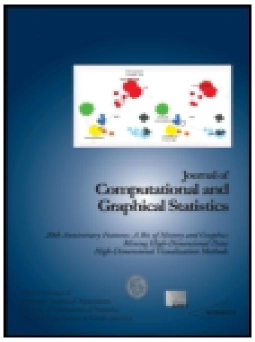

Taylor & Francis
Taylor & Francis Group

Journal of Computational and Graphical Statistics

ISSN: 1061-8600 (Print) 1537-2715 (Online) Journal homepage: https://www.tandfonline.com/loi/ucgs20

Assessing Convergence and Mixing of MCMC Implementations via Stratification

Rajib Paul , Steven N. MacEachern & L. Mark Berliner

To cite this article: Rajib Paul , Steven N. MacEachern & L. Mark Berliner (2012) Assessing Convergence and Mixing of MCMC Implementations via Stratification, Journal of Computational and Graphical Statistics, 21:3, 693-712, DOI: 10.1080/10618600.2012.663293

To link to this article: https://doi.org/10.1080/10618600.2012.663293

View supplementary material

Accepted author version posted online: 24 Feb 2012.
Published online: 16 Aug 2012.

Submit your article to this journal

Article views: 752

View related articles

Citing articles: 3 View citing articles

Supplementary materials for this article are available online.
Please go to www.tandfonline.com/r/JCGS

Assessing Convergence and Mixing of MCMC Implementations via Stratification

Rajib PAUL, Steven N. MACEACHERN, and L. Mark BERLINER

Some posterior distributions lead to Markov chain Monte Carlo (MCMC) chains that are naturally viewed as collections of subchains. Examples include mixture models, regime-switching models, and hidden Markov models. We obtain MCMC-based estimators of posterior expectations by combining different subgroup (subchain) estimators using stratification and poststratification methods. Variance estimates of the limiting distributions of such estimators are developed. Based on these variance estimates, we propose a test statistic to aid in the assessment of convergence and mixing of chains. We compare our diagnostic with other commonly used methods. The approach is illustrated in two examples: a latent variable model for arsenic concentration in public water systems in Arizona and a Bayesian hierarchical model for Pacific sea surface temperatures. Supplementary materials, which include MATLAB codes for the proposed method, are available online.

Key Words: Batch-means methods; Bootstrap; Convergence diagnostics; Delta method; Functional central limit theorem; Mixing; Stationarity.

1. INTRODUCTION

In Bayesian analysis, we often rely on Markov chain Monte Carlo (MCMC) simulation to provide information regarding complex, intractable posterior distributions. The idea is to construct a Markov chain having a stationary distribution that coincides with the target posterior distribution. If we simulate the chain for a sufficiently long time, we achieve approximate convergence to the stationary distribution. The resulting sample is a reasonable approximation to a sample from the target posterior distribution. Hence, a critical problem is the assessment of convergence to the stationary distribution.

A related issue involves the degree of mixing of the chain. Operationally, if the chain explores all important regions of its stationary distribution, we say that the chain is well mixed. Several formal definitions of mixing are in the literature, although we believe the most common notion is α-mixing: a Markov chain is α-mixing if the supremum over all

Rajib Paul is Assistant Professor, Department of Statistics, Western Michigan University, Kalamazoo, MI 49008 (E-mail: *rajib.paul@wmich.edu*). Steven N. MacEachern is Professor, Department of Statistics, The Ohio State University, Columbus, OH 43210 (E-mail: *snm@stat.osu.edu*). L. Mark Berliner is Professor, Department of Statistics, The Ohio State University, Columbus, OH 43210 (E-mail: *mb@stat.osu.edu*).

Journal of Computational and Graphical Statistics, Volume 21, Number 3, Pages 693–712
DOI: 10.1080/10618600.2012.663293

sets A and B of the absolute difference $|P(\theta_n \in A, \theta_0 \in B) - P(\theta_n \in A)P(\theta_0 \in B)|$ tends to zero as n grows. Intuitively, this means that the initial condition of the chain "washes out" and that values of the chain separated by large time intervals should "appear to be independent." Also, note that if P is a mixture, that is, $P(\theta_n \in A) = \sum_j p_j P_j(\theta_n \in A)$, then $P(\theta_n \in A \mid \theta_0 \in B)$ should eventually reflect this mixture for all B. That is, well-mixed chain must explore the important modes (i.e., visit the high-probability regions of the P_j), and must spend approximately the right amount of time in each mode (i.e., provide consistent estimates of the p_j). (The idea of this argument applies to all multimodal P.) Finally, the two notions are linked: if successive states of the chain are highly correlated, we expect a slow rate of convergence to asymptotic independence and a slow rate of visitation to the various modes of the P_j.

Determining whether a chain has adequately converged to its stationary distribution and mixed well is a difficult problem. Several methods for addressing this issue appear in the literature (see Section 1.1). Our investigations have suggested that the available procedures do a better job of assessing convergence than they do of assessing mixing. We believe that this is in part due to the difficulty in formally defining what one means by adequate mixing. It is relatively common for analysts in a team of researchers to have different opinions as to whether a chain has mixed well enough for the results to be presented. For example, trace plots and running means plots are widely used in data-analytic assessments of convergence. However, trace plots may be misleading. In particular, differentiating between nonconvergence and convergence to multimodal distributions is difficult, unless we know the number and locations of the modes, as in the development of Gelman and Rubin (1992).

Taking this view into account, we recast the problem. Instead of searching for the shortest chain that appears to be adequate for inference, our goal is to be able to say that the vagaries of the particular realization of the chain that we have generated have relatively little impact on our inferences. We hope to have a long enough chain (i.e., enough data from our simulation) so that our inference is stable and relatively insensitive to the technique used to make that inference. Pursuing this line, we propose a computationally simple technique to assess convergence and mixing by comparing estimates of some quantity.

The method that we propose is based on stratification and the use of batch means (Jones et al. 2006; Flegal, Haran, and Jones 2008). Specifically, we compare properties of estimates based on two different, but consistent estimators of the mean of the stationary distribution: (1) the usual sample mean based on the full series and (2) the appropriately weighted average of stratified sample means. We demonstrate that under stationarity and mixing, the asymptotic distributions of the estimated variances of the two estimators are equal. Our diagnostic quantifies the discrepancy between properties of the two estimators by comparing estimates of the variances of the estimators' asymptotic distributions. If the estimated variances are "close," we find no evidence for nonconvergence or poor mixing. This step is formalized by formulating the problem in the language of hypothesis testing, and by implementing the "test" using a bootstrap approach.

The definition of the strata can be tailored to match prerun expectations regarding concerns for mixing. For example, strata may be defined to be favored components or regions of the parameter space in the context of a mixture model. The method also allows for poststratification. These ideas are illustrated in the examples.

Our method involves the stratification of a single series and, hence, does not require multiple chains. Of course, there are desirable aspects to inspection of multiple runs. Hence, we also discuss the use of our approach when multiple chains are available in Section 5.2. Finally, we remark that this article is devoted to the assessment of convergence only. We do not consider MCMC-derived estimates nor supporting techniques such as thinning.

1.1 REVIEW: CONVERGENCE DIAGNOSTICS

We provide a very brief review. In-depth presentations can be found in Cowles and Carlin (1996) and Robert and Casella (2004, chap. 12).

Perhaps the most commonly used convergence diagnostic is the Gelman–Rubin statistic (Gelman and Rubin 1992). Implementation requires sample runs from multiple chains. The key quantity is the ratio of the resulting between- and within-chain variances. If the within-chain variance dominates the between-chain variance, the ratio approaches 1, which suggests that the chains have reached approximate stationarity. Geweke (1992) proposed a method for assessing convergence by testing the equality of the means of first $n_f\%$ and last $n_l\%$ of the iterates of an MCMC run. The test relies on the assumption of independence between the sample means based on the first $n_f\%$ and the last $n_l\%$ of the iterates. Typically, n_f is chosen to be 10 and n_l is set at 50. The variances of the sample means are estimated using a spectral density method.

Another method was proposed by Raftery and Lewis (1992). Suppose we focus on estimating the posterior probability $P(\theta \leq \theta^*|Y)$, where θ is the parameter of interest, θ^* is a specified quantity, and Y denotes the data. They required that the estimator should be within $\pm r$ of the true value with probability s. Based on this requirement, their method estimated the length of the burn-in period of the chain and the minimum number of iterations needed to achieve the probability requirement.

Other methods have been proposed by Roberts (1992), Ritter and Tanner (1992), and Zellner and Min (1995). These methods can be applied to assess convergence for a multivariate posterior distribution. However, these methods are constructed for application to Gibbs samplers and are very demanding computationally. Also, see Liu, Liu, and Rubin (1992) for a related approach.

Heidelberger and Welch (1983) provided approximate confidence intervals for posterior means based on spectral density and Brownian bridge arguments. Their method was applied sequentially to assess "burn-in." If the test supports the null hypothesis of convergence of the chain, a confidence interval for the mean is readily available. This method is based on the assumption of ϕ-mixing or geometrically ergodic chains.

Mykland, Tierney, and Yu (1995) developed a method based on regeneration theory. They used "scaled regeneration quantile (SRQ) plots" to assess stationarity. The method requires expression of the MCMC algorithm in a very specific and often infeasible fashion. Also, see Jones et al. (2006) and Flegal, Haran, and Jones (2008) for approaches using regeneration theory and batch-means methods.

1.2 OUTLINE

Section 2 presents the details of our approach. In Sections 3.1 and 3.2, we apply our method to illustrative examples. Section 4 provides a numerical assessment of the power of our test. Section 5 presents discussion regarding implementation of our approach and a summary.

2. A NEW METHOD

Let θ be a parameter of interest. Though not reflected in the notation, note that θ could be a selected function of the original state variable or a scalar function of a multivariate state vector. Assume we have a sample of size N, $\theta_1, \ldots, \theta_N$, generated from an MCMC implementation with stationary distribution π. We divide the sample into K batches, each of size n (i.e., $N = K \times n$). For $k = 1, \ldots, K$, let $\overline{\theta}_{[k]}$ be the kth batch mean. A natural estimator of the mean, $E^{\pi}(\theta)$, is the sample mean,

$$E_1 = \frac{1}{N} \sum_{i=1}^{N} \theta_i = \frac{1}{K} \sum_{k=1}^{K} \overline{\theta}_{[k]}. \tag{1}$$

We construct a second estimator, E_2, based on stratification. Partition the parameter space Θ into J "nontrivial," disjoint strata, $A_1, \ldots, A_J$. By "nontrivial," we mean that each stratum has positive probability under π. Define two quantities for the kth batch as follows:

$$\overline{Z}^{P}_{[k]j} = \frac{1}{n} \sum_{i=nk+1}^{n(k+1)} Z^{P}_{ij}, \tag{2}$$

$$\overline{Z}^{\theta}_{[k]j} = \frac{1}{n} \sum_{i=nk+1}^{n(k+1)} Z^{\theta}_{ij}, \tag{3}$$

where the index j denotes the jth stratum, $j = 1, \ldots, J$, and

$$Z^{P}_{ij} = I(\theta_i \in A_j),$$
$$Z^{\theta}_{ij} = \theta_i I(\theta_i \in A_j),$$

where $I(\cdot)$ is the indicator function: $I(A) = 1$ if the event A occurs, and $I(A) = 0$ otherwise. The quantities in (2) are estimates of strata probabilities based on each batch separately. Similarly, the quantities in (3) are batch- and stratum-specific mean estimates.

Let $\overline{\mathbf{Z}}$ be the $J(2K - 1) \times 1$ vector of the $\overline{Z}_{[k]}$'s. A simple estimate of the probability under π that θ belongs to the jth stratum is $\overline{P}_j = \frac{1}{N} \sum_{i=1}^{N} Z^{P}_{ij} = \frac{1}{K} \sum_{k=1}^{K} \overline{Z}^{P}_{[k]j}$. We define $E_2 = g_2(\overline{\mathbf{Z}})$, where

$$g_2(\overline{\mathbf{Z}}) = \frac{1}{K} \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{\overline{P}_j}{\overline{Z}^{P}_{[k]j}} \overline{Z}^{\theta}_{[k]j}. \tag{4}$$

Note that we can write the sample mean E_1 in terms of function, $g_1(\overline{\mathbf{Z}})$, of the $\overline{Z}_{[k]j}$'s as follows:

$$E_1 = g_1(\overline{\mathbf{Z}}) = \frac{1}{K} \sum_{j=1}^{J} \sum_{k=1}^{K} \overline{Z}^{\theta}_{[k]j}. \tag{5}$$

For a well-mixed chain, we expect the estimated probabilities of the strata based on the full sample to be close to those estimated from each of the batches. In that situation, the ratios $\overline{P}_j / \overline{Z}^{P}_{[k]j}$ in (4) will be close to 1, suggesting that E_2 and E_1 will be approximately equal.

Next, we turn to the variances of E_1 and E_2. Note that in the expression for E_2 in (4), the batch proportions in the denominator may take on zero values. Hence, the variance of E_2 does not exist (and E_2 is not fully defined). However, since all strata have positive π probability, E_2 is "asymptotically well defined" under mild conditions and the variance of the asymptotic distribution (as $n \to \infty$) of E_2 is finite. This asymptotic variance will be estimated and the result is then compared with the estimated variance of the asymptotic distribution of E_1.

2.1 DESCRIPTION OF THE METHOD

The development proceeds in four steps.

Step 1. In the Appendix, we develop theory and assumptions to justify an asymptotic distribution for $\overline{\mathbf{Z}}$:

$$\sqrt{n}\left(\overline{\mathbf{Z}} - E^{\pi}(\overline{\mathbf{Z}})\right) \xrightarrow{D} \mathrm{MVN}(\mathbf{0}, \Sigma_Z) \text{ as } n \to \infty, \tag{6}$$

where Σ_Z is the covariance matrix of asymptotic distribution of $\overline{\mathbf{Z}}$.

Step 2. Relying on the delta method (Lehmann and Casella 1998, p. 60), we obtain the following result.

Theorem 1. The asymptotic distribution of E_2 is

$$\sqrt{n}\left(E_2 - g_2(\overline{\mathbf{Z}}_o)\right) \simeq N\left(\mathbf{d}_2^t(E^{\pi}(\overline{\mathbf{Z}}) - g_2(\overline{\mathbf{Z}}_{\mathrm{o}})), \mathbf{d}_2^t\boldsymbol{\Sigma}_Z\mathbf{d}_2\right), \tag{7}$$

where $\overline{\mathbf{Z}}_{\mathrm{o}}$ is the observed value of $\overline{\mathbf{Z}}$ and the gradient $\mathbf{d}_2 = g_2'(\overline{\mathbf{Z}}) = dE_2/d\overline{\mathbf{Z}}$ is evaluated at $\overline{\mathbf{Z}}_{\mathrm{o}}$. The asymptotic distribution of E_1 is

$$\sqrt{n}\left(E_1 - g_1(\overline{\mathbf{Z}}_{\mathrm{o}})\right) \simeq N\left(\mathbf{d}_1^t(E^{\pi}(\overline{\mathbf{Z}}) - g_1(\overline{\mathbf{Z}}_{\mathrm{o}})), \mathbf{d}_1^t\boldsymbol{\Sigma}_Z\mathbf{d}_1\right), \tag{8}$$

where $\mathbf{d}_1 = g'{}_1(\overline{\mathbf{Z}}) = dE_1/d\overline{\mathbf{Z}}$ is evaluated at $\overline{\mathbf{Z}}_{\mathrm{o}}$.

Note that $\mathbf{d}_1 = [\mathbf{1}_{(J-1)K\times 1}, \frac{1}{K}\mathbf{1}_{JK\times 1}]$ and $\mathbf{d}_2 = [\{dE_2/d\overline{Z}^P_{[k]j}\}, \{dE_2/d\overline{Z}^\theta_{[k]j}\}]$, where

$$\begin{aligned}\frac{dE_2}{d\overline{Z}^P_{[k]j}} &= \frac{1}{K^2}\sum_{k=1}^{K}\frac{\overline{Z}^\theta_{[k]j}}{\overline{Z}^P_{[k]j}} - \frac{1}{K}\frac{\overline{P}_j\overline{Z}^\theta_{[k]j}}{(\overline{Z}^P_{[k]j})^2} - \frac{1}{K^2}\sum_{k=1}^{K}\frac{\overline{Z}^\theta_{[k]J}}{1-\sum_{j=1}^{J-1}\overline{Z}^P_{[k]j}}\\ &\quad + \frac{1}{K}\frac{(1-\sum_{j=1}^{J-1}\overline{P}_j)\overline{Z}^\theta_{[k]J}}{(1-\sum_{j=1}^{J-1}\overline{Z}^P_{[k]j})^2},\\ \frac{dE_2}{d\overline{Z}^\theta_{[k]j}} &= \frac{1}{K}\frac{\overline{P}_j}{\overline{Z}^P_{[k]j}}.\end{aligned}$$

Step 3. Under the assumptions in the Appendix, application of the ergodic theorem implies that

$$\mathbf{d}_2 - \mathbf{d}_1 \xrightarrow{\text{a.s.}} 0 \quad \text{as } n \to \infty. \tag{9}$$

Let V_i denote the variance of the asymptotic distribution of E_i, $i = 1, 2$.

Theorem 2. $V_1 = V_2$.

Proof. Let $\mathbf{d}$ be the gradient $dE_2/d\overline{\mathbf{Z}}$ evaluated at $E(\overline{\mathbf{Z}})$. Applying the delta method to E_2, we find that $V_2 = \mathbf{d}^t \boldsymbol{\Sigma}_Z \mathbf{d}$. Recalling (8), we have that $V_1 = \mathbf{d}_1^t \boldsymbol{\Sigma}_Z \mathbf{d}_1$. A simple algebra shows $\mathbf{d}_1 = \mathbf{d}$, which implies the result.

Natural estimates of the variances of the asymptotic distributions of E_1 and E_2 are

$$\widehat{V}_1 = \frac{1}{n}\mathbf{d}_1^t \widehat{\Sigma}_Z \mathbf{d}_1 \tag{10}$$

and

$$\widehat{V}_2 = \frac{1}{n}\mathbf{d}_2^t \widehat{\Sigma}_Z \mathbf{d}_2, \tag{11}$$

respectively, where $\widehat{\Sigma}_Z$ is an estimator of Σ_Z. Note that if $\hat{\Sigma}_Z$ is a consistent estimator of Σ_Z, then $\hat{V}_2$ is consistent for V_2 by (9).

As explained in the Appendix, as the batch size n tends to ∞, the batch means are asymptotically uncorrelated under mild conditions (e.g., an absolutely summable autocovariance function). Hence, Σ_Z approaches a block-diagonal matrix $\Sigma \otimes I_K$, where Σ is a symmetric, positive-definite $(2J-1) \times (2J-1)$ matrix and I_K is the identity matrix of rank K. There are at least two ways to estimate Σ: (1) apply renewal theory (e.g., Hobert et al. 2002) and (2) use batch-based methods (e.g., Jones et al. 2006). Application of renewal theory requires writing the MCMC algorithm in a specific format (split-chain method). Since this is a daunting task for complex models, we use the batch-means method here. The diagonal elements of Σ can be estimated by

$$\hat{\sigma}^P_{j,j} = \frac{n}{K-1}\sum_{k=1}^{K}\left(\overline{Z}^P_{[k]j} - \overline{Z}^P_j\right)^2, \tag{12}$$

$$\hat{\sigma}^\theta_{j,j} = \frac{n}{K-1}\sum_{k=1}^{K}\left(\overline{Z}^\theta_{[k]j} - \overline{Z}^\theta_j\right)^2, \tag{13}$$

where $\overline{Z}^P_j = \frac{1}{K}\sum_{k=1}^{K}\overline{Z}^P_{[k]j}$ and $\overline{Z}^\theta_j = \frac{1}{K}\sum_{k=1}^{K}\overline{Z}^\theta_{[k]j}$. The off-diagonal elements are estimated by

$$\hat{\sigma}^P_{i,j} = \frac{n}{K-1}\sum_{k=1}^{K}\left(\overline{Z}^P_{[k]i} - \overline{Z}^P_i\right)\left(\overline{Z}^P_{[k]j} - \overline{Z}^P_j\right), \tag{14}$$

$$\hat{\sigma}^{P\theta}_{i,j} = \frac{n}{K-1}\sum_{k=1}^{K}\left(\overline{Z}^P_{[k]i} - \overline{Z}^P_i\right)\left(\overline{Z}^\theta_{[k]j} - \overline{Z}^\theta_j\right), \tag{15}$$

$$\hat{\sigma}^\theta_{i,j} = \frac{n}{K-1}\sum_{k=1}^{K}\left(\overline{Z}^\theta_{[k]i} - \overline{Z}^\theta_i\right)\left(\overline{Z}^\theta_{[k]j} - \overline{Z}^\theta_j\right). \tag{16}$$

Damerdji (1994) and Jones et al. (2006) showed that, under certain assumptions, if n and K both tend to ∞, then these estimators are consistent.

Step 4: Diagnostic Procedure. As shown in the Appendix, the asymptotic distributions of $n\widehat{V}_1$ and $n\widehat{V}_2$ coincide under stationarity and mixing. To quantify departures from those properties, we construct a parametric bootstrap (see Davison and Hinkley 1997, chap. 2; Casella and Berger 2002, chap. 10) test to compare $\widehat{V}_1$ and $\widehat{V}_2$. Motivated by the asymptotic normality result given in (6), we approximate the distribution of $\widehat{V}_1$ as follows. Samples are generated from a normal distribution, with mean $\hat{E}(\overline{\mathbf{Z}})$ and covariance matrix $\hat{\Sigma}_Z$. We then compute $\widehat{V}_1$ as given in (10) for each sample. The resulting sample of $\widehat{V}_1$s is used to estimate the distribution of $\widehat{V}_1$. Specifically, if the observed value of $\widehat{V}_2$ falls between the $p_{0.05\alpha}$ and $p_{1-0.50\alpha}$ bootstrap-estimated quantiles of $\widehat{V}_1$, then, at level α, we find no evidence of nonstationarity or poor mixing.

Remark 1. Our test is based on generating samples of $\widehat{V}_1$ rather than $\widehat{V}_2$ because $\widehat{V}_1$ is more stable than $\widehat{V}_2$.

2.2 BOOTSTRAP SAMPLE SIZE

We develop rough suggestions regarding the selection of the bootstrap sample size based on our experience with examples.

As simple examples of a Markov chain that can display various levels of dependence, we considered first-order autoregressive [AR(1)] processes of the form

$$X_t = \alpha X_{t-1} + \epsilon_t, \tag{17}$$

where $|\alpha| < 1$ and $\{\epsilon_t\}$ is a white-noise process. Recall that the correlation between X_t and X_{t+s} is α^s. We chose α and the variance of the noise in such a way that the stationary distribution of the chain is a zero-mean Gaussian distribution with variance 1.

We set $\alpha = 0.2$ and generated 50 independent chains, each of size 120,000. Each sample was split into 30 batches of size 4000. In each of the 50 cases, we performed our test based on four different bootstrap sample sizes: 1000, 10,000, 25,000, and 50,000.

From the results in Figure 1, we observe that in all cases, our test supports the claim that the chain is well mixed. The black dots in these plots indicate the observed values of $\hat{V}_2$. Note that the lengths of confidence intervals vary little as the bootstrap sample size varies.

Next, we repeated the entire experiment with $\alpha = 0.998$. All of our tests rejected the claim that the chain is well mixed. Figure 2 shows the bootstrap intervals. In these plots, all sample values of $\hat{V}_2$ lie to the right of the bootstrap intervals and are off the plot. Again, the lengths of the confidence intervals are similar for the four bootstrap sample sizes.

From the evidence in these examples and further studies not presented here, we recommend a default bootstrap sample size of 1000.

3. EXAMPLES

3.1 LATENT VARIABLE MODEL FOR ARSENIC CONCENTRATIONS

Craigmile et al. (2009) developed a latent variable model for arsenic (As) concentrations in 1161 public water systems (PWS) in Arizona. The data are measurements of log(As)

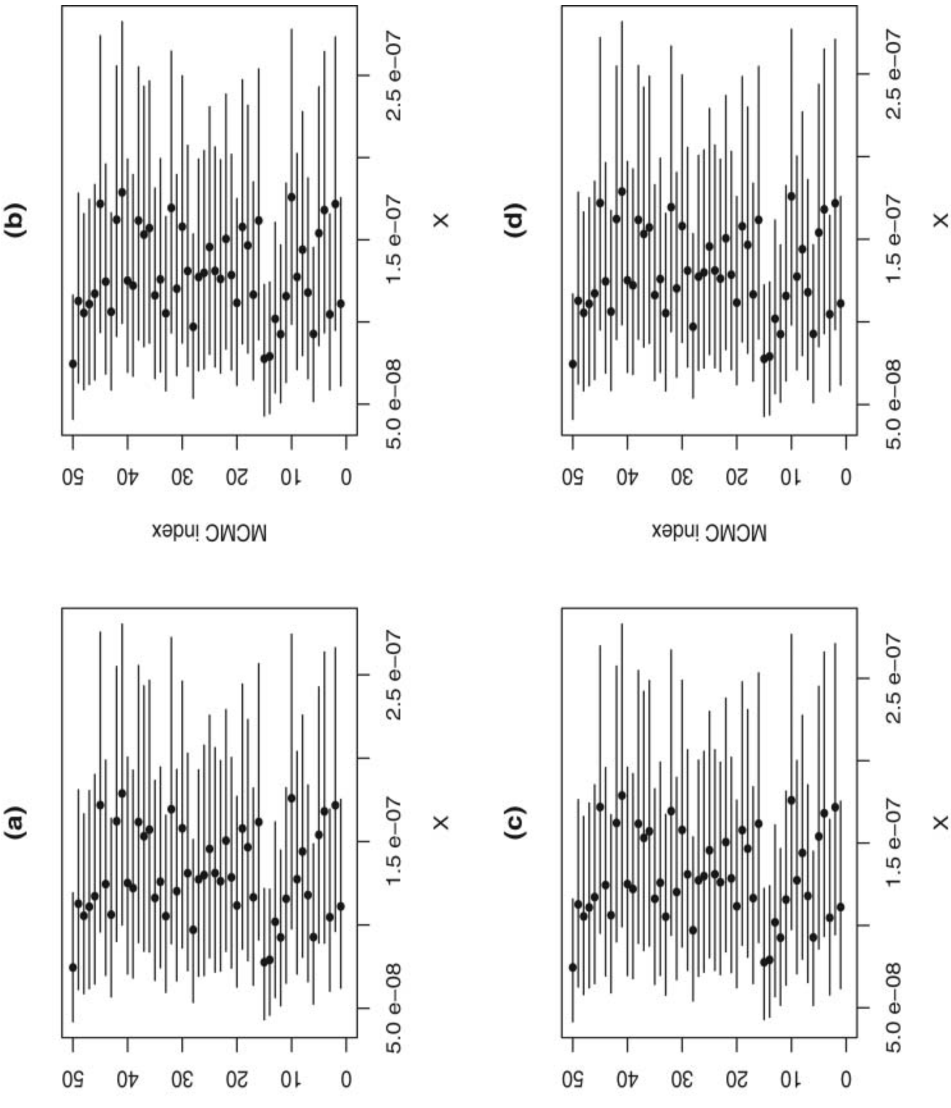

Figure 1. Acceptance regions constructed for the model in (17) with $\alpha = 0.2$ using four bootstrap sample sizes: (a) 1000, (b) 10,000, (c) 25,000, and (d) 50,000.

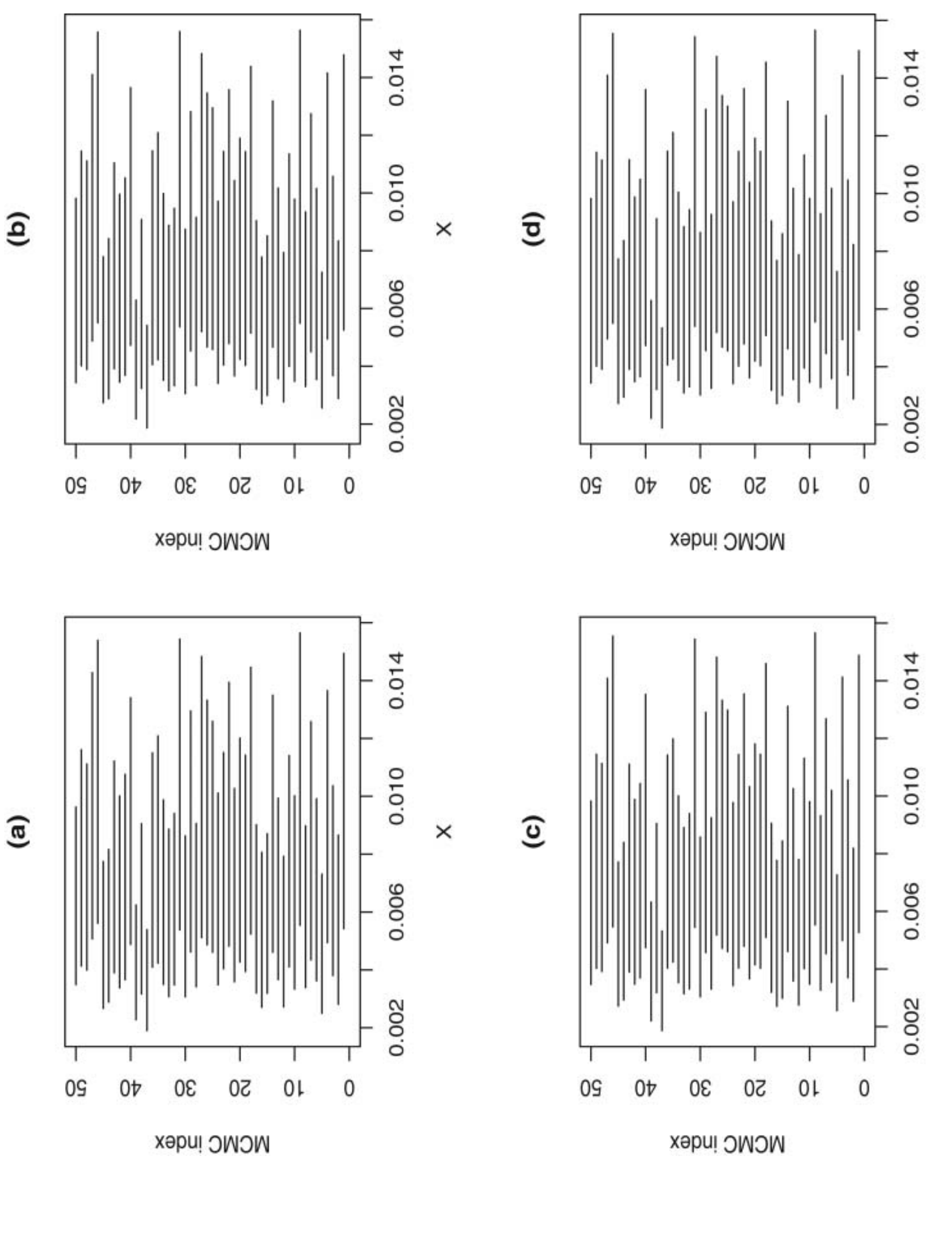

Figure 2. Acceptance regions constructed for the model in (17) with $\alpha = 0.998$ using four bootstrap sample sizes: (a) 1000, (b) 10,000, (c) 25,000, and (d) 50,000.

concentrations, $\{Y_{il} : i = 1, \ldots, 1161, l = 1, \ldots, n_i\}$, where i indexes PWS and l indexes observations at a given PWS. The model has three main parts: (1) a data model for the observed log(As) concentrations; (2) a process model for the true log(As) concentrations, denoted by X_i, $i = 1, \ldots, 1161$; and (3) prior distributions for all unknown parameters introduced in the modeling.

For each PWS i, Craigmile et al. (2009) assumed Gaussian data models for the $\{Y_{il} : l = 1, \ldots, n_i\}$, with unknown mean X_i. The process model is a Gaussian conditional autoregressive (CAR) model with constant mean. Let $\overline{Y}_i$ denote the sample mean of the measurements for PWS i, $i = 1, \ldots, 1161$. For each i, the range of X_i was stratified as $X_i \leq \overline{Y}_i$ and $X_i > \overline{Y}_i$. We then performed our diagnostic tests on some initial MCMC runs separately for each PWS. For 945 PWSs, the $\hat{V}_2$s fall within the acceptance region, but for the other 216 PWSs, the $\hat{V}_2$s are outside the acceptance region. Hence, we conclude that the chain has not converged or mixed well. See Craigmile et al. (2009) for further discussion of this issue.

3.2 Regime-Switching Model for Pacific Sea Surface Temperatures

Berliner, Wikle, and Cressie (2000) developed a Bayesian dynamic forecasting model for the prediction of gridded tropical Pacific sea surface temperature (SST) monthly anomalies with 7-month leads. They used a principal components regression model with time-varying coefficients. Their modeling included allowance for regimes associated with warmer-than-normal (i.e., “El Nino”), normal, and cooler-than-normal (i.e., “La Nina”) states. We labeled these regimes as $R = 1, 2$, and 3, respectively. The time-varying regression coefficients followed mixture models with three components corresponding to these regimes.

Using the algorithms and models of Berliner, Wikle, and Cressie (2000), we ran an MCMC analysis based on data from January 1971 through February 2008 to predict SST anomalies for September 2008. We considered two parameters for illustrative purposes: (1) the regression coefficient of the leading principal component for September 2008, denoted as a, and (2) the regime indicator R.

Our method cannot be applied directly to the values of the regime indicator R, since the computed variances within each stratum would be equal to 0. We defined three strata for the regression coefficient a based on the values of R. That is, letting m index the MCMC sample, we defined $A_j = \{a^{(m)} : R^{(m)} = j)\}$, $j = 1, 2, 3$. After discarding the first 1000 iterates, we split $N = 10,000$ iterates into $K = 20$ batches of equal size $n = 500$. The acceptance region obtained from the bootstrap was $(0.0136, 0.0503)$ and $\hat{V}_2 = 0.0304$. Hence, we found no evidence of nonconvergence or poor mixing. The plot in Figure 3 also indicates that all the three regimes are visited regularly.

4. POWER OF THE TEST

In this section, we provide numerical assessments of the power of our test. Qualitatively, our null hypothesis is the claim that the chain has converged and mixed well. We quantify this by forming a null hypothesis that the estimates of the variances of the asymptotic distributions of two estimators (E_1 and E_2) are stochastically equivalent.

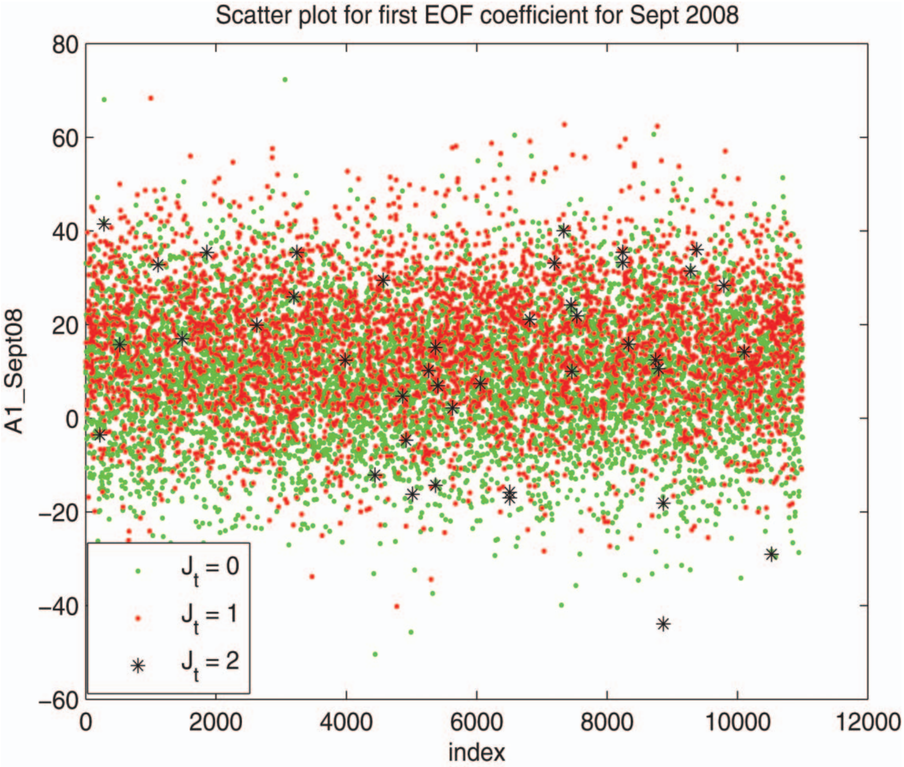

Figure 3. Plot of MCMC output for the coefficient a of the lead principal component for predicting SST for September 2008.

As a simple illustration, we considered the performance of diagnostics for an autoregressive process of order 1 defined by

$$X_t = 0.995X_{(t-1)} + e_t, \tag{18}$$

where $e_t \sim N(0, (1 - 0.995^2))$. It is straightforward to verify that the stationary distribution is an $N(0, 1)$ distribution. Note that this process is an example of a stationary, but slow-mixing chain, due to the high levels of autocorrelation. As evidence against mixing in this example, we present density estimates using 80,000 iterates of 10 generated chains in Figure 5. The thick, black curve is the density of the stationary distribution. The differences in density estimates show that there is substantial variation in the chains. In this example, we view the "correct" answer to be that the chain has not mixed adequately.

We simulated 1000 independent chains from (18), each of length 80,000, with initial states X_0 generated from the stationary distribution. Figure 4 shows the trace plot for one of the chains.

Table 1 compares the results obtained from conducting our test, Geweke's test, and the Gelman–Rubin test on the 1000 simulated chains. Geweke's tests were conducted using the first 10% and the last 50% of the iterates. The Gelman–Rubin tests were conducted on each chain by splitting the chain into eight groups, each of length 9000. To roughly approximate the required independence, we excluded 1000 iterates between groups and acted as if each

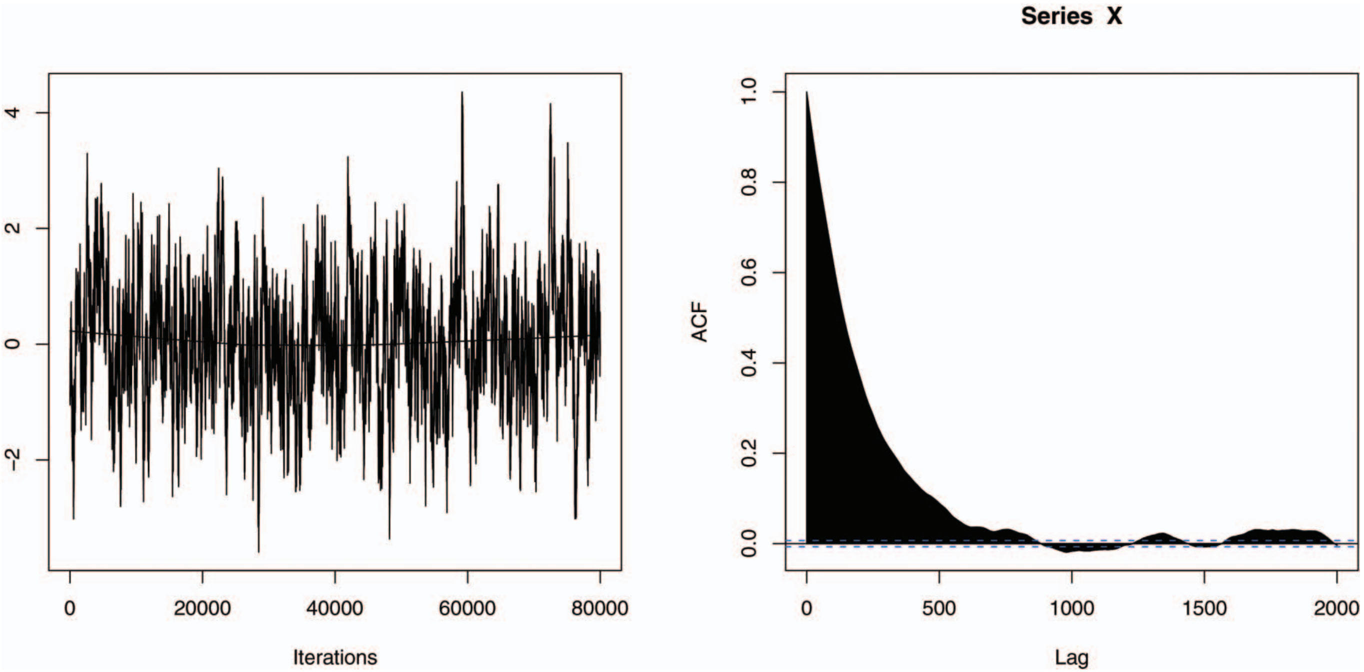

Figure 4. Trace plot and estimated autocorrelation function of an MCMC chain generated using (18). The online version of this figure is in color.

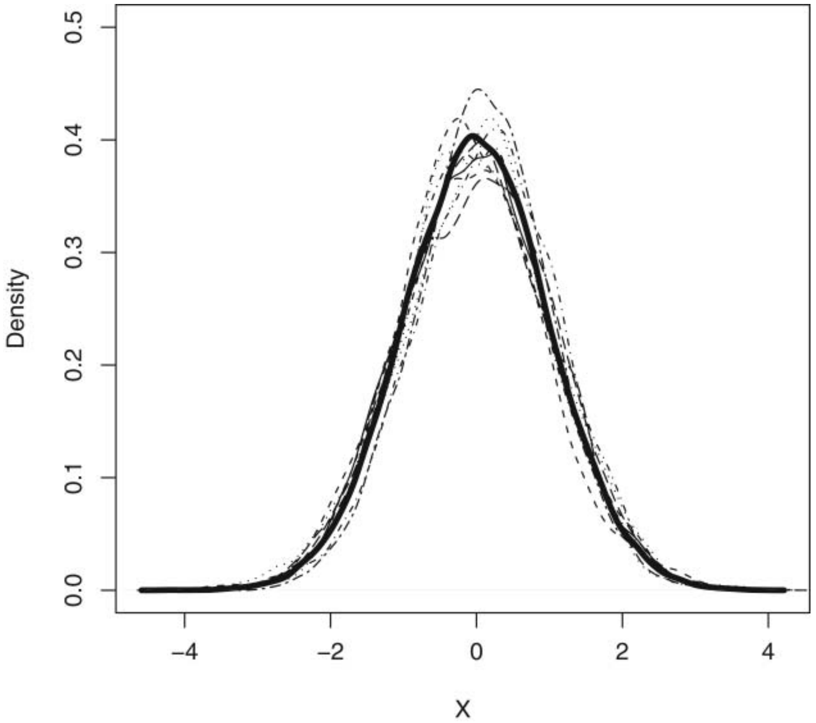

Figure 5. Density estimates from 10 simulated chains generated using (18). The thick, black line curve is the density of the stationary distribution.

of the resulting eight groups were eight independent chains. Our test was conducted on each of the 1000 runs using 20 batches, each of size 4000. We stratified the X space based on the conditions $X > 2$ and $X \leq 2$.

The diagonal entries of Table 1 are the proportions of the cases where each method supported the null hypothesis (we set the error rate to be 0.05). The off-diagonal entries show the numbers of cases in which the indicated methods both supported the null hypothesis. The Gelman–Rubin test fails to reject the null in all 1000 cases. The Geweke's test fails to reject the null in 82.4% of the cases. Of course, neither of these tests was developed to assess mixing. Our test fails to reject the null in only 2.2% of the cases, indicating that it does indeed assess both mixing and stationarity.

Table 1. Acceptance rates for Geweke's test, the Gelman–Rubin test, and our test. The off-diagonal values indicate the numbers of cases in which the two indicated tests both accepted the null hypothesis

Method	Geweke	Gelman–Rubin	Our test
Geweke	824/1000	824/1000	19/1000
Gelman–Rubin		1000/1000	22/1000
Our test			22/1000

Table 2. Estimated acceptance rates for our test applied to an autoregressive process using strata $X > c$ and $X \leq c$ for selected values of c. Each percentage is based on 1000 independent simulations

c	−2.50	−2.00	−1.50	−1.00	0	1.00	1.50	2.00	2.50
Acceptance rate	0	0.008	0.372	0.841	1	0.927	0.316	0.008	0

Next, we conducted our test using different stratification schemes, namely using the conditions $X > c$ and $X \leq c$, for c being −2.50 to 2.50 at an interval of 0.5. Table 2 presents the proportions of the cases for which we cannot reject the null hypothesis that we have a well-mixed chain as c varies (the level of all tests was 0.05). Note that the rejection rate increased as c increased. This suggests that the use of small c results in comparatively low power. However, use of too large a value of c may result in too few MCMC iterates occurring in the stratum $X > c$. Balance between these situations is desirable, though achieving it may require both "art" and preliminary computations, when feasible. Our default suggestion for a "first attempt" is to stratify the parameter space into three groups: (1) $X \leq x_{0.10}$, (2) $x_{0.10} < X < x_{0.90}$, and (3) $X \geq x_{0.90}$, where x_r denotes the rth quantile of the MCMC iterates of X.

We performed another simulation experiment to develop some guidance regarding the selection of the number of batches. We fixed the stratification to be

$$X > 2 \text{ and } X \leq 2. \tag{19}$$

We selected two sample sizes $N = 60{,}000$ and 300,000, hoping that the first size is too small to expect satisfactory convergence, while the latter is large enough for convergence to be achievable. For each N, we simulated three independent sets of 1000 independent chains and we examined our test results for $k = 15$, 30, and 60. The empirical acceptance rates are reported in Table 3. In concert with our expectations, our test rejects the claim of convergence in nearly all cases based on $N = 60{,}000$. The results are mixed when $N = 300{,}000$. Although the experiment is limited, the results appear to be in concert with traditional saws regarding the role of sample size in testing. Since the "effective" sample size is k, the results suggest that if k is too small, our test has low power, while if k is very large, we can expect to reject the null even when it is reasonably valid. Our recommendation is to keep the number of batches around 30.

Table 3. Estimated acceptance rates for our test applied to an autoregressive process and strata defined in (19). Each percentage is based on 1000 independent simulations

Series length N	k	n	Acceptance rate
60,000	15	4000	4%
	30	2000	0%
	60	1000	0%
300,000	15	20,000	100%
	30	10,000	54%
	60	5000	0%

5. DISCUSSION

5.1 SELECTION OF n, K, AND THE STRATA

Our method requires specification of the following control parameters: n (the number of batches), K (the batch size), and J (the number of strata). Since our derivations depend on several asymptotic results, where n and K both tend to infinity, it is essential that we choose large values of n and K. Jones et al. (2006) suggested that one can choose $n = K = \sqrt{N}$, where N is the length of the full chain. From experience with application of our method in a range of examples, we suggest that K be close to 30.

Next, J and the strata can be selected in various ways. First, exploratory data analysis (EDA) can provide useful information on stratification. Simple histograms or dendrogram plots of the data can reveal the number of components in the mixture model. If the number of variables is too large for simple plots, various methods to identify the number and pattern of clusters may be of interest. Many software packages are available to identify clusters, such as R-software library mixAK (Komarèk 2009) and R-function kmeans (Hartigan and Wong 1979). Second, as in Section 3.2, we may suggest the regimes from the scientific knowledge of the variables to be analyzed.

A third suggestion is to poststratify based on sample quantiles or likelihood-based selections (e.g., maximum likelihood estimates). Many MCMC algorithms do a decent job in sampling from high-probability regions (modes), but perform poorly in the tails. To account for this, we suggest stratifying the parameter space into three groups: (1) $X \leq x_{0.10}$, (2) $x_{0.10} < X < x_{0.90}$, and (3) $X \geq x_{0.90}$, where x_r denotes the rth quantile of the MCMC iterates of X. Finally, if resources permit, we recommend the use of preliminary, pilot runs of the MCMC algorithm to establish useful stratifications.

5.2 MULTIPLE CHAINS

It is common for MCMC practitioners to run multiple chains. Having several chains allows for a better visual assessment of burn-in. Further, one hopes that multiple chains, started at a wide range of values, will help uncover the modes of the stationary distribution. Although our method was essentially developed for realizations of a single chain, it can be extended easily to multiple chains. To apply our approach when we have M independent chains, we can treat each of these chains as a batch. Alternatively, we can combine M chains and then split the result into K batches. If a proper subset of the M chains is trapped in a particular region, which is included in the sojourns of the other chains, we expect our method to reject the hypothesis of well mixing.

As an illustrative example, we consider the bivariate mixture model considered in Robert and Casella (2004, p. 476). According to this model, they considered a pair (X, Y) distributed according to a mixture of bivariate normal distribution:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim 0.15\, N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 3 & 1 \\ 1 & 7 \end{pmatrix}\right) + 0.85\, N\left(\begin{pmatrix} 100 \\ 100 \end{pmatrix}, \begin{pmatrix} 5 & 2 \\ 2 & 5 \end{pmatrix}\right),$$

where $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a bivariate normal distribution, with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

Table 4. Empirical and true probabilities of strata for the bivariate normal mixture model

Chain	$X \leq 0$	$0 < X \leq 100$	$X > 100$	$Y \leq 0$	$0 < Y \leq 100$	$Y > 100$
1	0	0	1.0000	0	0	1.0000
2	0.0008	0.6840	0.3152	0	0.5974	0.4026
3	0.0010	0.6862	0.3128	0	0.5926	0.4074
4	0.4772	0.5228	0	0.5014	0.4986	0
5	0.0022	0.6776	0.3202	0	0.5876	0.4124
6	0	0	1.0000	0	0	1.0000
7	0.4990	0.5010	0	0.5064	0.4936	0
8	0.0028	0.6882	0.3090	0	0.5966	0.4034
9	0.5016	0.4984	0	0.4928	0.5072	0
10	0.4864	0.5136	0	0.4962	0.5038	0
True	0.075	0.5	0.425	0.075	0.5	0.425

Using Gibbs sampling, we simulated 30 independent chains, each of length 10,000, with initial values generated from

$$0.5\, N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}100 & 0\\0 & 100\end{pmatrix}\right) + 0.5\, N\left(\begin{pmatrix}100\\100\end{pmatrix}, \begin{pmatrix}100 & 0\\0 & 100\end{pmatrix}\right).$$

We stratified the X space into three groups as $X \leq 0$, $0 < X \leq 100$, and $X > 100$; the Y space was also stratified using the same scheme. Table 4 summarizes the proportions of samples in each stratum for the first 10 chains. In each case, after discarding the initial 5000 iterations, the summaries were computed using 5000 iterations.

We observe that the chains do become trapped in particular regions. We apply our method by treating 30 chains as 30 batches and using the aforementioned stratification scheme. Our test results in indicating poor mixing.

5.3 ASSESSING BURN-IN

Our method can be applied sequentially to assess burn-in of an MCMC implementation. First, select a window size, say w, and reject the first w iterations. We compute the ratio

$$\frac{\hat{V}_2^w}{\hat{V}_1^w} \tag{20}$$

based on the remaining output. Next, reject the first $2w$ iterations and compute $\hat{V}_2^{2w}/\hat{V}_1^{2w}$. We continue this process until the ratios $\hat{V}_2^{Bw}/\hat{V}_1^{Bw}$, $B = 1, 2, \ldots$, are reasonably close to 1. The point at which this occurs is a plausible value for the burn-in time.

As an illustration, consider an AR(1) process of the form

$$X_{t+1} = 0.99 X_t + \epsilon_{t+1}, \tag{21}$$

where the ϵ_{t+1}s are iid $N(0, 0.0001^2)$ random variables. The stationary distribution of this chain is $N(0, 0.0001^2/(1 - 0.99^2))$. We set $X_0 = 10{,}000$, generated 100,000 samples, and partitioned the samples into 20 batches of size 5000. We stratified the X space using two strata: $X > 0$ and $X \leq 0$. We set $w = 100$ and kept the number of batches fixed at $n = 20$ by adjusting the batch sizes accordingly. The plot in Figure 6 shows stabilization of the ratios after discarding around 1100 initial iterations.

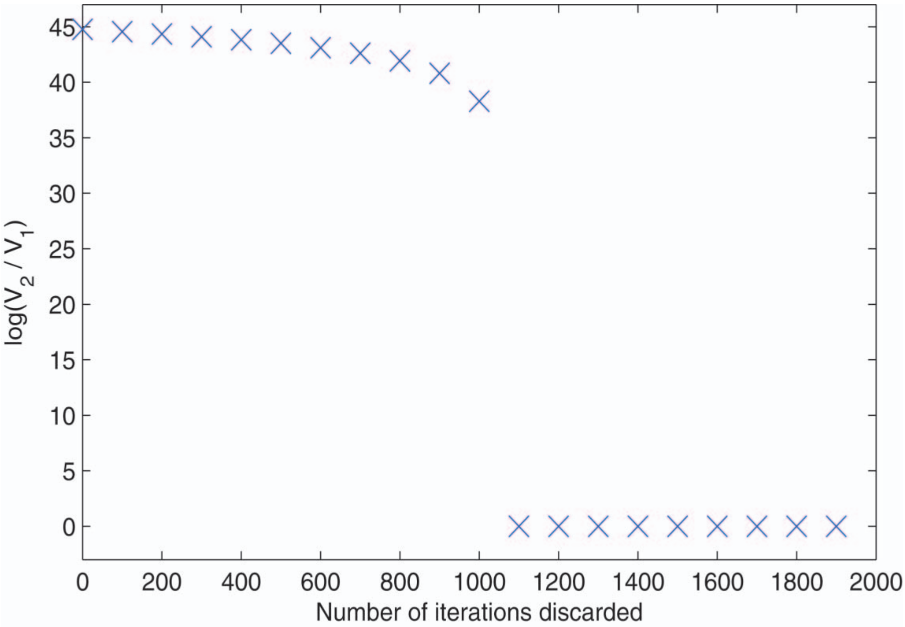

Figure 6. Assessing burn-in for the AR(1) process described in (21) using a window size of 100. The online version of this figure is in color.

5.4 SUMMARY

We have proposed a new diagnostic for assessing convergence to stationarity, as well as mixing, of a Markov chain. The distinguishing feature of our approach is the contrast between properties of two different estimators applied to the same stream of data, that is, the same realization of the chain. The technique can also be applied to multiple chains. We examined properties of the estimators, or rather our bootstrap-based estimates of these properties. These second-order features are more sensitive measures of differences between the estimators' properties.

The key notion used in developing one of the estimators is stratification. Appropriately weighted, stratum-specific results are combined and the results are compared with non-stratified results. The strata can be selected based on the analyst's understanding of aspects of the posterior distribution. Poststratification can also be used.

As with other diagnostics, the user of our approach has several testing parameters to choose. We offered some suggestions along these lines. We also indicated how our diagnostic can be adapted to estimate burn-in times for a chain.

Obtaining convergence diagnostics suitable for high-dimensional parameter spaces remains a serious challenge to users of MCMC. As with other "essentially univariate" diagnostics, one can, of course, apply our procedure to selected univariate parameter streams. However, the appearance of convergence and mixing for many univariate streams need not suggest overall convergence to the multivariate stationary distribution. Quite generally, the combination of many dependent (especially in MCMC) univariate tests is fraught with difficulties.

APPENDIX

The following numbered steps coincide with those used in Section 2.1.

Step 1. Consider the following form of the Ergodic Theorem (e.g., Robert and Casella 2004):

Theorem A.1. Assume $\theta_1, \dots, \theta_N$ are realizations of a Harris recurrent Markov chain with a σ-finite invariant measure π. If $f, h \in L^1(\pi)$ with $\int h(\theta) d\pi(\theta) \neq 0$, then

$$\lim_{N\to\infty} \frac{\frac{1}{N}\sum_{t=1}^{N} f(\theta_t)}{\frac{1}{N}\sum_{t=1}^{N} h(\theta_t)} = \frac{\int f(\theta) d\pi(\theta)}{\int h(\theta) d\pi(\theta)}.$$

Under the conditions of this theorem, for any fixed number of batches, K, $E_1 \overset{a.s.}{\to} E^{\pi}(\theta)$ and $E_2 \overset{a.s.}{\to} E^{\pi}(\theta)$ as $n \to \infty$.

Step 2. We apply the Markov Chain Central Limit Theorem (MCCLT) to obtain the result which is stated in Equation (7). For a detailed statement and various versions of the MCCLT, see Jones et al. (2006) and Robert and Casella (2004). For a Harris recurrent ergodic chain application of MCCLT requires the chain to be geometrically or polynomially ergodic, or α-mixing. Assuming that at least one of these required conditions holds, we apply the MCCLT on $\{Z^{P}_{[k]j}, Z^{\theta}_{[k]j}\}$, and then use the Cramer–Wold device (Basu 2004, chap. 9) to conclude that

$$\sqrt{n}\left(\overline{\mathbf{Z}} - E(\overline{\mathbf{Z}})\right) \overset{D}{\to} MVN(\mathbf{0}, \Sigma_Z) \text{ as } n \to \infty.$$

Step 3. As the batch size n tends to ∞, the batch means are asymptotically uncorrelated under mild conditions. This result is proved by Law and Carson (1979) as stated in the following lemma:

Lemma A.2 (Law and Carson). Let $\{\theta_i\}_{i=1,\dots,N}$ be a sample of size N. Split this sample into K batches of equal size n. Define $C_l = Cov(\theta_t, \theta_{t+l})$, for $l = 1, \dots, N-1$ and $C_j(n) = Cov(\overline{\theta}_{[t]}, \overline{\theta}_{[t+j]})$, for $j = 1, \dots, K-1$. $\overline{\theta}_{[t]}$ denotes the tth batch mean. Let $\rho_j(n) = \frac{C_j(n)}{C_0(n)}$. If $\sum_l^{\infty} C_l < \infty$, then $\rho_j(n) \to 0$ as $n \to \infty$ for all $j = 1, \dots, K-1$.

Step 4. To show that the asymptotic distribution of $n\hat{V}_1$ and $n\hat{V}_2$ coincide, we apply the following form of Slutsky's Theorem.

Theorem A.3 (Bunke and Bunke 1986, result A 4.19). Let $\{\mathbf{X}_n\}$ be a sequence of $q \times s$ random matrices and $\{\mathbf{Y}_n\}$ be a sequence of s-dimensional random vector. If $\mathbf{X}_n \overset{P}{\to} \mathbf{C}$, where $\mathbf{C}$ is a $q \times s$ matrix and $\mathbf{Y}_n \overset{D}{\to} \mathbf{Y}$, then $\mathbf{X}_n\mathbf{Y}_n \overset{D}{\to} \mathbf{CY}$.

Note that

$$n\hat{V}_1 = nd_1\hat{\Sigma}_Z d_1 = n\sum_i\sum_j d_{1i}d_{1j}\hat{\sigma}_{ij}$$

and

$$n\hat{V}_2 = nd_2\hat{\Sigma}_Z d_2 = n\sum_i\sum_j d_{2i}d_{2j}\hat{\sigma}_{ij},$$

where d_{ki} is the ith element of d_k vector and $\hat{\sigma}_{ij}$ is the (i, j)th element of the matrix $\hat{\Sigma}_Z$. Applying the previous theorem, we conclude that the asymptotic (as n tends to ∞) distributions of $n\hat{V}_1$ and $n\hat{V}_2$ are the same.

SUPPLEMENTARY MATERIALS

Illustrative examples: We illustrate the MATLAB codes of our proposed method using a simple autoregressive model of order one. (Codes_Read_Me.pdf)

MATLAB codes: MATLAB codes for the proposed method when the parameter space is stratified into two strata. (strata1.m)

MATLAB codes: MATLAB codes for the proposed method when the parameter space is stratified into three strata. (strata2.m)

ACKNOWLEDGMENTS

We are grateful to Ms. Emily (Lei) Kang for her invaluable help with implementing MCMC code for the Pacific Sea Surface Temperature Model, and the associate editor and two referees for their comments and suggestions. This work was supported in part by the National Science Foundation under grant numbers SES-0437251, DMS-0605041, and ATM-0724403, and by the Office of Naval Research under grant N00014-07-1-0512.

[Received May 2009. Revised April 2011.]

REFERENCES

Basu, A. K. (2004), Measure Theory and Probability, India: Prentice-Hall. [710]

Berliner, L. M., Wikle, C. K., and Cressie, N. (2000), “Long-Lead Prediction of Pacific SSTs via Bayesian Dynamic Modeling,” *Journal of Climate*, 13, 3953–3968. [702]

Bunke, H., and Bunke, O. (1986), *Statistical Inference in Linear Models, Statistical Methods of Model Building*, Chichester, UK: Wiley. [710]

Casella, G., and Berger, R. L. (2002), *Statistical Inference,* (2nd ed.), Stamford, CT: Thomson Learning. [699]

Cowles, M. K., and Carlin, B. P. (1996), “Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review,” *Journal of the American Statistical Association*, 91 (434), 883–904. [695]

Craigmile, P. F., Calder, C. A., Li, H., Paul, R., and Cressie, N. (2009), “Hierarchical Model Building, Fitting, and Checking: A Behind-the-Scenes Look at a Bayesian Analysis of Arsenic Exposure Pathways” (with discussion), *Bayesian Analysis*, 4, 1–62. [699,702]

Damerdji, H. (1994), “Strong Consistency of the Variance Estimator in Steady-State Simulation Output Analysis,” *Mathematics of Operations Research*, 19, 494–512. [699]

Davison, A. C., and Hinkley, D. V. (1997), *Bootstrap Methods and Their Application*, Cambridge: Cambridge University Press. [699]

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. [694,695]

Gelman, A., and Rubin, D. B. (1992), “Inference From Iterative Simulation Using Multiple Sequences” (with discussion), *Statistical Science*, 7, 457–511. [694,695]

Geweke, J. (1992), “Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments,” in *Bayesian Statistics* (Vol. 4), eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford: Oxford University Press, pp. 169–193. [695]

Hartigan, J. A., and Wong, M. A. (1979), "A *k*-Means Clustering Algorithm," *Applied Statistics*, 28, 100–108. [707]

Heidelberger, P., and Welch, P. D. (1983), "Simulation Run Length Control in the Presence of an Initial Transient," *Operations Research*, 31, 1109–1140. [695]

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. [698]

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. [694,695,698,707,710]

Komarèk, A. (2009), "A New R Package for Bayesian Estimation of Multivariate Normal Mixtures Allowing for Selection of the Number of Components and Interval-Censored Data," *Computational Statistics and Data Analysis*, 53, 3932–3947. [707]

Law, A. M., and Carson, J. S. (1979), "A Sequential Procedure for Determining the Length of a Steady-State Simulation," *Operations Research*, 27 (5), 1011–1125. [710]

Lehmann, E. L., and Casella, G. (1998), *Theory of Point Estimation* (2nd ed.), New York: Springer-Verlag. [697]

Liu, C., Liu, J., and Rubin, D. B. (1992), "A Variational Control Variable for Assessing the Convergence of the Gibbs Sampler," in *Proceedings of the American Statistical Association, Statistical Computing Section*, pp. 74–78. [695]

Mykland, P., Tierney, L., and Yu, B. (1995), "Regeneration in Markov Chain Samplers," *Journal of the American Statistical Association*, 90 (429), 233–241. [695]

Raftery, A. E., and Lewis, S. M. (1992), "How Many Iterations in the Gibbs Sampler?," in *Bayesian Statistics* (Vol. 4), eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford: Oxford University Press, pp. 763–773. [695]

Ritter, B. D., and Tanner, M. A. (1992), "Facilitating the Gibbs Sampler: The Gibbs Stopper and the Griddy-Gibbs Sampler," *Journal of the American Statistical Association*, 87, 861–868. [695]

Robert, C. P., and Casella, G. (2004), *Monte Carlo Statistical Methods* (2nd ed.), New York: Springer. [695,707,710]

Roberts, G. O. (1992), "Convergence Diagnostics of the Gibbs Sampler," in *Bayesian Statistics* (Vol. 4), eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford: Oxford University Press, pp. 775–782. [695]

Zellner, A., and Min, C. K. (1995), "Gibbs Sampler Convergence Criteria," *Journal of the American Statistical Association*, 90, 921–927. [695]