{\theta}_m\right]^T \Phi^{-1}\left[\psi\left(\hat{F}_0^m(n/\epsilon^2)\right) - \tilde{\theta}_m\right] \Rightarrow T^2_{1,m-1},$$

where $T^2_{1,m-1}$ is Hotelling's T-squared distribution with dimensionality parameter 1 and $m-1$ degrees of freedom. This inspires a potential stopping rule based on the region

$$C_\alpha^{m,p}(n) = \left[\psi\left(\hat{F}_0^m(n/\epsilon^2)\right) - \tilde{\theta}_m\right]^T \Phi^{-1}\left[\psi\left(\hat{F}_0^m(n/\epsilon^2)\right) - \tilde{\theta}_m\right]. \tag{6.23}$$

The potential stopping rule is then

$$\kappa_\alpha^{m,p}(\epsilon) = \inf\left\{n : \mathrm{Vol}(C_\alpha^{m,p}(n))^{1/p} + a(n) < \epsilon\right\}. \tag{6.24}$$

A major challenge with (6.24) is identifying the form of $\mathrm{Vol}(C_\alpha^m(n))$ when n is taken to be a random stopping time. The univariate case uses a cancellation technique to avoid requiring a strongly consistent estimator of the process variance. It is not clear that this approach exists in the multivariate case as there is now a covariance to consider.

Bibliography

Acosta, F., Huber, L. M., and Jones, G. L. (2014). Markov chain monte carlo with linchpin variables. Technical report.

Andrews, D. (1991). Heteroskedasticity and autocorrelation consistent covariant matrix estimation. *Econometrica*, 59:817–858.

Babu, G. J. and Rao, C. R. (1988). Joint asymptotic distribution of marginal quantiles and quantile functions in samples from a multivariate population. *Journal of Multivariate Analysis*, 27:15–23.

Bahadur, R. R. (1966). A note on quantiles in large samples. *The Annals of Mathematical Statistics*, 37:577–580.

Billingsley, P. (1968). *Convergence of Probability Measures*. Wiley, New York.

Birkhoff, G. D. (1931). Proof of the ergodic theorem. *Proceedings of the National Academy of Sciences*, 17(12):656–660.

Blum, J. and Hanson, D. (1960). On the mean ergodic theorem for subsequences. *Bulletin of the American Mathematical Society*, 66:308–311.

Bradley, R. C. (1986). Basic properties of strong mixing conditions. In Eberlein, E. and Taqqu, M. S., editors, *Dependence in Probability and Statistics: A Survey of Recent Results*, pages 165–192. Birkhauser, Cambridge, MA.

Bradley, R. C. (2005). Basic properties of strong mixing conditions. a survey and some open questions. *Probability surveys*, 2:107–144.

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

Chan, K. W. and Yau, C. Y. (2017). Automatic optimal batch size selection for recursive estimators of time-average covariance matrix. *Journal of the American Statistical Association*, 112:1076–1089.

Chen, D.-F. R. and Seila, A. F. (1987). Multivariate inference in stationary simulation using batch means. In *Proceedings of the 19th conference on Winter simulation*, pages 302–304. ACM.

Dai, N. and Jones, G. L. (2017). Multivariate initial sequence estimators in Markov chain Monte Carlo. *Journal of Multivariate Analysis*, 159:184–199.

Dong, J. and Glynn, P. (2016). A new approach to sequential stopping for stochastic simulation. Technical report, Northwestern University, School of Indusrial Engineering.

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

Doss, H. and Hobert, J. P. (2010). Estimation of Bayes factors in a class of hierarchical random effects models using a geometrically ergodic MCMC algorithm. *Journal of Computational and Graphical Statistics*, 19:295–312.

Ferguson, T. S. (1998). Asymptotic joint distribution of sample mean and a sample quantile. Technical report, UCLA.

Fishman, G. S. (1996). *Monte Carlo: Concepts, Algorithms, and Applications*. Springer, New York.

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

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

Friedman, J., Hastie, T., and Tibshirani, R. (2009). glmnet: Lasso and elastic-net regularized generalized linear models. *R package version*, 1.

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

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., and Hothorn, T. (2018). mvtnorm: Multivariate Normal and t distributions. R package version 1.0-8.

Ghosh, J. K. (1971). A new proof of the Bahadur representation of quantiles and an application. *The Annals of Mathematical Statistics*, pages 1957–1961.

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

Gong, L. and Flegal, J. M. (2016). A practical sequential stopping rule for high-dimensional Markov chain Monte Carlo. *Journal of Computational and Graphical Statistics*, 25:684–700.

Hamada, M. S., Wilson, A. G., Reese, C. S., and Martz, H. F. (2008). *Bayesian Reliability*. Spring Series in Statistics. Springer-Verlag New York.

Heidelberger, P. and Welch, P. D. (1983). Simulation run length control in the presence of an initial transient. *Operations Research*, 31:1109–1144.

Hijmans, R. J., Phillips, S., Leathwick, J., and Elith, J. (2017). dismo: Species distribution modeling. R package version 1.1-4.

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.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67.

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

Ibragimov, I. A. and Linnik, Y. V. (1971). *Independent and Stationary Sequences of Random Variables*. Walters-Noordhoff, The Netherlands.

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

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

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

Jones, G. L. and Hobert, J. P. (2004). Sufficient burn-in for Gibbs samplers for a hierarchical random effects model. *The Annals of Statistics*, 32:784–817.

Jones, G. L., Roberts, G. O., and Rosenthal, J. S. (2012). Convergence of conditional Metropolis-Hastings samplers, with an application to inference for discretely-observed diffusions. Technical report, University of Minnesota, School of Statistics.

Kay, M. (2018). tidybayes: Tidy data and geoms for Bayesian models. R package version 1.0.3.

Koehler, E., Brown, E., and Haneuse, S. J.-P. A. (2009). On the assessment of Monte Carlo error in simulation-based statistical analyses. *The American Statistician*, 63:155–162.

Lin, P.-E., Wu, K.-T., and Ahmad, I. A. (1980). Asymptotic joint distribution of sample quantiles and sample mean with applications. *Communications in Statistics-Theory and Methods*, 9:51–60.

Liu, Y. and Flegal, J. (2018a). Optimal mean squared error bandwidth for spectral variance estimators in mcmc simulations. *ArXiv e-prints*.

Liu, Y. and Flegal, J. M. (2018b). Weighted batch means estimators in Markov chain Monte Carlo. *Electron. J. Statist.*, 12:3397–3442.

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.

Martin, A. D., Quinn, K. M., and Park, J. H. (2011). MCMCpack: Markov chain monte carlo in R. *Journal of Statistical Software*, 42(9):22.

Masry, E. and Györfi, L. (1987). Strong consistency and rates for recursive probability density estimators of stationary processes. *Journal of Multivariate Analysis*, 22:79–93.

Meyn, S. and Tweedie, R. (2009). *Markov Chains and Stochastic Stability*, volume 2. Cambridge University Press Cambridge.

Meyn, S. P. and Tweedie, R. L. (1993). *Markov Chains and Stochastic Stability*. Springer-Verlag, London.

Priestley, M. B. (1981). *Spectral Analysis and Time Series. (Vol. 1): Univariate Series*. Academic Press.

R Core Team (2013). R: A language and environment for statistical computing.

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. (2001). Optimal scaling for various metropolis-hastings algorithms. *Statist. Sci.*, 16(4):351–367.

Robertson, N., Flegal, J. M., Jones, G. L., and Vats, D. (2019). New visualizations for Monte Carlo simulations. *ArXiv e-prints*.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.*, 27(3):832–837.

Rosenthal, J. S. (2011). Optimal proposal distributions and adaptive MCMC. In Brooks, S., Gelman, A., Meng, X.-L., and Jones, G. L., editors, *Handbook of Markov chain Monte Carlo*. Chapman & Hall, Boca Raton.

Roussas, G. G. (1988). Nonparametric estimation in mixing sequences of random variables. *Journal of Statistical Planning and Inference*, 18:135–149.

Ryzin, J. V. (1969). On strong consistency of density estimates. *The Annals of Mathematical Statistics*, 40(5):1765–1772.

Seila, A. F. (1982). Multivariate estimation in regenerative simulation. *Operations Research Letters*, 1:153–156.

Sen, P. K. (1968). Asymptotic normality of sample quantiles for m-dependent processes. *Ann. Math. Statist.*, 39(5):1724–1730.

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

Serfling, R. J. (1981). *Approximation Theorems of Mathematical Statistics (Wiley Series in Probability and Statistics)*. Wiley-Interscience.

Shubin, M. (2015). Principles of posterior visualization [blog post]. Retrieved from https://ctg2pi.wordpress.com/2015/02/24/principles-of-posterior-visualization.

Silverman, B. W. (1986). Density estimation for statistics and data analysis. *Monographs on Statistics and Applied Probability*.

Stan Development Team (2018). RStan: the R interface to Stan. R package version 2.17.4.

Stigler, S. M. (1973). Studies in the History of Probability and Statistics. XXXII: Laplace, Fisher, and the discovery of the concept of sufficiency. *Biometrika*, 60:439–445.

Takahata, H. (1980). Almost sure convergence of density estimators for weakly dependent stationary processes. *Bulletin of Tokyo Gakugei University, Series IV*, 32.

Tan, A. and Hobert, J. P. (2009). Block Gibbs sampling for Bayesian random effects models with improper priors: Convergence and regeneration. *Journal of Computational and Graphical Statistics*, 18:861–878.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58:267–288.

Vats, D. and Flegal, J. M. (2018). Lugsail lag windows and their application to MCMC. *ArXiv e-prints*.

Vats, D., Flegal, J. M., and Jones, G. L. (2018). Strong consistency of multivariate spectral variance estimators in Markov chain Monte Carlo. *Bernoulli*, 24:1860–1909.

Vats, D., Flegal, J. M., and Jones, G. L. (2019). Multivariate output analysis for Markov chain Monte Carlo. *Biometrika*, 106:321–337.

Wang, X., Hu, S., and Yang, W. (2011). The Bahadur representation for sample quantiles under strongly mixing sequence. *Journal of Statistical Planning and Inference*, 141:655–662.

Yoshihara, K. (1995). The Bahadur representation of sample quantiles for sequences of strongly mixing random variables. *Statistics & Probability Letters*, 24:299–304.