

# Detecting Poor Convergence of Posterior Samplers due to Multimodality

Dawn B. Woodard<sup>†</sup>  
Department of Statistical Science  
Duke University  
Durham, NC 27708  
email: [dawn@stat.duke.edu](mailto:dawn@stat.duke.edu)

<sup>†</sup>Dawn B. Woodard is a Ph.D. student in Duke University's Department of Statistical Science, Box 90251, Durham, NC 27708  
email: [dawn@stat.duke.edu](mailto:dawn@stat.duke.edu)

## Abstract

Computation in Bayesian statistical models is often performed using iterative sampling techniques such as Markov chain Monte Carlo (MCMC). The convergence of the sampler to the posterior distribution is typically assessed using a set of standard diagnostics.

We give several examples showing that this approach may be insufficient when the posterior distribution is multimodal—that lack of convergence due to posterior multimodality can be undetected using the standard convergence diagnostics, including the Gelman-Rubin diagnostic that was introduced for exactly this problem. We show that the poor convergence can be detected using a validation technique that was originally proposed for detecting coding errors in MCMC software (Cook, Gelman and Rubin 2006). The validation method can succeed where convergence diagnostics fail, because it evaluates the convergence of the sampling algorithm for many data sets drawn from the model, rather than for the particular data set under consideration.

We first give the example of a mixture model with bimodal posterior distribution where one mode of the posterior has a much smaller “basin

of attraction” than the other. The narrower mode is extremely difficult to detect, both for a Gibbs sampler and for the Gelman-Rubin diagnostic applied to that sampler. Failure to diagnose that there is an undetected narrow mode then leads to overestimation of the posterior variance. We show that the same effect can occur for the popular stochastic search variable selection technique (George and McCulloch 1993), leading to incorrect inferences.

We then argue that the modified validation technique should be widely applied when using sampling methods such as MCMC for computation in models where posterior unimodality is not guaranteed, including stochastic search variable selection.

**Keywords:** convergence diagnostic, Markov chain, Markov chain validation, mixture model, stochastic search variable selection.

## 1. INTRODUCTION

Computation in Bayesian statistical models is typically performed using iterative sampling methods such as Markov chain Monte Carlo (MCMC) or adaptive MCMC techniques. Convergence of the simulation to the posterior distribution is assessed using a set of standard diagnostics. One of these, the Gelman-Rubin diagnostic (Gelman and Rubin 1992), was introduced in recognition of the fact that Markov chains that make only local moves can take a very long time to move between modes of the posterior. This diagnostic requires running multiple chains from “overdispersed” starting points. If the chains have not converged to the same distribution during the period of the simulation, then it is likely that the samplers are not efficiently moving among the modes of the distribution.

It has been observed that if the features of the posterior distribution are completely unknown, then convergence diagnosis for a single Markov chain

is inherently an intractable problem (Clifford 1993). The possible presence of modes with very small “basins of attraction” (Section 3.1) implies that a convergence diagnostic would have to exhaustively enumerate the space in order to guarantee that the chain had not missed any such mode. The use of multiple chains, as in the Gelman-Rubin approach, does not change the intractability problem.

We give examples of statistical models and associated samplers for which the presence of undetected modes causes all of the standard diagnostics to fail to detect lack of convergence, including the Gelman-Rubin diagnostic. We give two examples for which this can occur: a simple mixture model, and the popular stochastic search variable selection method (George and McCulloch 1993).

However, we show that a validation technique originally introduced for detecting coding errors in MCMC implementations (Cook et al. 2006) can effectively detect lack of convergence in cases such as these. Their method can be used to assess the convergence properties of the posterior sampler for many data sets drawn from the model. For this reason, it broadly evaluates the performance of the algorithm rather than its performance for the particular data set being analyzed. This distinction allows the validation method to detect poor convergence properties of a posterior sampler for models where convergence diagnostics fail to do so.

In Section 2 we describe the standard convergence diagnostics, the Cook et al. (2006) validation technique, and modification of that method to detect poor convergence. We then give the mixture model and stochastic search variable selection examples, showing the potential lack of convergence and its effect on inferences, in Section 3. We apply the modified validation technique to de-

test the poor convergence properties for these models in Section 4. Finally, we summarize the results in Section 5.

## 2. DIAGNOSIS AND VALIDATION METHODS

### 2.1 Convergence Diagnostics

First we describe the most commonly used convergence diagnostics. Let the parameter vector be denoted  $\Theta$ . For many problems,  $\Theta$  is finite-dimensional. Nonparametric models are the exception, and theoretically have an infinite-dimensional parameter vector; however, any computerized sampler for  $\Theta$  must use a finite-dimensional approximation or marginalize out the infinite-dimensional component. Therefore we consider  $\Theta$  to be finite-dimensional, and denote the dimension by  $p$ .

Trace (time-series) plots are commonly used to detect lack of convergence, as are autocorrelation plots. We describe two additional common diagnostics:

The Geweke Diagnostic. Consider any real-valued function  $g(\Theta)$  that we are interested in estimating (for example,  $g(\Theta) = \Theta_i$  for  $i \in \{1, \dots, p\}$ ). Geweke (1992) introduces a diagnostic for Markov chains that is computed using the samples of  $g(\Theta)$ ; the diagnostic should be normally distributed if the chain has converged. Typically the diagnostic is applied to several segments of the chain, and the resulting Z-scores are plotted as a function of the first iteration in each segment. If the chain has converged, most of the Z-scores should fall within a 95% band (-2 to 2).

The Gelman-Rubin Diagnostic. Gelman and Rubin (1992) suggest simulating multiple chains as described in the introduction. They construct a *potential scale reduction factor* (PSRF), interpreted as the factor by which the

estimated variance of  $g(\Theta)$  could be reduced if the chain were simulated for more iterations. If the PSRF is much larger than 1, it suggests that the accuracy of the estimate of  $g(\Theta)$  could be much improved by simulating the chain for longer. As suggested by Brooks and Gelman (1998), the PSRF can be applied to segments of the chain, and plotted as a function of the last iteration in the segment. If the chains have converged during the burn-in period, then the PSRF values should be close to one for all of the segments. Brooks and Gelman (1998) also show how to construct a multivariate PSRF that assesses convergence for the entire vector  $\Theta$  simultaneously.

## 2.2 Validation

Next we describe the Cook et al. (2006) validation method and modify it for detecting poor convergence. Cook et al. (2006) propose a method for checking whether a Markov chain sampler is drawing from the correct posterior distribution, and apply this method for detecting the presence of coding errors. Using the notation  $\pi_Z$  to denote the distribution of a random variable  $Z$  and  $\pi_{X|Z=z}$  to denote the conditional distribution of  $X$  given  $Z = z$ , their validation method is composed of the following steps:

1. Draw  $\Theta^*$  from the prior distribution  $\pi_\Theta$ .
2. Draw  $Y^*$  from the data model  $\pi_{Y|\Theta=\Theta^*}$ .
3. Draw a (possibly dependent) sample  $\{\Theta^{(k)} : k = 1, \dots, N\}$  from the posterior distribution  $\pi_{\Theta|Y=Y^*}$  by simulating the Markov chain.

Assume that  $g(\Theta)$  has a continuous posterior distribution function ( $\pi_{g(\Theta)|Y=Y^*}(\{s\}) = 0$  for every  $s \in \mathbb{R}$ ). Cook et al. (2006) calculate the empirical quantile of  $g(\Theta^*)$

in the posterior sample  $\{g(\Theta^{(k)}) : k = 1, \dots, N\}$ :

$$q(\Theta^*, \{\Theta^{(k)}\}) = \frac{1}{N} \sum_{k=1}^N \mathbf{I}(g(\Theta^{(k)}) < g(\Theta^*)).$$

They show that, assuming ergodicity of the Markov chain, the marginal distribution of the random variable  $q(\Theta^*, \{\Theta^{(k)}\})$  approaches a uniform distribution as  $N \rightarrow \infty$  if the Markov chain implementation is correct. Their argument rests on the fact that, conditional on  $Y^*$ , the random variable  $\Theta^*$  is distributed according to  $\pi_{\Theta|Y=Y^*}$  (since  $(\Theta^*, Y^*)$  is a draw from  $\pi_{(\Theta, Y)}$ ), and the fact that the samples  $\{\Theta^{(k)} : k = 1, \dots, N\}$  are also distributed according to  $\pi_{\Theta|Y=Y^*}$ .

Cook et al. (2006) restrict their result to Markov chains. However, the proof requires only guaranteed convergence of ergodic averages of certain bounded measurable functions, a property which must hold for any sampling technique that is to be used for general posterior inference. Therefore the validation method can be applied to any such posterior sampler. For instance, many adaptive MCMC methods are known to have this property, such as the equi-energy sampler on a finite state space (Atchade and Liu 2006).

After  $n$  independent replications of steps 1-3, obtaining empirical quantiles  $\{q_j : j = 1, \dots, n\}$ , Cook et al. (2006) suggest testing uniformity using the function

$$f(\{q_j : j = 1, \dots, n\}) = \sum_{j=1}^n \Phi^{-1}(q_j)^2$$

where  $\Phi$  is the cumulative normal distribution function. As  $N \rightarrow \infty$  the distribution of  $f(\{q_j\})$  approaches a chi-squared distribution with  $n$  degrees of freedom, if the Markov chain implementation is correct. The authors define the p-value  $P(f > f(\{q_j\}))$  where  $f \sim \chi_n^2$ , and use  $n = 20$  replications.

Although not pointed out by Cook et al. (2006), the quantiles will also be uniformly distributed if the sampler simply draws from the prior distribution, ignoring the data. When applying the validation method, this case should be

ruled out. This can be done, for instance, by comparing the posterior samples from the first two replications to ensure that they are not drawn from the same distribution.

Modifying the Validation Method to Detect Poor Convergence. Since the quantiles should be uniformly distributed if  $N$  is large enough and if the code is correct, departure from uniformity when the code is correct indicates lack of convergence. The method by Cook et al. (2006) can therefore be applied in order to detect poor convergence properties of a sampling algorithm for data sets drawn from the model.

However, departure from uniformity due to poor convergence can be very subtle, and can take a different form than departure from uniformity due to a coding error. We therefore modify the validation method of Cook et al. (2006) in order to improve its sensitivity in this context. We take at least  $n = 200$  replications in order to detect more subtle departures from uniformity. We use the original p-value  $P(f > f(\{q_j\}))$ , and additionally the p-value  $P(f < f(\{q_j\}))$ . If the empirical quantiles are clustered close to zero or one, then the first p-value is small, while if the quantiles are clustered towards 0.5 the second p-value is small. We give several examples of the latter in Section 4.

It could be argued that if we make the validation test sensitive enough, we might reject due to lack of convergence that is too small to be important. This could occur; however, the same argument could be made for many convergence diagnostics. For these convergence diagnostics, including the Geweke diagnostic, the standard is that if lack of convergence can be detected while maintaining low type-I error then it is excessive. We will use the validation method to find lack of convergence in situations where convergence diagnostics do not. The interpretation is that these samplers need to be run for longer

than one would conclude by using convergence diagnostics alone.

### 3. EXAMPLES WITH UNDETECTED LACK OF CONVERGENCE

#### 3.1 A Mixture Model

Consider the following mixture model, where  $N_p$  is the  $p$ -dimensional normal density and  $I_p$  is the  $p$ -dimensional identity matrix:

$$(Y_i|\theta) \sim \frac{1}{2}N_p(\theta_i, \phi_1^{-1}I_p) + \frac{1}{2}N_p(\theta_i, \phi_2^{-1}I_p) \quad (1)$$

$$\theta_i \stackrel{\text{iid}}{\sim} N_p(0, \tau^{-1}I_p) \quad (2)$$

for  $i = 1, \dots, n$ . This is a variant of normal mean estimation with the mean modeled as a random effect. Here the model includes two mixture components with different variances; this might be used when the data is known to be either high-variance or low-variance depending on the value of an unobserved environmental factor with two levels.

We take  $\tau, \phi_1, \phi_2 > 0$  to be fixed for simplicity of analysis, although in practice they would typically be unknown and consequently assigned prior distributions. With these parameters fixed the posterior distribution of  $\theta_i$  is a mixture of normals, and depends only on  $y_i$ :

$$\begin{aligned} (\theta_i|Y = y) &\sim q_i N_p(y_i[\tau + \phi_1]^{-1}\phi_1, [\tau + \phi_1]^{-1}I_p) \\ &+ (1 - q_i) N_p(y_i[\tau + \phi_2]^{-1}\phi_2, [\tau + \phi_2]^{-1}I_p) \end{aligned} \quad (3)$$

for a particular value of  $q_i \in (0, 1)$ . Notice that the covariance matrices of the two mixture components are proportional. If  $\phi_1 \neq \phi_2$  then one of the normal mixture components is narrower than the other in each dimension.

Overlooking the conjugacy, one might obtain samples from the posterior by adding a latent parameter indicating the mixture component and using a

Gibbs sampler. Call the latent component indicator for the  $i$ th data point  $Z_i$ ; by defining  $\Pr(Z_i = 0) = \Pr(Z_i = 1) = 0.5$ ,  $(Y_i|\theta, Z_i = 1) \sim N_p(\theta_i, \phi_1^{-1}\mathbf{I}_p)$ , and  $(Y_i|\theta, Z_i = 0) \sim N_p(\theta_i, \phi_2^{-1}\mathbf{I}_p)$  we obtain the marginal model given in (1). Then the distribution of  $\theta_i$  conditional on  $Y$  and  $Z$  is

$$(\theta_i|Y = y, Z_i = 1, Z_{[-i]}) \sim N_p(y_i[\tau + \phi_1]^{-1}\phi_1, [\tau + \phi_1]^{-1}\mathbf{I}_p)$$

$$(\theta_i|Y = y, Z_i = 0, Z_{[-i]}) \sim N_p(y_i[\tau + \phi_2]^{-1}\phi_2, [\tau + \phi_2]^{-1}\mathbf{I}_p)$$

where  $Z_{[-i]}$  is all of the components of  $Z$  excluding the  $i$ th. The conditional distribution of  $Z_i$  given  $Y$  and  $\theta$  is

$$\Pr(Z_i|\theta, Y = y) \propto 1(Z_i = 1)N_p(y_i; \theta_i, \phi_1^{-1}\mathbf{I}_p) + 1(Z_i = 0)N_p(y_i; \theta_i, \phi_2^{-1}\mathbf{I}_p).$$

The Gibbs sampler alternates draws from the distribution of  $(\theta_i|Y, Z)$  for each  $i$  with draws from the distribution of  $(Z_i|\theta, Y)$  for each  $i$ .

In order to apply the Gelman-Rubin diagnostic, one can apply EM to search for modes of the posterior distribution of  $\theta$  and construct the mixture of  $t$  distributions as described in Section 2.1. The starting values for EM can be drawn from the prior distribution of  $\theta$  so that they are well-dispersed. For each Gibbs chain, the initial value for  $\theta$  is drawn from the mixture of  $t$  distributions, and the Gibbs sampler then starts with a draw from the conditional distribution of  $(Z_i|\theta, Y)$  for each  $i$ .

Unfortunately, the Gibbs sampler described above converges very slowly in some situations, and the lack of convergence can be undetected by the Gelman-Rubin diagnostic as well as the other convergence diagnostics.

Simulation example. To illustrate this behavior, take  $n = 1$ ,  $\phi_1 = 1$ ,  $\phi_2 = 10000$ ,  $\tau = 1$ , and  $p = 8$ . In order to generate data that is typical under the model,  $\theta_1$  is sampled from the prior given in (2), and the data  $Y_1$  is

simulated from the data model given in (1), conditional on  $\theta_1$ . Gibbs sampling is then applied to obtain samples from the posterior. We used 20 EM runs and constructed the  $t$  mixture for the Gelman-Rubin diagnostic, using one degree of freedom to ensure overdispersion. The initial values for each of ten Gibbs chains were then drawn from this mixture, and the chains were simulated for an initial burn-in period of 1000 iterations plus a sampling period of 10000 iterations.

For data sets generated as described above, the EM algorithm consistently finds only a single mode of the posterior distribution, namely that corresponding to the mixture component  $Z_1 = 1$ . The convergence diagnostics then do not detect any lack of convergence, including the Gelman-Rubin diagnostic. These diagnostics are illustrated in Figure 1 for a typical data set, where  $\theta_{11}$  denotes the first element of the vector  $\theta_1$ . A time series plot of  $\theta_{11}$  from the first chain does not show any trend in the mean or variability of the samples. The autocorrelation plot of the same samples shows negligible autocorrelation. The Geweke-Brooks plot of the same samples has all of the Z-scores within the 95% confidence band. The Gelman-Rubin-Brooks plot of  $\theta_{11}$ , which uses all ten chains, shows that the potential scale reduction factor approaches one quickly and stays close to one for the remaining iterations of the chains. Diagnostic plots for the other elements of the vector  $\theta_1$  are similar.

A statistical practitioner in this situation would typically be satisfied that the chains have converged, and perform inferences on  $\theta_1$  based on the posterior samples. However, these chains are far from having converged, and some of the resulting inferences are incorrect. None of the chains ever leaves the wide mixture component given by  $Z_1 = 1$ , so the (Monte Carlo) estimated probability that  $Z_1 = 0$  using the samples from the Gibbs chains is zero.

However, analytical calculation shows that the true posterior probability that  $Z_1 = 0$  well above zero (the exact value depends on  $y_1$ ).

This lack of convergence affects the posterior estimate of the covariance matrix of  $(\theta_1|Y_1)$ . The Monte Carlo estimate of this covariance matrix using the Gibbs samples is close to  $[\tau + \phi_1]^{-1}\mathbf{I}_p$ . However, the true posterior covariance matrix has substantially smaller diagonal entries than this matrix, since the narrow mixture component given by  $Z_1 = 0$  has covariance matrix  $[\tau + \phi_2]^{-1}\mathbf{I}_p$  where  $\phi_2$  is much larger than  $\phi_1$ .

The failure of EM to detect the mode corresponding to the mixture component  $Z_1 = 0$  is due to the fact that the basin of attraction of this mode under EM is much smaller than that of the mode corresponding to  $Z_1 = 1$ . By basin of attraction of a particular mode under EM we mean the subset of the parameter space for which EM converges to that mode. Using other mode-finding algorithms the basin of attraction of the mode with  $Z_1 = 0$  is still likely to be much smaller than that of the mode with  $Z_1 = 1$ . The other mode-finding algorithms would then probably also perform poorly in this context, finding only the mode of the diffuse mixture component.

Numerous other convergence diagnostics have also been proposed (see Cowles and Carlin (1996) for a review), all of which assess convergence of the Markov chain algorithm to the particular posterior distribution of interest. We applied some of these diagnostics to the example of this section, which also failed to detect the lack of convergence. Due to the difficulty of detecting the narrow mixture component it is unlikely that any convergence diagnostic could be constructed that would succeed in this context.

### 3.2 Stochastic Search Variable Selection

Stochastic search variable selection (SSVS) is a method for covariate selection in linear regression, introduced by George and McCulloch (1993). They take the standard linear regression model:

$$Y_i|\beta \stackrel{\text{iid}}{\sim} N(X_i'\beta, \sigma^2) \tag{4}$$

and use the prior distribution for  $\beta$  given by

$$\beta_j|\gamma_j \stackrel{\text{iid.}}{\sim} (1 - \gamma_j)N(0, \tau_j^2) + \gamma_jN(0, c_j^2\tau_j^2) \quad j = 1, \dots, J \tag{5}$$

where  $\gamma_j \in \{0, 1\}$ . The quantities  $c_j$  and  $\tau_j$  are taken to be fixed positive values such that  $\tau_j$  is small and  $c_j$  is much larger than one for each  $j$ . This yields the interpretation that if  $\gamma_j = 0$ ,  $\beta_j$  is approximately zero, so the components for which  $\gamma_j = 1$  correspond to the important coefficients in the model. George and McCulloch (1993) also assign a prior distribution to  $\sigma^2$ ; however, to simplify application of the Gelman-Rubin diagnostic we take  $\sigma^2$  to be fixed.

George and McCulloch (1993) note that a variety of prior distributions are possible for the set of  $\gamma_j$ . One possibility is:

$$\gamma_j \stackrel{\text{iid}}{\sim} \text{Bern}(p)$$

for  $p$  fixed.

Chipman (1996) proposes the following alternative prior distribution for the set of  $\gamma_j$ . If one of the covariates in the model is a factor with several levels, then several of the  $\beta_j$  might correspond to the levels of this single covariate. In this situation, a prior distribution could be used that sets the  $\gamma_j$  corresponding to this covariate to be equal to each other. This can be described by introducing a constant  $Z_j \in \{1, \dots, K\}$  that indicates with which covariate  $\beta_j$  is associated, where  $K \leq J$  is the number of covariates. Then one could

assign the following prior for  $\gamma$ :

$$\begin{aligned} \gamma_j &= \alpha_{Z_j} \quad \forall j \\ \alpha_k &\stackrel{\text{iid}}{\sim} \text{Bern}(p) \quad \text{for } k = 1, \dots, K. \end{aligned} \tag{6}$$

Using the data model (4) and the prior specification given by (5) and (6), a straightforward Gibbs sampler can be employed, sampling in turn from the conditional posterior distributions of  $(\gamma|\beta, Y)$  and  $(\beta|\gamma, Y)$ . The forms of these conditional posterior distributions are detailed in George and McCulloch (1993).

In order to apply the Gelman-Rubin diagnostic, one can use EM to find modes of the posterior distribution of  $\beta$ . To obtain starting values for EM, a reasonable approach is to use draws from the posterior distribution of  $\beta$  under the standard Bayesian regression model (which has the prior  $\beta_j \sim N(0, c_j^2 \tau_j^2)$  instead of the mixture prior (5)).

George and McCulloch note that the Gibbs sampler described above can mix very slowly, moving rarely between the mixture components  $\gamma_j = 0$  and  $\gamma_j = 1$ . For the prior distribution (6), this corresponds to the sampler rarely switching between  $\alpha_k = 0$  and  $\alpha_k = 1$ . More troubling than this potential lack of convergence, we will show that the lack of convergence can be difficult to detect, so that a practitioner might mistakenly use samples from an unconverged chain for inference. We illustrate this behavior as follows.

Simulation example. We take a single covariate with ten levels, so that  $J = 10$ ,  $K = 1$ ,  $Z_j = 1$  for each  $j$ , and the ten  $\beta_j$  parameters correspond to the ten levels. We set  $\tau_j^2 = 0.0001$  and  $c_j^2 = 10000$  for all  $j$ . We also take  $\sigma^2 = 100$  so that the data has high variance relative to the variance of  $\beta$ , and is only weakly informative about the value of  $\beta$ . We use ten data points, one

for each level of the covariate, so that  $X_i = (0, \dots, 0, 1, 0, \dots, 0)$  is the vector of length  $J$  with all zeros except in the  $i$ th position. We also set  $p = 0.5$ .

In order to obtain data that is typical under our model, we sample  $\gamma$  and then  $\beta|\gamma$  from their prior distributions (6) and (5), then sample  $(Y_i|\beta)$  for  $i = 1, \dots, 10$  from the data model (4). To run each Gibbs chain we draw an initial value for  $\beta$  from the mixture of  $t$  distributions centered at the modes found by the EM algorithm, and run the chain for a burn-in period of 1000 iterations followed by a sampling period of 10000 iterations. We use 20 independent EM runs, construct the  $t$  mixture having one degree of freedom, then run ten Gibbs chains.

For data sets generated in the above fashion, the EM algorithm consistently finds only one mode, namely that corresponding to  $\alpha_1 = 1$ . Analogously to the example in Section 3.1, this is due to the fact that the mixture component corresponding to  $\alpha_1 = 0$  is much narrower (smaller variance) than that corresponding to  $\alpha_1 = 1$ . The  $t$  mixture then only has one component, centered at the mode corresponding to  $\alpha_1 = 1$ .

The Gibbs chains with initial values drawn from this  $t$  distribution then also consistently fail to discover the mixture component corresponding to  $\alpha_1 = 0$ . As a result, the Monte Carlo estimate of the posterior probability that  $\alpha_1 = 0$  is zero. However, this is incorrect; since the data is only weakly informative about the value of  $\beta$ , the true posterior probability that  $\alpha_1 = 0$  is close to its prior probability 0.5.

Convergence diagnostics based on the Gibbs chains do not detect the lack of convergence. Just as in the example of Section 3.1, time series, autocorrelation, Geweke-Brooks, and Gelman-Rubin-Brooks plots all appear satisfactory. The p-values associated with the Geweke diagnostic are large, and the Gelman-

Rubin potential scale reduction factor is close to one.

#### 4. DETECTING LACK OF CONVERGENCE USING VALIDATION

For the examples described in Sections 3.1 and 3.2, it is straightforward to draw from the prior distribution and from the data model. The validation method is then easily applied. We modified the BayesValidate library which is provided by Cook et al. (2006) to run their method in the statistical software R. Code for the examples given here is available on the website <http://people.orie.cornell.edu/woodard>.

For each chain simulated during the validation procedure, we verified in an automated fashion that the convergence diagnostics passed. In order to check that the autocorrelation was small, we verified that the effective sample size of each parameter was at least one-tenth of the number of samples. We also checked that the p-value associated with the Geweke diagnostic (after Bonferroni adjustment) was at least 0.01 for each parameter, and that the Gelman-Rubin multivariate potential scale reduction factor was less than 1.2.

We initially ran the validation method on a number of mixture models for which the likelihood is invariant to certain changes in the labeling of the parameters. These models included, for instance, several nonparametric Dirichlet process mixture models (Muller, Erkanli and West 1996; Ishwaran and Zarepour 2000). This choice was due to the well-known multimodal nature of these models and poor mixing properties of their Gibbs samplers. However, for these examples the validation method did not find lack of convergence for functions of the parameters that did not depend on the labeling. This result is consistent with the observation of Geweke (2007) that poor mixing of Gibbs sampling in the context of label-invariance does not affect inferences on such

functions.

However, for the examples described in this paper, the validation test rejects due to lack of convergence: histograms of the quantiles exhibit non-uniformity, and tests of uniformity are rejected. These results are consistent across independent runs. Unlike label-invariant mixture models, for which the modes are identical up to labeling, the mixture model examples given here have the property that one of the modes has a much smaller basin of attraction than another, and thus is very difficult to discover.

For the example of Section 3.1, histograms of the quantiles for the first and second elements of the vector  $\theta_1$  are shown in Figure 2. Rather than being flat, their distribution has a slight inverted-U shape. While this is a mild visual effect, a test of uniformity gives an extremely small p-value: the left-tail chisquared p-value described in Section 2.2 is  $10^{-6}$  after Bonferroni adjustment. As shown in Section 3.2, the variance of the elements of the vector  $\theta_1$  is overestimated in the Markov chain samples. The draws  $\theta_1^*$  (defined in Section 2.2) are then closer to the center of the estimated posterior distributions than they would be if the chains had converged. This effect leads to the inverted-U shape of the quantile distribution and to the small chisquared statistic.

The quantiles for the SSVS example are shown in Figure 3; they are visually highly nonuniform. Correspondingly, the p-value for the left-tail chisquared test is  $10^{-13}$ . The histogram shows a cluster of quantiles near 0.5; once again, this is due to overestimation of the posterior variance.

## 5. CONCLUSIONS

We have shown that by modifying the validation method of Cook et al. (2006), originally introduced for detecting coding errors in MCMC, one can effectively

detect poor convergence of a posterior sampler due to multimodality. The resulting method is not a convergence diagnostic, since it evaluates the convergence of the sampler for data sets drawn from the model rather than for the particular data set of interest, broadly evaluating the algorithm rather than a particular application of the algorithm. For this reason, it detects poor convergence properties of posterior samplers in situations where all of the standard convergence diagnostics fail to do so.

We have illustrated this behavior using a mixture model with associated Gibbs sampler. We showed that inherent poor convergence properties, undetected by convergence diagnostics, lead to invalid inferences. The poor convergence properties of the algorithm, and the failure of the convergence diagnostics to detect the problem, are both due to the fact that one of the modes of the posterior distribution has a much smaller basin of attraction than the other mode and is consequently extremely difficult to find. The validation method, evaluating the convergence of the algorithm for many typical data sets, is able to detect the convergence problem.

The same effect can occur for the popular stochastic search variable selection method of George and McCulloch (1993), also leading to incorrect inferences. Once again, the validation method discovers the problem when convergence diagnostics fail to do so.

The validation method by Cook et al. (2006) is trivial to apply in any situation where one can sample from the prior distribution and from the data model. Despite this simplicity, when modified as we have described it can be effective in detecting poor convergence of a sampling technique due to multimodality of the posterior distribution. It is applicable to Markov chain methods, adaptive MCMC methods, and other types of iterative samplers.

Since the validation method can detect both coding errors in the sampler implementation and poor convergence properties of the sampling algorithm, we recommend that it be routinely applied along with convergence diagnostics. As we have shown, its use is particularly important for models that do not have guaranteed posterior unimodality, including stochastic search variable selection.

## REFERENCES

- Atchade, Y., and Liu, J. S. (2006), “Discussion of ‘Equi-energy sampler’ by Kou, Zhou, and Wong,” *Annals of Statistics*, 34, 1620–1628.
- Brooks, S. P., and Gelman, A. (1998), “General methods for monitoring convergence of iterative simulations,” *Journal of Computational and Graphical Statistics*, 7, 434–455.
- Chipman, H. (1996), “Bayesian variable selection with related predictors,” *The Canadian Journal of Statistics*, 24, 17–36.
- Clifford, P. (1993), “Discussion on the meeting on the Gibbs sampler and other Markov chain Monte Carlo methods,” *J. of the Royal Statistical Society, Series B*, 55, 53–54.
- Cook, S. R., Gelman, A., and Rubin, D. B. (2006), “Validation of software for Bayesian models using posterior quantiles,” *Journal of Computational and Graphical Statistics*, 15, 675–692.
- 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.

- Gelman, A., and Rubin, D. B. (1992), “Inference from iterative simulation using multiple sequences,” *Statistical Science*, 7, 457–472.
- George, E. I., and McCulloch, R. E. (1993), “Variable selection via Gibbs sampling,” *Journal of the American Statistical Association*, 88, 881–889.
- Geweke, J. (1992), “Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments,” in *Bayesian Statistics 4*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford: Oxford University Press.
- Geweke, J. (2007), “Interpretation and inference in mixture models: simple MCMC works,” *Computational Statistics and Data Analysis*, 51, 3529–3550.
- Ishwaran, H., and Zarepour, M. (2000), “Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models,” *Biometrika*, 87(2), 371–390.
- Muller, P., Erkanli, A., and West, M. (1996), “Bayesian Curve Fitting Using multivariate normal mixtures,” *Biometrika*, 83(1), 67–79.

## List of Figures

Figure 1. The standard convergence diagnostics fail to detect lack of convergence for the mixture model. Top left: time series plot for  $\theta_{11}$  (first chain). Bottom left: autocorrelation plot for  $\theta_{11}$  (first chain). Top right: Geweke-Brooks plot for  $\theta_{11}$  (first chain). Bottom right: Gelman-Rubin-Brooks plot for  $\theta_{11}$  (all ten chains).

Figure 2. Histograms of the quantiles for the first and second elements of the vector  $\theta_1$  in the mixture model example.

Figure 3. Histograms of the quantiles for  $\beta_1$  and  $\beta_2$  in the SSVS example.

Figure 1:

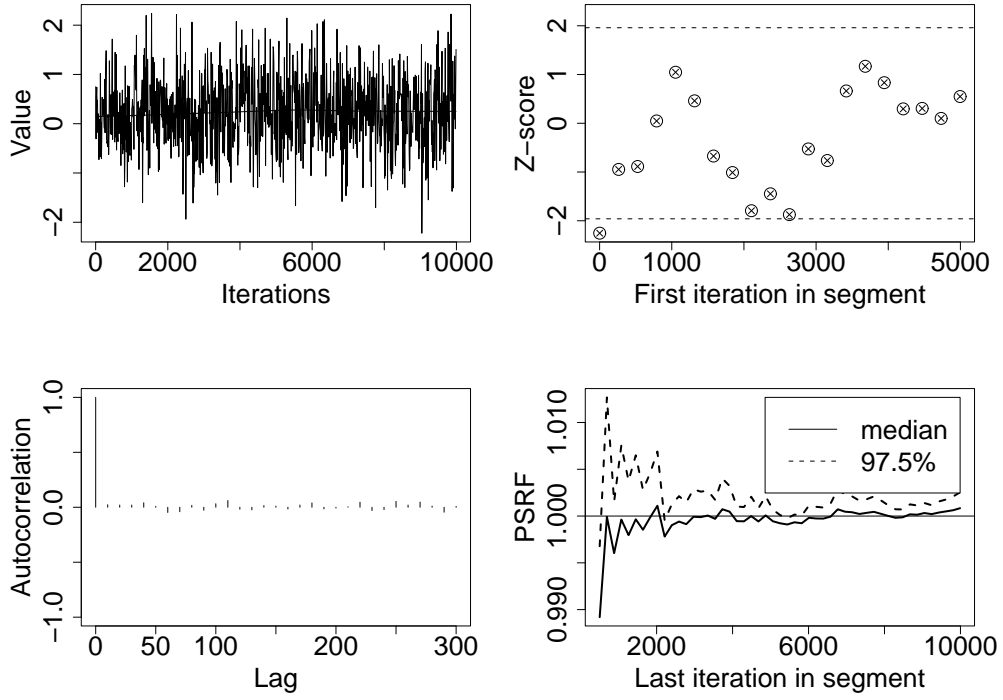


Figure 2:

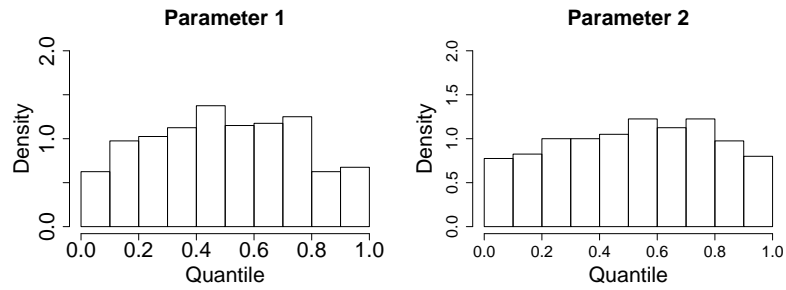


Figure 3:

