Chapter 1

Implementing Markov chain Monte Carlo: Estimating with confidence

James M. Flegal and Galin L. Jones

(June 30, 2010)

1.1 Introduction

Our goal is to introduce some of the tools useful for analyzing the output of a Markov chain Monte Carlo (MCMC) simulation. In particular, we focus on methods which allow the practitioner (and others!) to have confidence in the claims put forward. The following are the main issues we will address: (1) initial graphical assessment of MCMC output; (2) using the output for estimation; (3) assessing the Monte Carlo error of estimation; and (4) terminating the simulation.

Let π be a probability function with support $\mathsf{X} \subseteq \mathbb{R}^d$ (most often π will be a probability mass function or a probability density function) about which we wish to make an inference. This inference often is based on some feature of π. For example, if $g : \mathsf{X} \to \mathbb{R}$ a common goal is the calculation of[1]

$$E_\pi g = \int_{\mathsf{X}} g(x)\pi(dx) \ . \tag{1.1.1}$$

We will typically want the value of several features such as mean and variance parameters along with quantiles and so on. As a result, the features of interest form a p-dimensional vector which we call θ_π. Unfortunately, in practically relevant settings we often cannot calculate any of the components of θ_π analytically or even numerically. Thus we are faced with a classical statistical problem; given a probability distribution π we want to estimate several fixed, unknown features of it. For ease of exposition we focus on the case where θ_π is univariate but we will come back to the general case at various points throughout.

Consider estimating an expectation as in (1.1.1). The basic MCMC method entails constructing a Markov chain $X = \{X_0, X_1, X_2, \ldots\}$ on X having π as its invariant distribution. (See Geyer (2010) for an introduction to MCMC algorithms.) Then we simulate X for a finite number of steps, say n, and use the observed values to estimate $E_\pi g$ with a sample average

$$\bar{g}_n := \frac{1}{n}\sum_{i=0}^{n-1} g(x_i) \ . \tag{1.1.2}$$

The use of this estimator is justified through the Markov chain strong law of large numbers (SLLN)[2]: If $E_\pi|g| < \infty$, then $\bar{g}_n \to E_\pi g$ almost surely as $n \to \infty$. From a practical point of view this means we can obtain an accurate estimate of $E_\pi g$ with a sufficiently long simulation.

Outside of toy examples, no matter how long our simulation, there will be an unknown *Monte Carlo error*, $\bar{g}_n - E_\pi g$. While it is impossible to assess this error directly, we can obtain its approximate sampling distribution if a Markov chain central limit theorem (CLT) holds. That is, if

$$\sqrt{n}(\bar{g}_n - E_\pi g) \xrightarrow{d} \mathrm{N}(0, \sigma_g^2) \tag{1.1.3}$$

as $n \to \infty$ where $\sigma_g^2 \in (0, \infty)$. It is important to note that due to the correlation present in a Markov chain $\sigma_g^2 \neq \mathrm{var}_\pi g$ except in trivial cases. For now, suppose we have an estimator such that $\hat{\sigma}_n^2 \to \sigma_g^2$ almost surely as $n \to \infty$ (see Section 1.4 for some suitable techniques). This

[1] The notation in (1.1.1) avoids having separate formulas for the discrete case where it denotes $\sum_{x \in \mathsf{X}} g(x)\,\pi(x)$ and the continuous case where it denotes $\int_{\mathsf{X}} g(x)\,\pi(x)\,dx$.

[2] This is a special case of the Birkhoff Ergodic Theorem (p. 558 Fristedt and Gray, 1997).

allows construction of an asymptotically valid confidence interval for $E_\pi g$ with half-width

$$t_* \frac{\hat{\sigma}_n}{\sqrt{n}} \tag{1.1.4}$$

where t_* is an appropriate quantile.

Most importantly, calculating and reporting the Monte Carlo standard error (MCSE), $\hat{\sigma}_n/\sqrt{n}$, allows everyone to judge the reliability of the estimates. In practice this is done in the following way. Suppose after n simulations our estimate of $E_\pi g$ is $\bar{g}_n = 1.3$. Let h_α denote the half-width given in (1.1.4) of a $(1-\alpha)100\%$ interval. We can be confident in the "3" in our estimate if $1.3 \pm h_\alpha \subseteq [1.25, 1.35)$. Otherwise, reasonable values such as 1.2 or 1.4 could be obtained by rounding. If the interval is too wide for our purposes, then more simulation should be conducted. Of course, we would be satisfied with a wider interval if we only wanted to trust the "1" or the sign of our estimate. Thus the interval estimator (1.1.4) allows us to describe the confidence in the reported estimate, and moreover, including an MCSE with the point estimate allows others to assess its reliability. Unfortunately, this is not currently standard practice in MCMC (Flegal et al., 2008).

The rest of this chapter is organized as follows. In Section 1.2 we consider some basic techniques for graphical assessment of MCMC output then Section 1.3 contains a discussion of various point estimators of θ_π. Next, Section 1.4 introduces techniques for constructing interval estimators of θ_π. Then Section 1.5 considers estimating marginal densities associated with π and Section 1.6 further considers stopping rules for MCMC simulations. Finally, in Section 1.7 we give conditions for ensuring the CLT at (1.1.3). The computations presented in our examples were carried out using the R language. See Flegal and Jones (2010c) for an `Sweave` file from which the reader can reproduce all of our calculations.

1.2 Initial Examination of Output

As a first step it pays to examine the empirical finite-sample properties of the Markov chain being simulated. A few simple graphical methods are often used in the initial assessment of the simulation output. These include scatterplots, histograms, time series plots, autocorrelation plots and plots of sample means. We will content ourselves with an illustration of some of these techniques; see Geyer (2010) for further discussion. Consider the following toy example which we will return to several times.

Example 1 (Normal AR(1) Markov chains)**.** *The normal AR(1) time series is defined by*

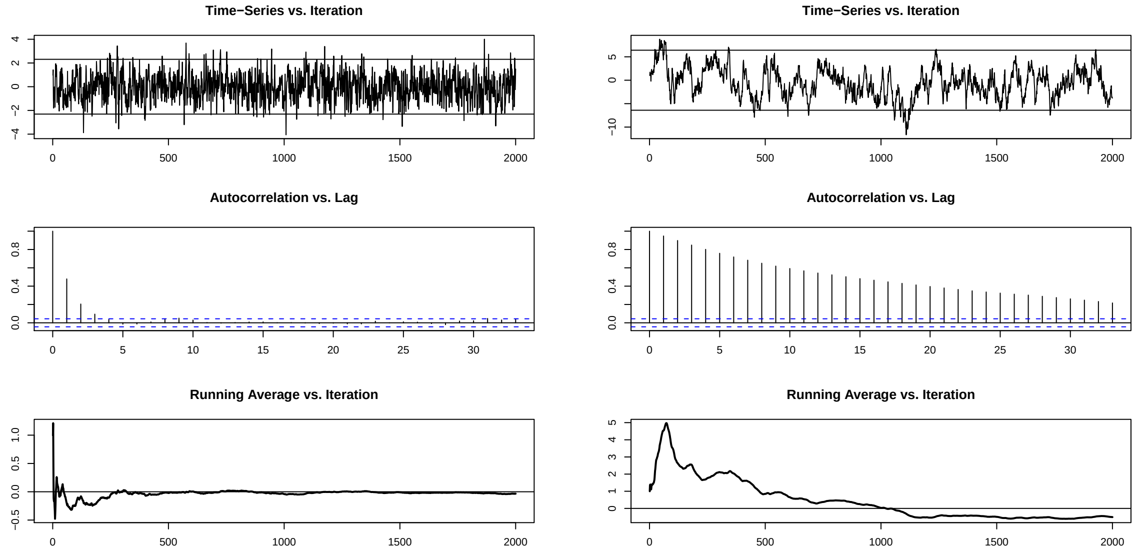
$$X_{n+1} = \rho X_n + \epsilon_n \tag{1.2.1}$$

where the ϵ_n are i.i.d. N(0,1) and $\rho < 1$. This Markov chain has invariant distribution $N(0, 1/(1-\rho^2))$.

As a simple numerical example, consider simulating the chain (1.2.1) *in order to estimate the mean of the invariant distribution, that is $E_\pi X = 0$. While this is a toy example, it is quite useful because ρ plays a crucial role in the behavior of this chain. Figure 1.1 contains plots based on single sample path realizations starting at $X_1 = 1$ with $\rho = 0.5$ and $\rho = 0.95$. In each figure the top plot is a time series plot of the observed sample path. The mean of the target distribution is 0 and the horizontal lines are 2 standard deviations above and below the mean. Comparing the time series plots it is apparent that while we may be getting a representative sample from the invariant distribution, when $\rho = 0.95$ the sample is highly correlated. This is also apparent from the autocorrelation (middle) plots in both figures. When $\rho = 0.5$ the autocorrelation is negligible after about lag 4 but when $\rho = 0.95$ there is a substantial autocorrelation until about lag 30. The impact of this correlation is apparent in the bottom two plots which plot the running estimates of the mean versus iterations in the chain. The true value is displayed as the horizontal line at 0. Clearly, the more correlated sequence requires many more iterations to achieve a reasonable estimate. From these plots, we can see that the simulation with $\rho = 0.5$ may have been run long enough while the simulation with $\rho = 0.95$ likely hasn't.*

In the example, the plots were informative because we were able to draw horizontal lines depicting the true values. In practically relevant MCMC settings—where the truth is unavailable—it is hard to know when to trust these plots. Nevertheless, they can still be useful since a Markov chain that is mixing well would tend to have time series and autocorrelation plots that look like Figure 1.1a while time series and autocorrelation plots like the one in Figure 1.1b would indicate a potentially problematic simulation. Also, plots of current parameter estimates (with no reference to a standard error) versus iteration are not as helpful since they provide little information as to the quality of estimation.

In simulating a d-dimensional Markov chain to simultaneously estimate the p-dimensional vector θ_π of features of π, p can be either greater than or less than d. When either d or p are large, the standard graphical techniques are obviously problematic. That is, even if each component's time series plot indicates good mixing one should not necessarily infer that the



(a) $\rho = 0.5$

(b) $\rho = 0.95$

Figure 1.1: Initial output examination for AR(1) model.

chain has converged to its joint target distribution. In addition, if either d or p is large it will be impractical to look at plots of each component. These issues have received very little attention in the MCMC literature.

1.3 Point Estimates of θ_π

In this section, we consider two specific cases of θ_π, estimating a univariate expectation $E_\pi g$ and estimating a quantile of one of the univariate marginal distributions from π.

1.3.1 Expectations

Suppose $\theta_\pi = E_\pi g$ and assume $E_\pi |g| < \infty$. Recall from Section 1.1 that there is a SLLN and hence it is natural to use the sample average $\bar{g}_n$ to estimate θ_π. Alternatively, we could use point estimates of θ_π obtained through the use of burn-in or averaging conditional expectations.

Consider the use of burn-in. Usually, the simulation is not started with a draw from π since otherwise we would just do ordinary Monte Carlo. It follows that marginally each $X_i \nsim \pi$ and $E_\pi g \neq E[g(X_i)]$. Thus $\bar{g}_n$ is a biased estimator of $E_\pi g$. In the current setting, we have that $X_n \xrightarrow{d} \pi$ as $n \to \infty$ so, in order to diminish the effect of this "initialization bias" an alternative estimator may be employed

$$\bar{g}_{n,B} = \frac{1}{n} \sum_{i=B}^{n+B-1} g(x_i)$$

where B denotes the burn-in or amount of simulation discarded. By keeping only the draws obtained after $B-1$ we are effectively choosing a new initial distribution that is "closer" to π. The SLLN still applies to $\bar{g}_{n,B}$ since if it holds for any initial distribution it holds for every initial distribution. Of course, one possible (maybe even likely) consequence of using burn-in is that $\text{var}(\bar{g}_{n,B}) \geq \text{var}(\bar{g}_{n+B,0})$, that is, the bias decreases but the variance increases for the same total simulation effort. Obviously, this means that using burn-in could result in an estimator having larger mean-squared error than one without burn-in. Moreover, without some potentially difficult theoretical work (Jones and Hobert, 2001; Latuszynski and Niemiro, 2009; Rosenthal, 1995; Rudolf, 2009) it is not clear what value of B should be chosen. Popular approaches to determining B include simply discarding a fraction of the total run length (see e.g. Gelman and Rubin, 1992) or are based on convergence diagnostics (for a review see Cowles and Carlin, 1996). Unfortunately, there simply is no guarantee that any of these diagnostics will detect a problem with the simulation and, in fact, using them can introduce bias (Cowles et al., 1999).

Now consider the estimator obtained by averaging conditional expectations. To motivate this discussion, suppose the target is a function of two variables $\pi(x,y)$ and we are interested in estimating the expectation of a function of only one of the variables, say $g(x)$. Let $(X,Y) = \{(X_0,Y_0), (X_1,Y_1), (X_2,Y_2), \ldots\}$ denote the Markov chain, $m_Y(y)$ denote the marginal density, and $f_{X|Y}(x|y)$ denote the conditional density. Notice

$$E_\pi g = \int \int g(x)\pi(x,y)dxdy = \int \left[\int g(x) f_{X|Y}(x|y)dx \right] m_Y(y)dy$$

so by letting

$$h(y) = \int g(x) f_{X|Y}(x|y)dx$$

we can appeal to the SLLN again to see that as $n \to \infty$

$$\bar{h}_n = \frac{1}{n} \sum_{i=0}^{n-1} h(y_i) = \frac{1}{n} \sum_{i=0}^{n-1} \int g(x) f_{X|Y}(x|y_i)dx \overset{a.s.}{\to} E_\pi g \; .$$

This estimator is conceptually the same as $\bar{g}_n$ in the sense that both are sample averages and the Markov chain SLLN applies to both. The estimator $\bar{h}_n$ is often called the *Rao-Blackwellised* (RB) estimator[3] of $E_\pi g$ (Casella and Robert, 1996). A natural question is which of $\bar{g}_n$, the sample average, or $\bar{h}_n$, the Rao-Blackwellised estimator, is better? It is obvious that $\bar{h}_n$ will sometimes be impossible to use if $f_{X|Y}(x|y)$ is not available in closed form or if the integral is intractable. Hence $\bar{h}_n$ will not be as generally practical as $\bar{g}_n$. However, there are settings, such as in data augmentation (Hobert, 2010) where $\bar{h}_n$ is theoretically and empirically superior to $\bar{g}_n$; see Liu et al. (1994) and Geyer (1995) for theoretical investigation of these two estimators.

Example 2. *This example is also considered in Hobert (2010). Suppose our goal is to estimate the first two moments of a Students t distribution with 4 degrees of freedom and having density*

$$m(x) = \frac{3}{8}\left(1+\frac{x^2}{4}\right)^{-5/2} .$$

There is nothing about this that requires MCMC since we can easily calculate that $E_m X = 0$ and $E_m X^2 = 2$. Nevertheless, we will use a data augmentation algorithm based on the joint density

$$\pi(x,y) = \frac{4}{\sqrt{2\pi}} y^{\frac{3}{2}} e^{-y\left(2+x^2/2\right)}$$

so that the full conditionals are $X|Y \sim N(0, y^{-1})$ and $Y|X \sim Gamma(5/2, 2+x^2/2)$. Consider the Gibbs sampler that updates X then Y so that a one-step transition looks like $(x', y') \to (x, y') \to (x, y)$. Suppose we have obtained n observations $\{x_i, y_i\,;\, i = 0, \ldots, n-1\}$ from running the Gibbs sampler. Then the standard sample average estimates of $E_m X$ and $E_m X^2$ are

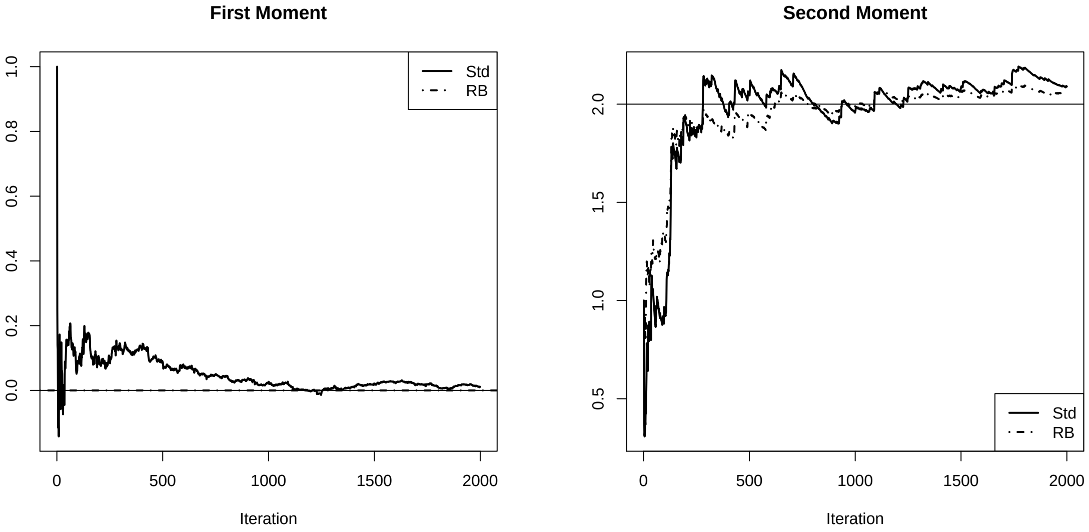
$$\frac{1}{n}\sum_{i=0}^{n-1} x_i \qquad \textit{and} \qquad \frac{1}{n}\sum_{i=0}^{n-1} x_i^2\ , \ \textit{respectively.}$$

Further, the Rao-Blackwellised estimates are easily computed. Since $X|Y \sim N(0, y^{-1})$ the Rao-Blackwellised estimate of $E_m X$ is 0! On the other hand, the RB estimate of $E_m X^2$ is

$$\frac{1}{n}\sum_{i=0}^{n-1} y_i^{-1} .$$

As an illustration of these estimators we simulated 2000 iterations of the Gibbs sampler and

[3] This as an unfortunate name since it is only indirectly related to the Rao-Blackwell Theorem, but the name has stuck in the literature.



(a) $E_m X$ (b) $E_m X^2$

Figure 1.2: Estimators of the first two moments for Example 2. The horizontal line denotes the truth, the solid curves are the running sample averages while the dotted curves are the running RB sample averages.

plotted the running values of the estimators in Figure 1.2. In this example, the RB averages are less variable than the standard sample averages.

It is not the case that RB estimators are always better than sample means. Whether they are better depends on the expectation being estimated as well as the properties of the MCMC sampler. In fact, Liu et al. (1994) and Geyer (1995) give an example where the RB estimator is provably worse than the sample average. RB estimators are more general than our presentation suggests. Let h be any function and set

$$f(x) = E[g(X)|h(X) = h(x)]$$

so that $E_\pi g = E_\pi f$. Thus, by the Markov chain SLLN with probability 1 as $n \to \infty$,

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) = \frac{1}{n}\sum_{i=1}^{n} E[g(X)|h(X_i) = h(x_i)] \to E_\pi g \ .$$

As long as the conditional distribution $X|h(x)$ is tractable, RB estimators are available.

1.3.2 Quantiles

It is common to report estimated quantiles in addition to estimated expectations. Actually, what is nearly always reported is not a multivariate quantile but rather quantiles of the univariate marginal distributions associated with π. This is the only setting we consider. Let F be a marginal distribution function associated with π. Then the qth quantile of F is

$$\phi_q = F^{-1}(q) = \inf\{x : F(x) \geq q\}, \quad 0 < q < 1 \; . \tag{1.3.1}$$

There are many potential estimates of ϕ_q but we consider only the inverse of the empirical distribution function from the observed sequence. First define $\{X_{(1)}, \ldots, X_{(n)}\}$ as the order statistics of $\{X_0, \ldots, X_{n-1}\}$, then the estimator of ϕ_q is given by

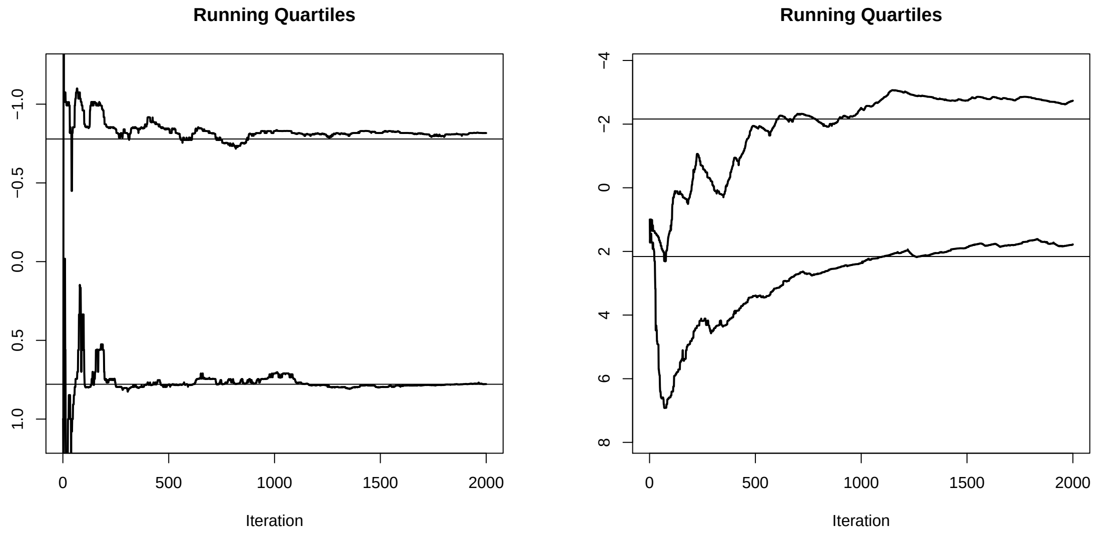
$$\hat{\phi}_{q,n} = X_{(j+1)} \text{ where } \frac{j}{n} \leq q < \frac{j+1}{n} \; . \tag{1.3.2}$$

Example 3 (Normal AR(1) Markov chains)**.** *Consider again the time series defined at* (1.2.1)*. Our goal in this example is to illustrate estimating the first and third quartiles, denoted Q_1 and Q_3. The true values of Q_1 and Q_3 are $\pm\Phi^{-1}(0.75)/\sqrt{1-\rho^2}$ where Φ is the cumulative distribution function of a standard normal distribution.*

Using the same realization of the chain as in Example 1, Figures 1.3a and 1.3b show plots of the running quartiles versus iteration when $\rho = 0.5$ and $\rho = 0.95$ respectively. It is immediately apparent that estimation is more difficult when $\rho = 0.95$ and hence the simulation should continue. Also, without the horizontal lines, these plots would not be as useful. Recall a similar conclusion was reached for estimating the mean.

1.4 Interval Estimates of θ_π

In our examples we have known the truth, enabling us to draw the horizontal lines on the plots which allow us to gauge the quality of estimation. Obviously, the true parameter value is unknown in practical settings and hence the size of the Monte Carlo error is unknown. For this reason, when reporting a point estimate of θ_π, a Monte Carlo standard error should be included so that the reader can assess the quality of the reported point estimates. In this section we address how to calculate MCSEs and construct asymptotically valid interval estimates of θ_π.



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

Figure 1.3: Plots for AR(1) model of running estimates of Q_1 and Q_3. The horizontal lines are the true quartiles.

1.4.1 Expectations

Suppose $\theta_\pi = E_\pi g$ which will be estimated with $\bar{g}_n = \bar{g}_{n,0}$, that is with no burn-in. However, using burn-in presents no theoretical difficulties since, as with the SLLN, if the CLT holds for any initial distribution then it holds for every initial distribution. Thus the use of burn-in does not affect the existence of a CLT, but it may affect the quality of the asymptotic approximation. If $\hat{\sigma}_n^2$ is an estimate of σ_g^2, then one can form a confidence interval for $E_\pi g$ with half-width

$$t_* \frac{\hat{\sigma}_n}{\sqrt{n}} \tag{1.4.1}$$

where t_* is an appropriate quantile. Thus the difficulty in finding interval estimates is in estimating σ_g^2 which requires specialized techniques to account for correlation in the Markov chain. We restrict attention to strongly consistent estimators of σ_g^2. Some interval estimation techniques do not require consistent estimation of σ_g^2 (see Schruben, 1983) but we need it for the methods presented later in Section 1.6. The methods yielding strongly consistent estimators include batch means methods, spectral variance methods and regenerative simulation. Alternatives include the initial sequence methods of Geyer (1992), however the theoretical properties of Geyer's estimators are not well understood. We will focus on batch means as it is the most generally applicable method; for more on spectral methods see Flegal and Jones (2010a) while Hobert et al. (2002) and Mykland et al. (1995) study regenerative simulation. There are many variants of batch means, here we emphasize overlapping batch means (OLBM).

Overlapping Batch Means

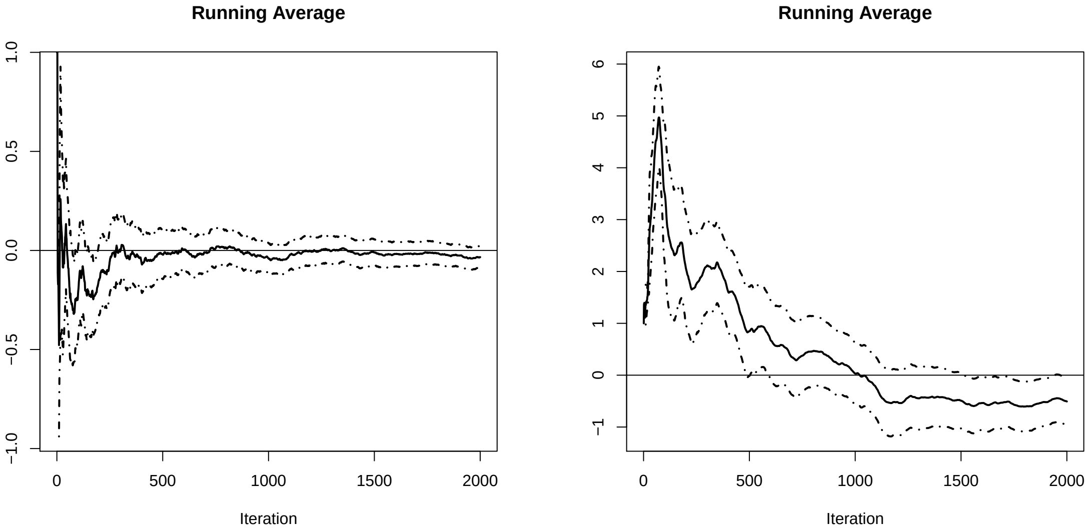
As the name suggests, in OLBM we divide the simulation into overlapping batches of length b_n say. For example, if $b_n = 3$, then $\{X_0, X_1, X_2\}$ and $\{X_1, X_2, X_3\}$ would be the first two overlapping batches. In general, there are $n - b_n + 1$ batches of length b_n, indexed by j running from 0 to $n - b_n$. Let $\bar{Y}_j(b_n) := b_n^{-1} \sum_{i=0}^{b_n-1} g(X_{j+i})$ for $j = 0, \ldots, n - b_n$. Then the OLBM estimator of σ_g^2 is

$$\hat{\sigma}^2_{OLBM} = \frac{n b_n}{(n - b_n)(n - b_n + 1)} \sum_{j=0}^{n-b_n} (\bar{Y}_j(b_n) - \bar{g}_n)^2 \; . \tag{1.4.2}$$

Batch means estimators are not generally consistent for σ_g^2 (Glynn and Whitt, 1991). However, roughly speaking, Flegal and Jones (2010a) show that if the Markov chain mixes quickly and b_n is allowed to increase as the overall length of the simulation does, then $\hat{\sigma}^2_{OLBM}$ is a strongly consistent estimator of σ_g^2. It is often convenient to take $b_n = \lfloor n^\nu \rfloor$ for some $0 < \nu < 3/4$, and $\nu = 1/2$ may be a reasonable default. However, ν values yielding strongly consistent estimators are dependent on the number of finite moments of g with respect to the target π and the mixing conditions of the Markov chain. These conditions are similar to those required for a Markov chain CLT, see Flegal and Jones (2010a); Jones (2004); and Jones et al. (2006). Finally, when constructing the interval (1.4.1), t_* is a quantile from a Student's t distribution with $n - b_n$ degrees of freedom.

Example 4 (Normal AR(1) Markov chains). *Recall the AR(1) model defined at* (1.2.1). *Using the same realization of the chain as in Example 1, that is, 2000 iterations with* $\rho \in \{0.5, 0.95\}$ *starting from* $X_1 = 1$, *we consider estimating the mean of the invariant distribution, i.e.* $E_\pi X = 0$. *Utilizing OLBM with* $b_n = \lfloor n^{1/2} \rfloor$, *we calculated an MCSE and resulting 80% confidence interval. Figures 1.4a and 1.4b show the running means versus iteration for* $\rho = 0.5$ *and* $\rho = 0.95$ *respectively. The dashed lines correspond to upper and lower 80% confidence bounds. Notice that for the larger value of* ρ *it takes longer for the MCSE to stabilize and begin decreasing. After 2000 iterations for* $\rho = 0.5$ *we obtained an interval of* -0.034 ± 0.056 *while for* $\rho = 0.95$ *the interval is* -0.507 ± 0.451.

Many of our plots are based on simulating only 2000 iterations. We chose this value



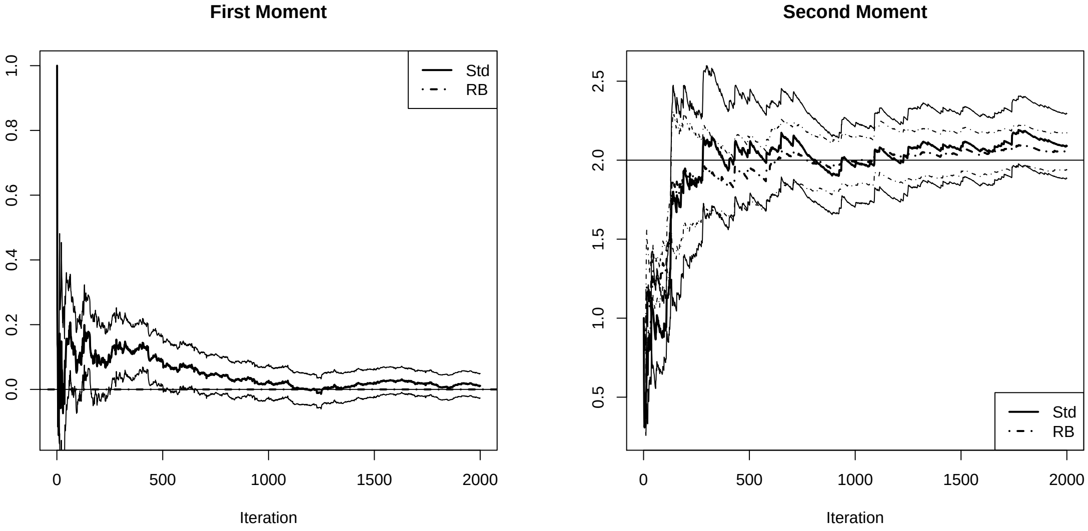
(a) $\rho = 0.5$ (b) $\rho = 0.95$

Figure 1.4: Plots for AR(1) model of running estimates of the mean along with confidence intervals calculated via OLBM. the horizontal line denotes the truth.

strictly for illustration purposes. An obvious question is has the simulation been run long enough? That is, are the interval estimates sufficiently narrow after 2000 iterations? In the $\rho = .5$ case, the answer is 'maybe' while in the $\rho = .95$ case it is clearly 'no'. Consider the final interval estimate of the mean with $\rho = 0.5$, that is, $-0.034 \pm 0.056 = (-0.090, 0.022)$. If the user is satisfied with this level of precision, then 2000 iterations is sufficient. On the other hand, when $\rho = .95$ our interval estimate is $-0.507 \pm 0.451 = (-0.958, -0.056)$, indicating we cannot trust any of the significant figures reported in the point estimate.

Recall the RB estimators of Section 1.3.1. It is straightforward to use OLBM to calculate the MCSE for these estimators since the conditional expectations being averaged define the sequence of batch means $\bar{Y}_j(b_n)$.

Example 5. *Recall Example 2 where the first two moments of a Students t distribution with 4 degrees of freedom were estimated using sample average and RB estimators. Using the same Markov chain, an 80% CI is calculated via OLBM with $b_n = \lfloor n^{1/2} \rfloor$ at each iteration. Figure 1.5a shows the running estimate of $E_\pi X$ versus iteration and includes confidence bounds for the sample average estimator. Recall, the RB estimator is exact so there is no uncertainty in this estimate. Figure 1.5b shows the running estimate of $E_\pi X$ versus iteration with confidence bounds for both estimates. Here it is provable that the RB estimator has a*



(a) $E_m X$

(b) $E_m X^2$

Figure 1.5: Estimators of the first two moments from a t distribution with 4 degrees of freedom. The horizontal line denotes the truth, the solid curves are the running sample averages with confidence bounds while the dotted curves are the running RB sample averages with confidence bounds.

smaller asymptotic variance than the sample average estimator (Geyer, 1995). This is clearly reflected by the narrower interval estimates.

Parallel Chains

To this point, our recipe seems straightforward: given a sampler, pick a starting value and run the simulation for a sufficiently long time using the SLLN and the CLT to produce a point estimate and a measure of its uncertainty. A variation of this procedure relies on simulating multiple independent, or parallel, chains. Debate between a single long run and parallel chains began in the early statistics literature on MCMC (see e.g. Gelman and Rubin, 1992; Geyer, 1992), even earlier in the operations research and physics literature (see e.g. Bratley et al., 1987; Fosdick, 1959; Kelton and Law, 1984), and continues today (Alexopoulos et al., 2006; Alexopoulos and Goldsman, 2004). The main idea of parallel chains is to run r independent chains using different starting values where each chain is the same length using the same burn-in. This yields r independent estimates of $E_\pi g$, namely $\bar{g}_{n,B,1}, \bar{g}_{n,B,2}, \ldots, \bar{g}_{n,B,r}$. The grand mean would then estimate $E_\pi g$—although Glynn and Heidelberger (1991) have shown that an alternative estimator may be superior—and our estimate of its performance, i.e. σ_g^2, would be the usual sample variance of the $\bar{g}_{n,B,i}$.

This approach has some intuitive appeal in that estimation avoids some of the serial correlation inherent in MCMC and it is easily implemented when more than one processor is available. Moreover, there is value in trying a variety of initial values for any MCMC experiment. It has also been argued that by choosing the r starting points in a widely dispersed manner there is a greater chance of encountering modes that one long run may have missed. Thus, for example, some argue that using independent replications results in "superior inferential validity" (Gelman and Rubin, 1992, p. 503). However, there is no agreement on this issue, indeed, Bratley et al. (1987, p. 80) "are skeptical about [the] rationale" of some proponents of independent replications. Notice that the total simulation effort using independent replications is $r(n+B)$. To obtain good estimates of σ_g^2 will require r to be large which will require $n+B$ to be small for a given computational effort. If we use the same value of B as we would when using one long run this means that each $\bar{g}_{n,B,i}$ will be based on a comparatively small number n of observations. Using more than one chain will also enhance the initialization bias so that a careful choice of B can be quite important to the statistical efficiency of the estimator of $E_\pi g$ (Glynn and Heidelberger, 1991). Moreover, since each run will be comparatively short there is a reasonable chance that a given replication will not move far from its starting value. Alexopoulos and Goldsman (2004) have shown that this can result in much poorer estimates (in terms of mean square error) of $E_\pi g$ than a single long run. On the other hand, if we can find a variety of starting values that are from a distribution very close to π, then independent replications may indeed be superior. This should not be surprising since independent draws directly from π are clearly desirable.

There is an important caveat to the above analysis. There are settings (see e.g. Caffo et al., 2010) where it is prohibitively difficult (or time-consuming) to produce a sufficiently large Monte Carlo sample without parallel computing. This has received limited attention in MCMC settings (Brockwell, 2006; Rosenthal, 2000) but perhaps deserves more.

1.4.2 Functions of Moments

Suppose we are interested in estimating $\phi(E_\pi g)$ where ϕ is some function. If ϕ is continuous, then $\phi(\bar{g}_n) \to \phi(E_\pi g)$ with probability 1 as $n \to \infty$ making estimation of $\phi(E_\pi g)$ straightforward. Also, a valid Monte Carlo error can be obtained via the delta method (Ferguson, 1996; van der Vaart, 1998). Assuming (1.1.3) the delta method says that if ϕ is continuously differentiable in a neighborhood of $E_\pi g$ and $\phi'(E_\pi g) \neq 0$, then as $n \to \infty$

$$\sqrt{n}(\phi(\bar{g}_n) - \phi(E_\pi g)) \xrightarrow{d} \mathrm{N}(0, [\phi'(E_\pi g)]^2 \sigma_g^2) .$$

Thus if our estimator of σ_g^2, say $\hat{\sigma}_n^2$ is strongly consistent, then $[\phi'(\bar{g}_n)]^2 \hat{\sigma}_n^2$ is strongly consistent for $[\phi'(E_\pi g)]^2 \sigma_g^2$.

Example 6. *Consider estimating $(E_\pi X)^2$ with $(\bar{X}_n)^2$. Let $\phi(x) = x^2$ and assume $E_\pi X \neq 0$ and a CLT as in* (1.1.3)*. Then as $n \to \infty$*

$$\sqrt{n}\left((\bar{X}_n)^2 - (E_\pi X)^2\right) \xrightarrow{d} N(0, 4(E_\pi X)^2 \sigma_x^2)$$

and we can use OLBM to consistently estimate σ_x^2 with $\hat{\sigma}_n^2$ which means that $4(\bar{X}_n)^2 \hat{\sigma}_n^2$ is a strongly consistent estimator of $4(E_\pi X)^2 \sigma_x^2$.

From this example we see that the univariate delta method makes it straightforward to handle powers of moments. The multivariate delta method allows us to handle more complicated functions of moments. Let T_n denote a sequence of d-dimensional random vectors and θ be a d-dimensional parameter. If as $n \to \infty$,

$$\sqrt{n}(T_n - \theta) \xrightarrow{d} \mathrm{N}(\mu, \Sigma)$$

and ϕ is continuously differentiable in a neighborhood of θ and $\phi'(\theta) \neq 0$, then as $n \to \infty$

$$\sqrt{n}(\phi(T_n) - \phi(\theta)) \xrightarrow{d} \mathrm{N}(\phi'(\theta)\mu, \phi'(\theta)\Sigma\phi'(\theta)^T) .$$

Example 7. *Consider estimating $var_\pi g = E_\pi g^2 - (E_\pi g)^2$ with, setting $h = g^2$,*

$$\frac{1}{n}\sum_{i=1}^{n} h(X_i) - \left(\frac{1}{n}\sum_{i=1}^{n} g(X_i)\right)^2 := \hat{v}_n .$$

Assume

$$\sqrt{n}\left(\begin{pmatrix} \bar{g}_n \\ \bar{h}_n \end{pmatrix} - \begin{pmatrix} E_\pi g \\ E_\pi g^2 \end{pmatrix}\right) \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_g^2 & c \\ c & \sigma_h^2 \end{pmatrix}\right)$$

where $c = E_\pi g^3 - E_\pi g E_\pi g^2$. Let $\phi(x, y) = y - x^2$. Then as $n \to \infty$

$$\sqrt{n}(\hat{v}_n - var_\pi g) \xrightarrow{d} N(0, 4(E_\pi g)(\sigma_g^2 E_\pi g - E_\pi g^3 + E_\pi g E_\pi g^2) + \sigma_h^2) .$$

Since it is easy to use OLBM to construct strongly consistent estimators of σ_g^2 and σ_h^2 the following formula gives a strongly consistent estimator of the variance in the asymptotic normal distribution for $\hat{v}_n$

$$4(\bar{g}_n)(\hat{\sigma}^2_{g,n}\,\bar{g}_n - \bar{j}_n + \bar{g}_n\bar{h}_n) + \hat{\sigma}^2_{h,n}$$

where $j = g^3$.

1.4.3 Quantiles

Suppose our goal is to estimate ϕ_q with $\hat{\phi}_{q,n}$ defined at (1.3.1) and (1.3.2), respectively. We now turn our attention to constructing an interval estimate of ϕ_q. It is tempting to think that bootstrap methods would be appropriate for this problem. Indeed there has been a substantial amount of research into bootstrap methods for stationary time series which would be appropriate for MCMC settings (see e.g. Bertail and Clémençon, 2006; Bühlmann, 2002; Datta and McCormick, 1993; Politis, 2003). Unfortunately, our experience has been that these methods are *extremely* computationally intensive and have inferior finite-sample properties compared to the method presented below.

As above, we assume the existence of an asymptotic normal distribution for the Monte Carlo error, that is, there is a constant $\gamma_q^2 \in (0, \infty)$ such that as $n \to \infty$

$$\sqrt{n}(\hat{\phi}_{q,n} - \phi_q) \xrightarrow{d} \mathrm{N}(0, \gamma_q^2)\ . \tag{1.4.3}$$

Flegal and Jones (2010b) give conditions under which (1.4.3) obtains. Just as when we were estimating an expectation, we find ourselves in the position of estimating a complicated constant γ_q^2. We focus on the use of the subsampling bootstrap method (SBM) in this context. The reader should be aware that our use of the term "subsampling" is quite different than the way it is often used in the context of MCMC in that we are not deleting any observations of the Markov chain.

Subsampling Bootstrap

This section will provide a brief overview of SBM in the context of MCMC and illustrate its use for calculating the MCSE of $\hat{\phi}_{q,n}$. While this section focuses on quantiles, SBM methods apply much more generally; the interested reader is encouraged to consult Politis et al. (1999).

The main idea for SBM is similar to OLBM in that we are taking overlapping batches (or subsamples) of size b_n from the first n observations of the chain $\{X_0, X_1, \ldots, X_{n-1}\}$. There are $n - b_n + 1$ such subsamples. Let $\{X_i, \ldots, X_{i+b_n-1}\}$ be the ith subsample with corresponding ordered subsample $\{X^*_{(1)}, \ldots, X^*_{(b_n)}\}$. Then define the quantile based on the ith subsample as

$$\phi^*_i = X^*_{(j+1)} \text{ where } \frac{j}{b_n} \le q < \frac{j+1}{b_n} \text{ for } i = 0, \ldots, n - b_n \; .$$

The SBM estimate of γ^2_q is then

$$\hat{\gamma}^2_q = \frac{b_n}{n - b_n + 1} \sum_{i=0}^{n-b_n+1} (\phi^*_i - \bar{\phi}^*)^2 \; , \tag{1.4.4}$$

where

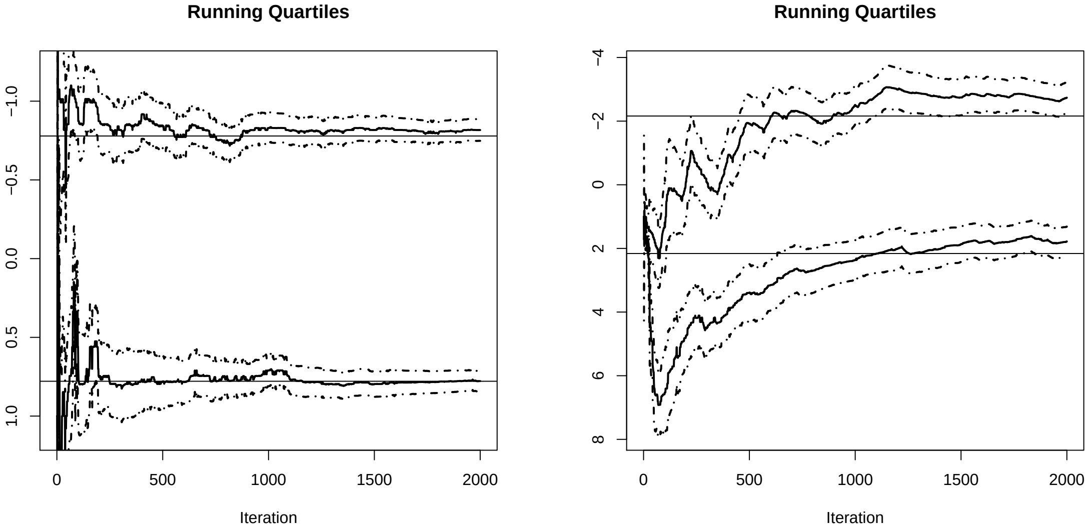
$$\bar{\phi}^* = \frac{1}{n - b_n + 1} \sum_{i=0}^{n-b_n+1} \phi^*_i \; .$$

Politis et al. (1999) give conditions that ensure this estimator is strongly consistent but their conditions could be difficult to check in practice. SBM implementation requires choosing b_n such that as $n \to \infty$ we have $b_n \to \infty$ and $b_n/n \to 0$. A natural choice is $b_n = \lfloor \sqrt{n} \rfloor$.

Example 8 (Normal AR(1) Markov chains)**.** *Using the AR(1) model defined at* (1.2.1) *we again consider estimating the first and third quartiles, denoted* Q_1 *and* Q_3*. Recall that the true values for the quartiles are* $\pm\Phi^{-1}(0.75)/\sqrt{1-\rho^2}$*, respectively.*

Figure 1.6 shows the output from the same realization of the chain used previously in Example 3 but this time the plot includes an interval estimate of the quartiles. Figure 1.6a shows a plot of the running quartiles versus iteration when $\rho = 0.5$*. In addition, the dashed lines show the 80% confidence interval bounds at each iteration. These intervals were produced with SBM using* $b_n = \lfloor \sqrt{n} \rfloor$*. At around 200 iterations, the MCSE (and hence interval estimates) seem to stabilize and begin to decrease. At 2000 iterations, the estimates for* Q_1 *and* Q_3 *are* -0.817 ± 0.069 *and* 0.778 ± 0.065 *respectively. Figure 1.6b shows the same plot when* $\rho = 0.95$*. At 2000 iterations, the estimates for* Q_1 *and* Q_3 *are* -2.74 ± 0.481 *and* 1.78 ± 0.466 *respectively.*

Are the intervals sufficiently narrow after 2000 iterations? In both cases ($\rho = 0.5$ *and* $\rho = .95$*) the answer is likely 'no'. Consider the narrowest interval which is the one for*



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

Figure 1.6: Plots for AR(1) model of running estimates of Q_1 and Q_3 along with 80% pointwise confidence intervals calculated via SBM. The horizontal lines denote the true values.

Q_3 with $\rho = .5$, that is, $0.778 \pm 0.065 = (0.713, 0.843)$ which indicates that we can at most trust the sign and the leading 0 of the estimate, 0.778. On the other hand, certainly this is evidence that the true quantile is between .71 and .85. Note that in a real problem we would not have the horizontal line in the plot depicting the truth.

SBM is applicable much more generally than presented here and, in fact, essentially generalizes the method of overlapping batch means previously discussed in the context of estimating an expectation. The subsample mean is $\bar{Y}_j(b_n)$ and the resulting estimate of σ_g^2 is

$$\hat{\sigma}^2_{SBM} = \frac{b_n}{n - b_n + 1} \sum_{j=0}^{n-b_n} (\bar{Y}_j(b_n) - \bar{Y}^*)^2 \tag{1.4.5}$$

where

$$\bar{Y}^* = \frac{1}{n - b_n + 1} \sum_{i=0}^{n-b_n+1} \bar{Y}_j(b_n) \, .$$

It is straightforward to establish that the OLBM estimate defined in (1.4.2) is asymptotically equivalent to the SBM estimate defined at (1.4.5).

1.4.4 Multivariate Estimation

While we have largely focused on the univariate setting, recall from Section 1.1 that a typical MCMC experiment is conducted with the goal of estimating a p-dimensional vector of parameters, θ_π, associated with the d-dimensional target π. Generally, θ_π could be composed of expectations, quantiles and so on and p could be either much larger or much smaller than d. Suppose each component $\theta_{\pi,i}$ can be estimated with $\hat{\theta}_{n,i}$ so that $\hat{\theta}_n = (\hat{\theta}_{n,1}, \dots, \hat{\theta}_{n,1}) \to \theta_\pi$ almost surely as $n \to \infty$. It is natural to seek to establish the existence of an asymptotic distribution of the Monte Carlo error $\hat{\theta}_n - \theta_\pi$ and then use this distribution to construct asymptotically valid confidence regions. To our knowledge this problem has not been investigated. However, it has received some attention in the case where θ_π consists only of expectations; we know of one paper in the statistics literature (Kosorok, 2000) and a few more in operations research including Muñoz and Glynn (2001), Seila (1982), and Yang and Nelson (1992). Currently, the most common approach is to ignore the multiplicity issue and simply construct the MCSE for each component of the Monte Carlo error. If p isn't too large then a Bonferroni correction could be used but this is clearly less than optimal. This is obviously an area in MCMC output analysis that could benefit from further research.

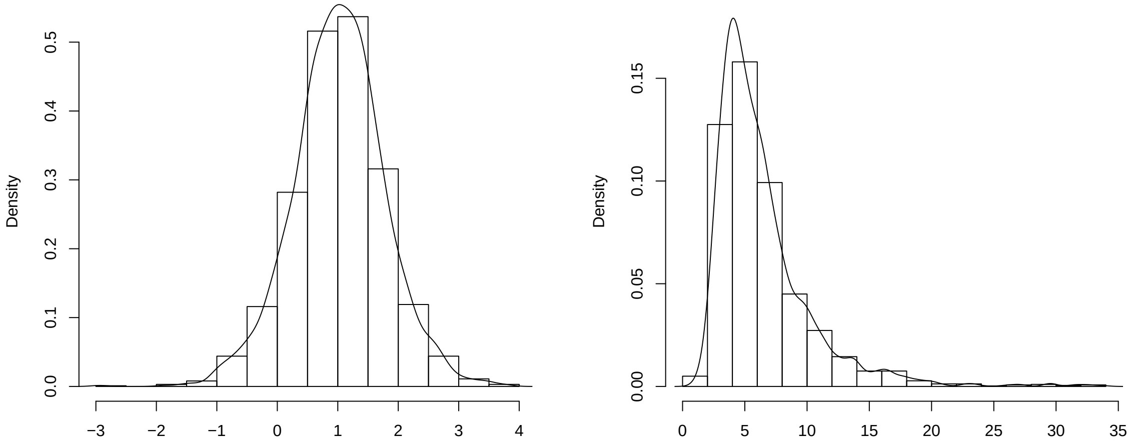
1.5 Estimating Marginal Densities

A common inferential goal is the production of a plot of a marginal density associated with π. In this section we cover two methods for doing this. We begin with a simple graphical method then introduce a clever method due to Wei and Tanner (1990) that reminds us of the Rao-Blackwellisation methods of Section 1.3.

A histogram approximates the true marginal by the Markov chain SLLN. Moreover, because histograms are so easy to construct with existing software they are popular. Another common approach is to report a nonparametric density estimate or smoothed histogram. It is conceptually straightforward to construct pointwise interval estimates for the smoothed histogram using SBM. However, outside of toy examples, the computational cost is typically prohibitive.

Example 9. *Suppose $Y_i|\mu, \theta \sim N(\mu, \theta)$ independently for $i = 1, \dots, m$ where $m \geq 3$ and assume the standard invariant prior $\nu(\mu, \theta) \propto \theta^{-1/2}$. The resulting posterior density is*

$$\pi(\mu, \theta|y) \propto \theta^{-(m+1)/2} e^{-\frac{m}{2\theta}(s^2 + (\bar{y} - \mu)^2)}$$



(a) Marginal density of μ (b) Marginal density of θ

Figure 1.7: Histograms for estimating the marginal densities of Example 9.

where s^2 is the usual biased sample variance. It is easy to see that $\mu|\theta, y \sim N(\bar{y}, \theta/m)$ and that $\theta|\mu, y \sim IG((m-1)/2, m[s^2 + (\bar{y} - \mu)^2]/2)$, and hence a Gibbs sampler is appropriate. We consider the Gibbs sampler that updates μ then θ so that a one-step transition is given by $(\mu', \theta') \to (\mu, \theta') \to (\mu, \theta)$ and use this sampler to estimate the marginal densities of μ and θ.

Now suppose $m = 11$, $\bar{y} = 1$ and $s^2 = 4$. We simulated 2000 *realizations of the Gibbs sampler starting from $(\mu_0, \lambda_0) = (1, 1)$. The marginal density plots were created using the default settings for the* `density` *function in R and are shown in Figure 1.7 while an estimated bivariate density plot (created using R functions* `kde2d` *and* `persp`*) is given in Figure 1.8. It is obvious from these figures that the posterior is simple, so it is no surprise that the Gibbs sampler has been shown to converge in just a few iterations (Jones and Hobert, 2001).*

A clever technique for estimating a marginal is based on the same idea as RB estimators (Wei and Tanner, 1990). To keep the notation simple suppose the target is a function of only two variables, $\pi(x, y)$ and let m_X and m_Y be the associated marginals. Then

$$m_X(x) = \int \pi(x, y)dy = \int f_{X|Y}(x|y)m_Y(y)dy = E_{m_Y} f_{X|Y}(x|y)$$

suggesting that, by the Markov chain SLLN, we can get a functional approximation to m_X

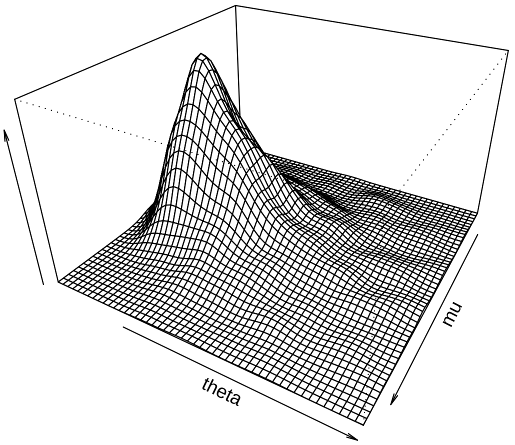


Figure 1.8: Estimated posterior density of Example 9.

since for each x as $n \to \infty$

$$\frac{1}{n}\sum_{i=0}^{n-1} f_{X|Y}(x|y_i) \to m_X(x) \; . \tag{1.5.1}$$

Of course, just as with RB estimators this will only be useful when the conditionals are tractable. Note also, that it is straightforward to use OLBM to get pointwise confidence intervals for the resulting curve; that is for each x we can calculate an MCSE of the sample average in (1.5.1).

Example 10. *Recall the setting of Example 9. We will focus on estimation of the marginal posterior density of $\mu|y$, i.e. $\pi(\mu|y)$. Note that*

$$\pi(\mu|y) = \int \pi(\mu|\theta, y)\pi(\theta|y)\, d\theta$$

so that by the Markov chain SLLN we can estimate $\pi(\mu|y)$ with

$$\frac{1}{n}\sum_{i=0}^{n-1} \pi(\mu|\theta_i, y)$$

which is straightforward to evaluate since $\mu|\theta_i, y \sim N(\bar{y}, \theta_i/m)$. Note that the resulting marginal estimate is a linear combination of normal densities. Using the same realization

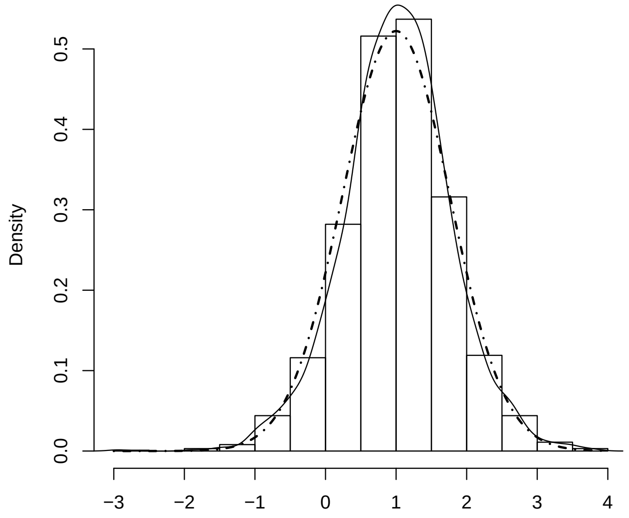


Figure 1.9: Estimates of the marginal density μ. The three estimates are based a histogram, smoothed marginal densities (solid line), and the method of Wei and Tanner (1990) (dashed line).

of the chain from Example 9 we estimated $\pi(\mu|y)$ *using this method. Figure 1.9 shows the results with our previous estimates. One can also calculate pointwise confidence intervals using OLBM which results in a very small Monte Carlo error (and is therefore not included in the plot). Notice the estimate based on* (1.5.1) *is a bit smoother than either the histogram or the smoothed histogram estimate but is otherwise quite similar.*

1.6 Terminating the Simulation

A common approach to stopping an MCMC experiment is to simulate for a fixed run length. That is, the simulation is terminated using a *fixed-time* rule. Notice that this makes Monte Carlo standard errors crucial to understanding the reliability of the resulting estimates. There are settings where, due to the nearly prohibitive difficulty of the simulation, a fixed-time rule may be the only practical approach. However, this is not the case for many MCMC experiments.

Perhaps the most popular approach to terminating the simulation is to simulate an initial Monte Carlo sample of size, say n_0. The output is examined and if the results are found to be unsatisfactory, the simulation is continued for another n_1 steps and the output reanalyzed. If the results are still unsatisfactory, the process is repeated. Notice that this is a sequential procedure that will result in a random total simulation effort.

When implementing this sequential procedure the examination of the output can take many forms; it is often based on the use of graphical methods such as those described in Section 1.2 or convergence diagnostics. We advocate terminating the simulation the first time the MCSE is sufficiently small. Equivalently, the simulation is terminated the first time the half-width of a confidence interval for θ_π is sufficiently small, resulting in a *fixed-width* rule. There is a substantial amount of research on fixed-width procedures in MCMC when θ_π is an expectation; see Flegal et al. (2008), Glynn and Whitt (1992) and Jones et al. (2006) and the references therein, but none that we are aware of when θ_π is not an expectation. Let $\hat{\sigma}_n^2$ be a strongly consistent estimator of σ_g^2 from (1.1.3). Given a desired half-width ϵ the simulation terminates the first time

$$t_* \frac{\hat{\sigma}_n}{\sqrt{n}} + p(n) \le \epsilon \tag{1.6.1}$$

where t_* is the appropriate quantile and $p(n)$ is a positive function such that $p(n) = o(n^{-1/2})$ as $n \to \infty$. Letting n^* be the desired minimum simulation effort, a reasonable default is $p(n) = \epsilon I(n \le n^*) + n^{-1}$. Glynn and Whitt (1992) established conditions ensuring the interval at (1.6.1) is asymptotically valid[4] in the sense that the desired coverage probability is obtained as $\epsilon \to 0$. The use of these intervals in MCMC settings has been extensively investigated and found to work well by Flegal et al. (2008), Flegal and Jones (2010a) and Jones et al. (2006).

Example 11 (Normal AR(1) Markov chains)**.** *Consider the normal AR(1) time series defined at* (1.2.1)*. In Example 4 we simulated 2000 iterations from the chain with* $\rho = 0.95$ *starting from* $X_0 = 1$ *and found that a 80% confidence interval for the mean of the invariant distribution was* -0.507 ± 0.451*.*

Suppose we wanted to continue our simulation until we were 80% confident our estimate was within .1 of the true value after a minimum of 1000 iterations–a fixed-width procedure. If we use OLBM to estimate the variance in the asymptotic distribution, then (1.6.1) *becomes*

$$t_* \frac{\hat{\sigma}_n}{\sqrt{n}} + 0.1 I(n \le 1000) + n^{-1} \le 0.1 \; .$$

[4] Glynn and Whitt (1992) also provide a counterexample to show that weak consistency of $\hat{\sigma}_n^2$ for σ_g^2 is not enough to achieve asymptotic validity.

where t_* is the appropriate quantile from a Student's t distribution with $n - b_n$ degrees of freedom. It would be computationally expensive to check this criterion after each iteration so instead we added 1000 iterations before recalculating the half-width each time. In this case, the simulation terminated after 6e4 iterations resulting in an interval estimate of -0.0442 ± 0.100. *Notice that this simple example required a relatively large simulation effort compared to what is often done in much more complicated settings but note that ρ is large. Further, either narrowing the interval or increasing the desired confidence level will require a larger simulation effort.*

1.7 Markov Chain Central Limit Theorems

Throughout we have assumed the existence of a Markov chain central limit theorem, see e.g. (1.1.3) and (1.4.3). In this section we provide a brief discussion of the conditions required for these claims; the reader can find much more detail in Chan and Geyer (1994), Jones (2004), Meyn and Tweedie (1993), Roberts and Rosenthal (2004) and Tierney (1994).

Implicitly, we assumed that the Markov chain X is *Harris ergodic*, that is, Harris recurrent and aperiodic. To fully explain these conditions would require a fair amount of Markov chain theory so we will content ourselves with providing references; the interested reader should consult Meyn and Tweedie (1993), Nummelin (1984) or Roberts and Rosenthal (2004). However, it is frequently trivial to verify Harris ergodicity (see e.g. Hobert, 2010; Tan and Hobert, 2009; Tierney, 1994).

Harris ergodicity alone is not sufficient for the Markov chain SLLN or a CLT. However, if X is Harris ergodic and $E_\pi|g| < \infty$, then the SLLN holds: $\bar{g}_n \to E_\pi g$ with probability 1 as $n \to \infty$. A CLT requires stronger conditions. In fact, it is important to be aware that there are simple non-pathological examples of Harris ergodic Markov chains which do not enjoy a CLT (Roberts, 1999). Let the conditional distribution of X_n given $X_0 = x$ be denoted $P^n(x, \cdot)$, that is,

$$P^n(x, A) = \Pr(X_n \in A \mid X_0 = x) .$$

Then Harris ergodicity implies that for every starting point $x \in \mathsf{X}$

$$\|P^n(x, \cdot) - \pi(\cdot)\| \downarrow 0 \quad \text{as} \quad n \to \infty \tag{1.7.1}$$

where $\| \cdot \|$ is the total variation norm. We will need to know the rate of the convergence in

(1.7.1) to say something about the existence of a CLT. Let $M(x)$ be a nonnegative function on X and $\gamma(n)$ be a nonnegative function on $\mathbb{Z}_+$ such that

$$\|P^n(x, \cdot) - \pi(\cdot)\| \le M(x)\gamma(n) \; . \tag{1.7.2}$$

When X is *geometrically ergodic* $\gamma(n) = t^n$ for some $t < 1$ while *uniform ergodicity* means X is geometrically ergodic and M is bounded. These are key sufficient conditions for the existence of an asymptotic normal distribution of the Monte Carlo error but these are not the only conditions guaranteeing a CLT. In particular, a CLT as at (1.1.3) holds if X is geometrically ergodic and $E_\pi g^{2+\delta} < \infty$ for some $\delta > 0$ or if X is uniformly ergodic and $E_\pi g^2 < \infty$. Moreover, geometric ergodicity is a key sufficient condition for the strong consistency of the estimators of σ_g^2 from (1.1.3). For example, Flegal and Jones (2010a) establish that when X is geometrically ergodic and $E_\pi g^{2+\delta} < \infty$ for some $\delta > 0$ the overlapping batch means method produces a strongly consistent estimator of σ_g^2. Geometric ergodicity is also an important sufficient condition for establishing (1.4.3) when estimating a quantile (Flegal and Jones, 2010b).

In general, establishing (1.7.2) directly is apparently daunting. However, if X is finite (no matter how large), then a Harris ergodic Markov chain is uniformly ergodic. When X is a general space there are constructive methods which can be used to establish geometric or uniform ergodicity; see Hobert (2010) and Jones and Hobert (2001) for accessible introductions. These techniques have been applied to many MCMC samplers. For example, Metropolis-Hastings samplers with state-independent proposals can be uniformly ergodic (Tierney, 1994). Standard random walk Metropolis-Hastings chains on $\mathbb{R}^d$, $d \ge 1$ cannot be uniformly ergodic but may still be geometrically ergodic; see Mengersen and Tweedie (1996). An incomplete list of other research on establishing convergence rates of Markov chains used in MCMC is given by Atchade and Perron (2007), Christensen et al. (2001), Geyer (1999), Jarner and Hansen (2000), Meyn and Tweedie (1994) and Neath and Jones (2009) who considered Metropolis-Hastings algorithms and Doss and Hobert (2010), Hobert and Geyer (1998), Hobert et al. (2002), Johnson and Jones (2010), Jones and Hobert (2004), Marchev and Hobert (2004), Roberts and Polson (1994), Roberts and Rosenthal (1999), Rosenthal (1995, 1996), Roy and Hobert (2007), Roy and Hobert (2010), Tan and Hobert (2009) and Tierney (1994) who examined Gibbs samplers.

1.8 Discussion

The main point of this chapter is that a Monte Carlo standard error should be reported along with the point estimate obtained from an MCMC experiment. At some level this probably seems obvious to most statisticians but it is not the case in the reporting of most MCMC-based simulation experiments. In fact, Doss and Hobert (2010) recently wrote

> Before the MCMC revolution, when classical Monte Carlo methods based on i.i.d. samples were used to estimate intractable integrals, it would have been deemed unacceptable to report a Monte Carlo estimate without an accompanying asymptotic standard error (based on the CLT). Unfortunately, this seems to have changed with the advent of MCMC.

While it is tempting to speculate on the reasons for this change, the fact remains that most currently published work in MCMC reports point estimates while failing to even acknowledge an associated MCSE; see also Flegal et al. (2008). Thus we have little ability to assess the reliability of the reported results. This is especially unfortunate since it is straightforward to compute a valid MCSE.

The only potentially difficult part of the method presented here is in establishing the existence of a Markov chain CLT. In essence this means simulating a Markov chain known to be geometrically ergodic and checking a moment condition. Given the amount of work that has been done on establishing geometric ergodicity for standard algorithms in common statistical settings, this is not the obstacle it was in the past. However, this remains an area rich with important open research questions.

Acknowledgments

This work supported by NSF grant DMS-0806178. The authors would also like to thank the anonymous referees for their helpful suggestions.

Bibliography

Alexopoulos, C., Andradóttir, S., Argon, N. T., and Goldsman, D. (2006). Replicated batch means variance estimators in the presence of an initial transient. *ACM Transactions on Modeling and Computer Simulation*, 16:317–328.

Alexopoulos, C. and Goldsman, D. (2004). To batch or not to batch? *ACM Transactions on Modeling and Computer Simulation*, 14(1):76–114.

Atchade, Y. F. and Perron, F. (2007). On the geometric ergodicity of Metropolis-Hastings algorithms. *Statistics*, 41:77–84.

Bertail, P. and Clémençon, S. (2006). Regenerative block-bootstrap for Markov chains. *Bernoulli*, 12:689–712.

Bratley, P., Fox, B. L., and Schrage, L. E. (1987). *A Guide to Simulation*. Springer–Verlag, New York.

Brockwell, A. E. (2006). Parallel Markov chain Monte Carlo by pre-fetching. *Journal of Computational and Graphical Statistics*, 15:246–261.

Bühlmann, P. (2002). Bootstraps for time series. *Statistical Science*, 17:52–72.

Caffo, B. S., Peng, R., Dominici, F., Louis, T., and Zeger, S. (2010). Parallel Bayesian MCMC imputation for multiple distributed lag models: A case study in environmental epidemiology. In *Handbook of Markov Chain Monte Carlo* (to appear). CRC, London.

Casella, G. and Robert, C. P. (1996). Rao-Blackwellisation of sampling schemes. *Biometrika*, 83:81–94.

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

Christensen, O. F., Moller, J., and Waagepetersen, R. P. (2001). Geometric ergodicity of Metropolis-Hastings algorithms for conditional simulation in generalized linear mixed models. *Methodology and Computing in Applied Probability*, 3:309–327.

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

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

Datta, S. and McCormick, W. P. (1993). Regeneration-based bootstrap for Markov chains. *The Canadian Journal of Statistics*, 21:181–193.

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* (to appear).

Ferguson, T. S. (1996). *A Course in Large Sample Theory.* Chapman & Hall / CRC, Boca Raton.

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

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

Flegal, J. M. and Jones, G. L. (2010b). Quantile estimation via Markov chain Monte Carlo. Technical report, University of California, Riverside, Department of Statistics.

Flegal, J. M. and Jones, G. L. (2010c). Sweave documentation for ”Implementing Markov chain Monte Carlo: Estimating with confidence”. http://arXiv.org/abs/1006.5690.

Fosdick, L. D. (1959). Calculation of order parameters in a binary alloy by the Monte Carlo method. *Physical Review*, 116:565–573.

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

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

Geyer, C. J. (1992). Practical Markov chain Monte Carlo (with discussion). *Statistical Science*, 7:473–511.

Geyer, C. J. (1995). Conditioning in Markov chain Monte Carlo. *Journal of Computational and Graphical Statistics*, 4:148–154.

Geyer, C. J. (1999). Likelihood inference for spatial point processes. In Barndorff-Nielsen, O. E., Kendall, W. S., and van Lieshout, M. N. M., editors, *Stochastic Geometry: Likelihood and Computation*, pages 79–140. Chapman & Hall/CRC, Boca Raton.

Geyer, C. J. (2010). Introduction to Markov chain Monte Carlo. In *Handbook of Markov Chain Monte Carlo* (to appear). CRC, London.

Glynn, P. W. and Heidelberger, P. (1991). Analysis of initial transient deletion for replicated steady-state simulations. *Operations Research Letters*, 10:437–443.

Glynn, P. W. and Whitt, W. (1991). Estimating the asymptotic variance with batch means. *Operations Research Letters*, 10:431–435.

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

Hobert, J. P. (2010). The data augmentation algorithm: Theory and methodology. In *Handbook of Markov Chain Monte Carlo* (to appear). CRC, London.

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.

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

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

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

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

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

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

Kelton, A. M. and Law, W. D. (1984). An analytical evaluation of alternative strategies in steady-state simulation. *Operations Research*, 32:169–184.

Kosorok, M. R. (2000). Monte Carlo error estimation for multivariate Markov chains. *Statistics and Probability Letters*, 46:85–93.

Latuszynski, K. and Niemiro, W. (2009). Rigorous confidence bounds for MCMC under a geometric drift condition. http://arXiv.org/abs/0908.2098v1.

Liu, J. S., Wong, W. H., and Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika*, 81:27–40.

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.

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

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

Meyn, S. P. and Tweedie, R. L. (1994). Computable bounds for geometric convergence rates of Markov chains. *The Annals of Applied Probability*, 4:981–1011.

Muñoz, D. F. and Glynn, P. W. (2001). Multivariate standardized time series for steady-state simulation output analysis. *Operations Research*, 49:413–422.

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

Neath, R. and Jones, G. L. (2009). Variable-at-a-time implementations of Metropolis-Hastings. Technical report, University of Minnesota, School of Statistics.

Nummelin, E. (1984). *General Irreducible Markov Chains and Non-negative Operators*. Cambridge University Press, London.

Politis, D. N. (2003). The impact of bootstrap methods on time series analysis. *Statistical Science*, 18:219–230.

Politis, D. N., Romano, J. P., and Wolf, M. (1999). *Subsampling*. Springer-Verlag Inc.

Roberts, G. O. (1999). A note on acceptance rate criteria for CLTs for Metropolis-Hastings algorithms. *Journal of Applied Probability*, 36:1210–1217.

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. (1999). Convergence of slice sampler Markov chains. *Journal of the Royal Statistical Society,* Series B, 61:643–660.

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

Rosenthal, J. S. (1995). Minorization conditions and convergence rates for Markov chain Monte Carlo. *Journal of the American Statistical Association*, 90:558–566.

Rosenthal, J. S. (1996). Analysis of the Gibbs sampler for a model related to James-Stein estimators. *Statistics and Computing*, 6:269–275.

Rosenthal, J. S. (2000). Parallel computing and Monte Carlo algorithms. *Far East Journal of Theoretical Statistics*, 4:201–236.

Roy, V. and Hobert, J. P. (2007). Convergence rates and asymptotic standard errors for Markov chain Monte Carlo algorithms for Bayesian probit regression. *Journal of the Royal Statistical Society, Series B*, 69(4):607–623.

Roy, V. and Hobert, J. P. (2010). On Monte Carlo methods for Bayesian multivariate regression models with heavy-tailed errors. *Journal of Multivariate Analysis* (to appear).

Rudolf, D. (2009). Error bounds for computing the expectation by Markov chain Monte Carlo. *Preprint.*

Schruben, L. (1983). Confidence interval estimation using standardized time series. *Operations Research*, 31:1090–1108.

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

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.

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

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, New York.

Wei, G. C. G. and Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85:699–704.

Yang, W.-N. and Nelson, B. L. (1992). Multivariate batch means and control variates. *Management Science*, 38:1415–1431.