

Output Assessment for Monte Carlo Simulations via the Score Statistic

Y. Fan[†], S.P. Brooks[‡] and A. Gelman[§]

Abstract

In this paper, we present several applications of the score statistic in the context of output assessment for Monte Carlo simulations. We begin by observing that the expected value of the score statistic U is zero, and that when the inverse of the information matrix $\mathbf{I} = E(\mathbf{U}\mathbf{U}^T)$ exists, the asymptotic distribution of $\mathbf{U}^T \mathbf{I}^{-1} \mathbf{U}$ is χ^2 . Thus we may monitor the sample mean of this statistic throughout a simulation as a means to determine whether or not the simulation has been run for a sufficiently long time.

We also demonstrate a second convergence assessment method based upon the idea of path sampling, but first demonstrate how the score statistic can be used to accurately estimate the stationary density using only a small number of simulated values. These methods provide a powerful suite of tools which can be generically applied when alternatives such as the Rao-Blackwell density estimator are not available. Our second convergence assessment method is based upon these density estimates. By running several replications of the chain, the corresponding estimated densities may be compared to assess how “close” the chains are to one another and to the true stationary distribution. We explain how this may be done using both L^1 and L^2 distance measures.

We first illustrate these new methods via the analysis of MCMC output arising from some simulated examples, emphasising the advantages of our methods over existing diagnostics. We further illustrate the utility of our methods with three examples: analysing a set of real time series data, a collection of censored survival data, and bivariate Normal data using a model with a non-identified parameter.

Key Words: convergence diagnostics; Markov chain Monte Carlo; density estimation; score statistic; path sampling

1 Introduction

Iterative simulations, especially Markov chain Monte Carlo (MCMC) methods, have been increasingly popular in statistical computation, most notably for drawing simulations from Bayesian posterior distributions, see Brooks (1998a) for example. In addition to any implementational difficulties and computing resources required, iterative simulation presents two problems beyond those of traditional statistical methods. First, when running an *iterative* algorithm, one must decide when to stop the iterations or, more precisely, one must judge how close the algorithm is to convergence after a finite number of iterations. Secondly, MCMC simulation converges to a target *distribution*, rather than a target point. This leads to many practical difficulties, not least of which is how to provide adequate summaries of the sampler output.

Of course, these problems apply quite generally to iterative simulation algorithms, not just to MCMC algorithms. For example Gelman (1992) discusses how importance sampling methods are in fact iterative and, in general, result in draws from the target distribution only in the limit as the number of iterations approaches infinity. One way of seeing this approximate nature of importance sampling is to note that ratio estimates of importance-weighted means, $\sum_{i=1}^n w_i h(\theta_i) / \sum_{i=1}^n w_i$ are unbiased only in the limit as $n \rightarrow \infty$, and that this convergence (as well as more practical issues of the variance of the estimate in a finite sample) depends upon the upper tail of the distribution of the weights w_i . Liu *et al.* (1993) note the duality between this “importance weight infinity” and the “waiting time infinity” of MCMC and rejection sampling.

In this paper we aim to address both aspects of the problem of assessing Monte Carlo output. We show how calculations based upon the classical log score statistic, that is the derivative of the log

[†]Department of Statistics, School of Mathematics, University of New South Wales, NSW, Australia (yanan@maths.unsw.edu.au)

[‡]Statistical Laboratory, CMS, University of Cambridge, Wilberforce Road, CB3 0WB, UK (Steve@statslab.cam.ac.uk)

[§]Andrew Gelman, Department of Statistics, Columbia University, New York, USA. (gelman@stat.columbia.edu)

density, can be used to construct robust and accurate density estimates from Monte Carlo sampler output. These methods can be generically applied when other popular methods based upon Rao-Blackwellisation, for example, cannot. Not only do these estimates provide useful summaries of the marginal distributions of the quantities being sampled, they also facilitate the identification of sample modes for example.

In this paper, we also introduce several new diagnostic techniques based upon the calculation of the classical score statistic. However, before introducing these methods, it is useful to briefly review existing methods to establish a context for these new developments.

Various methods have been proposed for assessing convergence without the analysis of simulation output. Perhaps the most obvious approach is to design the simulation algorithm to produce independent draws directly from the target distribution. Examples include rejection sampling using a proposal function that uniformly dominates the target density; coupling and regeneration methods in MCMC; and the “perfect simulation” method of Propp and Wilson (1996), in cases where it is computationally feasible. In each of these approaches, the time required to wait until the next independent draw is a random variable, which can limit the effectiveness of these methods if the waiting time is too long.

Theoretical (analytic) results bounding the difference between the simulation and target distributions after some specified number of iterations have also been developed. Reasonable results of this type have appeared only for some very simple models. See Cowles and Rosenthal (1998), for example. Though recent work has improved considerably the state of the art in this area, it is still unrealistic to expect this approach to become widely applicable in MCMC simulation except in certain special cases.

Thus, in the absence of practical analytic techniques, we assess convergence by analysing sampler output to assess the mixing properties of the simulation. Probably the most commonly-used convergence assessment techniques make use of the fact that most MCMC algorithms exhibit a random-walk behaviour in which a simulated chain gradually spreads out from its starting point to ergodically cover the space of the target distribution. Convergence occurs when the chain has fully spread to the target distribution, which can be judged in three basic ways. The first is to monitor trends. Given a single MCMC sequence, one can judge mixing by looking for trends in the simulation (Brooks 1998b); unfortunately, such an approach will not necessarily detect lack of convergence of a slowly-moving sequence (Gelman and Rubin, 1992b).

A second approach is to monitor autocorrelation. Efficiency of simulations can be judged by autocorrelations, and this approach can also be used to obtain approximately independent simulation draws (Raftery and Lewis, 1992). This approach however can also be fooled by very slow-moving series and thus is perhaps most effective as a measure of efficiency for an MCMC algorithm for which convergence has already been judged by other means. A third approach is to monitor mixing of sequences directly. Gelman and Rubin (1992a) (and subsequently Brooks and Gelman 1998a) proposed monitoring the mixing of simulated sequences by comparing the variance within each sequence to the total variance of the mixture of the sequences. This is an adaptation of statistical analysis of variance to the standard multiple-sequence approaches in statistical physics (see, e.g., Fosdick, 1959).

Interestingly, the approaches based upon detecting a lack of mixing are ineffective in monitoring convergence of non-Markov-chain iterative simulation methods such as importance sampling, for which successive draws are not nearby in the parameter space. This is another argument in favour of the use of MCMC in preference to other iterative simulation methods. In particular, the autocorrelation or locality of random-walk or state-space algorithms, which is generally perceived as a drawback (since it decreases the efficiency of simulations), is actually an advantage in convergence monitoring.

An alternative group of approaches that do not require random walk-type behaviour are based upon sequential testing of portions of simulation output in order to determine whether or not they could be considered to have been drawn from the same distribution. Methods of this sort sequentially discard an increasing proportion of the early simulated values and divide the remaining observations into three blocks. The observations in the first and third block are then compared and a formal procedure used to test the null hypothesis that the simulated observations are drawn from the same distribution. If the test is rejected then more of the early values are discarded and the testing procedure is repeated. If the test is accepted, then it is assumed that the discarded observations covered the burn-in period and that the remaining observations are all generated from the same (assumed to be the stationary) density. See Geweke (1992) and Heidelberger and Welch (1983), for

example.

Methods based upon functions of the simulation output that are related to the simulation algorithm in a known way, such as importance ratios, acceptance probabilities, transition probabilities, and posterior densities have also been developed. Importance ratios and acceptance probabilities have been useful in approximately evaluating the efficiency of importance sampling (Kong, 1992) and Metropolis algorithms (Gelman *et al* 1996) once convergence has been reached, but they do not seem very powerful in detecting poor convergence if used alone. More effective approaches combine importance ratios with other information, as in the methods of Roberts (1992) and Brooks *et al* (1997). These methods are based upon the comparison of density estimates obtained from different replications using appropriate distance measures. These are particularly powerful convergence assessment techniques but are typically hard to implement and difficult to interpret.

There has been a great deal of interest, for both theoretical and practical reasons, for summary measures based upon the target density function. This paper explores one such class of methods, based upon the score function. As we shall see, we can use some already-known identities from classical statistics and path sampling to develop a new class of convergence diagnostics which are both practical to implement and easy to interpret. We begin with what we call the score function diagnostic in Section 2 which uses the fact that the expected value of the score statistic U is zero with respect to the target distribution. We illustrate the performance of this method via a simulated example. In Section 3, we discuss how techniques based upon ideas from path sampling may be used to provide marginal density estimates for comparison between replications. In Section 4 we extend this idea and describe four separate density estimation techniques and, in Section 5, we discuss the application of these density estimates as convergence diagnostics with an illustrated comparison to existing methods. Finally in Section 6, we introduce three examples to illustrate the utility of our methods before ending with some discussions in Section 7.

2 The score function diagnostic

One approach to detecting lack of convergence is to estimate, using simulation, quantities that have known values under the target distribution. If $\boldsymbol{\theta}$ denotes the parameter vector sampled via iterative simulation, then we can use simulation draws to estimate $\mathbb{E}[U(\boldsymbol{\theta})]$ for any computable function U . Many diagnostic techniques are based upon monitoring functionals which converge to some specific value. However, in general this value is not known and so the resulting diagnostic is rather hard to interpret in that it may have settled to some value, but it is unclear whether or not it is the *true* value, see Roberts (1992), Gelman and Rubin (1992b). Of course, these problems would be removed if we knew what the true expectation of U was under the stationary distribution, and this paper is concerned with trying to find functions, or families of functions, for which this is the case.

One such function is the score function. If $\boldsymbol{\theta} \in E \subseteq \mathbb{R}^K$, and we let $\pi(\boldsymbol{\theta})$ denote the target distribution for the simulations, then we might take

$$U_k(\boldsymbol{\theta}) = \frac{\partial \log \pi(\boldsymbol{\theta})}{\partial \theta_k}, \quad k = 1, \dots, K.$$

Then, assuming the usual regularity conditions (Cox and Hinkley 1974; p.107) for the density π , $\mathbb{E}_\pi[U_k(\boldsymbol{\theta})] = 0$ for all $k = 1, \dots, K$. Further, if we let

$$\mathbf{U}(\boldsymbol{\theta}) = (U_1(\boldsymbol{\theta}), U_2(\boldsymbol{\theta}), \dots, U_K(\boldsymbol{\theta}))$$

and the information matrix

$$\mathbf{I}(\boldsymbol{\theta}) = E(\mathbf{U}(\boldsymbol{\theta})\mathbf{U}(\boldsymbol{\theta})^T)$$

is the variance-covariance matrix of the $U_k(\boldsymbol{\theta})$ s, with the (jk) th element defined as

$$\mathbf{I}_{jk}(\boldsymbol{\theta}) = E(U_j(\boldsymbol{\theta})U_k(\boldsymbol{\theta})) = E\left[\frac{\partial \log \pi(\boldsymbol{\theta})}{\partial \theta_j} \frac{\partial \log \pi(\boldsymbol{\theta})}{\partial \theta_k}\right] = E\left[-\frac{\partial^2 \log \pi(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_k}\right],$$

then, by the central limit theorem $\mathbf{U}(\boldsymbol{\theta})$ has an asymptotic multivariate Normal distribution so that

$$\mathbf{U}(\boldsymbol{\theta}) \sim N(0, \mathbf{I}(\boldsymbol{\theta}))$$

and therefore

$$\mathbf{U}(\boldsymbol{\theta})^T \mathbf{I}(\boldsymbol{\theta})^{-1} \mathbf{U}(\boldsymbol{\theta}) \sim \chi_K^2$$

provided that $\mathbf{I}(\boldsymbol{\theta})$ is non-singular so that the inverse $\mathbf{I}(\boldsymbol{\theta})^{-1}$ exists. (See Dobson 1990; p.50.)

If the density π is not regular (i.e., we are unable to interchange the order of integration and differentiation, as with a Uniform distribution for example, see Cox and Hinkley 1974; p112) we can always reparameterise π in order to restore regularity, as discussed by Cox and Hinkley (1974).

To assess convergence, we might monitor each of these U_k functions for samples $\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \dots$, until they appear to settle to around zero. We might also estimate the standard error of the $U_k(\boldsymbol{\theta})$ from multiple sequences, so as to determine whether or not observed values are “significantly” different from zero. In practice the diagnostic may be best implemented as follows. Given the output from separate chains $j = 1, \dots, J$, let \bar{U}_{kj} denote the mean of the $U_k(\boldsymbol{\theta})$ values for chain j , formed from the second half of the chain. Then, for each k , let μ_k and σ_k denote the empirical mean and standard deviation, respectively, of \bar{U}_{kj} calculated over the J replications.

Clearly, if draws really are from the stationary distribution, then simple manipulations reveal that μ_k is approximately Normal with mean zero and standard deviation σ_k/\sqrt{J} . Thus, we can plot the μ_k values over time (and for each k), together with “error bounds” $\mu_k \pm 2\sigma_k/\sqrt{J}$, which should cover the value zero.

Similarly, if we define

$$X^2 = J \sum_{k=1}^K \left(\frac{\mu_k}{\sigma_k} \right)^2,$$

then $X^2 \sim \chi_K^2$. Thus, we might also plot X^2 against time to gain an overall assessment of the convergence of all K parameters simultaneously. We shall refer to these diagnostics as diagnostics based on the univariate score functions.

In the case where the information matrix $\mathbf{I}(\boldsymbol{\theta})$ is non-singular so that an inverse exist, a more powerful multivariate diagnostic based on $\mathbf{U}(\boldsymbol{\theta})^T \mathbf{I}(\boldsymbol{\theta})^{-1} \mathbf{U}(\boldsymbol{\theta})$ statistic can be used instead of monitoring each of the $U_k(\boldsymbol{\theta})$ s separately. Given the output from chains $j = 1, \dots, J$, let $\bar{\mathbf{U}}_j(\boldsymbol{\theta})$ denote the mean of the $\mathbf{U}(\boldsymbol{\theta})^T \mathbf{I}(\boldsymbol{\theta})^{-1} \mathbf{U}(\boldsymbol{\theta})$ for chain j formed over the second half of the chain. Let μ and σ denote the empirical mean and standard deviation respectively of the $\bar{\mathbf{U}}_j(\boldsymbol{\theta})$ calculated over the J replications. It can be shown that μ is approximately normally distributed with mean K and standard deviation σ/\sqrt{J} . Thus plotting the μ values over time together with error bounds $\mu \pm 2\sigma/\sqrt{J}$ should cover the value K . We refer to these diagnostics as diagnostics based on multivariate score functions.

2.1 Toy Example

The main strength of our diagnostic method is that comparisons are made between the sample output and known quantities of the target distribution, whilst most other existing diagnostics tend to base comparisons solely between sample output.

To illustrate this point, we use a mixture of two bivariate Normal distributions

$$\pi = 0.8N \left[\begin{pmatrix} -2 \\ 6 \end{pmatrix}, \begin{pmatrix} 2 & 0.8 \\ 0.8 & 1 \end{pmatrix} \right] + 0.2N \left[\begin{pmatrix} 4 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 2 \end{pmatrix} \right].$$

Suppose that our sampler was able to sample only from one of the two modes in our target distribution, say, the smaller of the two modes, as was the case in this example. We took 5000 iterations of this Metropolis-Hastings sampler over 5 replications, each replication using the same starting value. Figure 1(a,b) show the outputs from Gelman and Rubin (1992a) diagnostic, and clearly, they indicate convergence. A few of the other well known convergence diagnostics were also tried on these sample outputs, and none were able to spot the lack of convergence.

Figure 1(c,d) show the univariate score diagnostics U_k calculated for the second half of the chains. Clearly while a lack of convergence in the x direction was detected, the diagnostic failed to pick up the lack of convergence in the y direction. Whilst in the multivariate X^2 case, Figure 1e, the diagnostic also indicates a lack of confidence in the convergence of the chain. The fact that the univariate score diagnostic failed to detect a lack of convergence highlights one drawback of this method, in that if the chain was stuck sampling in some (nearly) symmetric part of the target distribution, the diagnostic would fail to reveal the problem. This would happen, for example, in the case where the target

distribution is $N(0, 1)$ and we may be sampling incorrectly from $N(0, 2)$. We would expect that as the complexity of the target distribution increases with increasing dimension, it would be relatively rare for such cases to occur, at least not in all directions, as illustrated in our current example, the lack in convergence was detected in the x direction.

Figure 1f, where the multivariate diagnostic $U^T IU$ was calculated for the second half of the chains at every 100 iterations, shows that the diagnostic based on the multivariate score function U overcomes the problem discussed above. In Figure 1f, the expected value for the multivariate score diagnostic does not lie within the 95% confidence interval of the diagnostic. Thus we would recommend the multivariate score diagnostic to be used whenever the inverse can be found for the information matrix.

[Figure 1 about here.]

We provide further illustration of these methods in Section 6 but first introduce a second group of diagnostic techniques based upon the idea of path sampling.

3 Path sampling

Gelman and Meng (1998) review how, given two unnormalised densities $p_0(\theta)$ and $p_1(\theta)$ on the same support, with normalisation constants $c(0)$ and $c(1)$ respectively, the log of the ratio of these normalisation constants can be written as

$$\lambda = \log \left(\frac{c(1)}{c(0)} \right) = \mathbb{E}_{p(\theta, \tau)} \left(\frac{U(\theta, \tau)}{p(\tau)} \right),$$

where $p(\theta|\tau) = p_0^{1-\tau}(\theta)p_1^\tau(\theta)$, $\tau \in [0, 1]$ describes a geometric path between p_0 and p_1 ,

$$U(\theta, \tau) = \frac{\partial}{\partial \tau} \log p(\theta|\tau)$$

is the score function as discussed above, $p(\theta, \tau) = p(\theta|\tau)p(\tau)$, and $p(\tau)$ is an arbitrary prior density for $\tau \in [0, 1]$.

Thus, by drawing samples $\{(\theta^i, \tau^i) : i = 1, \dots, n\}$ from $p(\theta, \tau)$ we can estimate λ by

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^n \frac{U(\theta^i, \tau^i)}{p(\tau^i)}.$$

In practice, the sampling is most easily performed either by first sampling τ from $p(\tau)$, commonly a standard uniform, and then drawing θ from $p(\theta|\tau)$; or by specifying $p(\theta, \tau)$ directly and drawing (θ, τ) jointly using a Metropolis-Hastings algorithm. In the latter case, the marginal $p(\tau)$ may not itself be known, and λ will be estimated using numerical integration over the range of the sampled points, as described in Section 4 below.

This can be extended to general dimensions as follows. Suppose $\mathbf{t} \in \mathbb{R}^K$, and that we are interested in estimating $\lambda = \log [c(\mathbf{t}_1)/c(\mathbf{t}_0)]$, where $c(\mathbf{t}_i)$ is the normalisation constant for $p(\theta|\mathbf{t}_i)$. Then, if we let $\{\mathbf{t}(\tau) : \tau \in [0, 1]\}$ denote a continuous path from $\mathbf{t}_1 = \mathbf{t}_0$, and

$$U_k(\theta, \mathbf{t}) = \frac{\partial}{\partial t_k} \log p(\theta|\mathbf{t}); \quad \dot{t}_k(\tau) = \frac{\partial t_k(\tau)}{\partial \tau} \quad k = 1, \dots, K,$$

then

$$\lambda = \int_0^1 \int p(\theta|\mathbf{t}(\tau)) \sum_{k=1}^K U_k(\theta, \mathbf{t}(\tau)) \dot{t}_k(\tau) d\theta d\tau.$$

As before, this may be approximated by drawing observations $\{(\theta^i, \mathbf{t}(\tau^i)) : i = 1, \dots, n\}$, and setting

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K \dot{t}_k(\tau^i) U_k(\theta^i, \mathbf{t}(\tau^i)).$$

In order to construct a diagnostic, we will use this final result to construct estimates of the marginal distribution of parameters of interest. These estimates can be compared between replications, with any discrepancies indicating a lack of convergence.

Suppose that we have some target distribution $\pi(\boldsymbol{\theta})$, from which we have samples $\boldsymbol{\theta}^1, \dots, \boldsymbol{\theta}^n$, and that we are interested in estimating the marginal distribution for θ_k , which we shall denote by $\pi_k(\theta_k)$. Then we define

$$U_k(\boldsymbol{\theta}) = \frac{\partial}{\partial \theta_k} \log \pi(\boldsymbol{\theta}).$$

If $\pi(\boldsymbol{\theta})$ is known only up to some normalisation constant so that we have only the functional form $\tilde{\pi}(\boldsymbol{\theta})$, then we also have that

$$U_k(\boldsymbol{\theta}) = \frac{\partial}{\partial \theta_k} \log \tilde{\pi}(\boldsymbol{\theta}).$$

We may then estimate the marginal distribution from the sample by constructing a path sampling estimate as follows. Let $\lambda_k(\theta_k) = \log \tilde{\pi}_k(\theta_k)$ denote the log of the unnormalised marginal posterior distribution up to an arbitrary constant, and order the observations $\theta_k^{(1)} < \theta_k^{(2)} < \dots < \theta_k^{(m_k)}$, $m_k \leq n$, ignoring any duplicates which may arise in the context of the Metropolis algorithm for example. Now, suppose that there are n_k^i repeated observations for each $\theta_k^{(i)}$, $i = 1, \dots, m_k$ and that they occur at times $\tau_i^k(1), \tau_i^k(2), \dots, \tau_i^k(n_k^i)$, then define $\bar{U}_k(i)$ to be the mean of the $U_k(\boldsymbol{\theta})$ values for each of these replications, so that

$$\bar{U}_k(i) = \frac{1}{n_k^i} \sum_{j=1}^{n_k^i} U_k(\boldsymbol{\theta}_{\tau_i^k(j)}).$$

Then, we obtain the following lemma.

Lemma 3.1

$$\frac{\partial}{\partial \theta_k} \lambda_k(\theta_k) = \mathbb{E}(U_k(\boldsymbol{\theta})),$$

where the expectation is taken with respect to all of the other $K - 1$ components of $\boldsymbol{\theta}$.

Proof: See Appendix.

We may construct an estimate of $\pi_k(\theta_k)$ by using the $\bar{U}_k(i)$ either as empirical estimates of the gradient of $\lambda_k = \log \tilde{\pi}_k(\theta_k)$ at $\theta_k^{(i)}$, thus obtaining a piecewise exponential estimate of $\tilde{\pi}_k(\theta_k)$ or, by recalling that

$$\frac{\partial}{\partial \theta_k} \lambda_k(\theta_k) = \frac{\partial}{\partial \theta_k} \log \tilde{\pi}_k(\theta_k) = \frac{1}{\tilde{\pi}_k(\theta_k)} \frac{\partial}{\partial \theta_k} \tilde{\pi}_k(\theta_k) \quad (1)$$

as empirical estimates of the gradient of $\tilde{\pi}_k(\theta_k)$ directly. Alternatively, we can use a simple stepwise approximation rather than linear on the log-scale. We discuss all of these approaches in the next section.

4 Estimating the distribution function

In this section, we discuss various methods for estimating the marginal density function of some parameter θ_k , given samples from the full joint distribution, using path sampling.

4.1 Stepwise linear estimation

The basic idea here is that using the $\bar{U}_k(i)$, we can estimate the gradient of π_k at each of the points in the sample. We can then use these points to form a stepwise linear approximation to π_k , by arbitrarily setting the approximating function to be zero at $\theta_k^{(1)}$ and then using the gradient estimate at that point to obtain the value at the next point in the sample. We then take the value at the first point as being an approximation to π_k in the region between the two points. We can repeat this procedure to obtain a sequence of lines defined between successive $\theta_k^{(i)}$.

This approximation, with its arbitrary scale, can be used to obtain an estimate of the distribution function of π_k , by analytically integrating the stepwise linear approximation. The resulting estimate

of the distribution function is then normalised by setting its value at $\theta_k^{(m)}$ to be 1. Thus, we obtain a normalised estimate of the density π_k , which we can use for comparison. We proceed as follows.

By Lemma 3.1, the \bar{U}_k provide an empirical estimate of the gradient of λ_k at points in the sample, and may therefore be used to construct a stepwise approximation to $\log \pi_k$ and therefore π_k . We construct this piecewise linear approximation by arbitrarily setting $\hat{\lambda}_k(\theta_k^{(1)}) = 0$ and, for $i = 2, \dots, m$, define

$$\hat{\lambda}_k(\theta_k^{(i)}) = \hat{\lambda}_k(\theta_k^{(i-1)}) + (\theta_k^{(i)} - \theta_k^{(i-1)}) \times (\bar{U}_k(i) + \bar{U}_k(i-1)) / 2. \quad (2)$$

We thus have an unscaled approximation to the value of $\log \pi_k$ at each of the sample points, which can be used to construct a stepwise linear approximation to π_k for all points $\theta \in [\theta_k^{(1)}, \theta_k^{(m)}]$, given by

$$p_k(\theta) = \exp(\hat{\lambda}_k(\theta_k^{(i)})), \quad \text{for } \theta \in [\theta_k^{(i)}, \theta_k^{(i+1)}]. \quad (3)$$

Having obtained this stepwise linear approximation to π_k , we can then estimate the corresponding distribution function, by integrating p_k within the range $[\theta_k^{(1)}, \theta_k^{(m)}]$. This estimate is then normalised by dividing by the integral over the entire range so that the estimate of the distribution function becomes 1 at the last observed data point. The following lemma provides us with the normalisation constant that we require.

Lemma 4.1 *Given a piecewise linear function,*

$$p_1(t) = \begin{cases} a_i + b_i t & t \in (y_i, y_{i+1}] \\ 0 & t > y_m \text{ or } t < y_1 \end{cases}, \quad i = 1, \dots, m-1, \quad \text{then}$$

$$P_1(t) = \int_{-\infty}^t p_1(\tau) d\tau = \begin{cases} 0 & t \in (-\infty, y_1] \\ a_i(t - y_i) + b_i(t^2 - y_i^2)/2 + \sum_{j=1}^{i-1} a_j(y_{j+1} - y_j) & t \in (y_i, y_{i+1}], i = 1, \dots, m-1 \\ \sum_{j=1}^{m-1} a_j(y_{j+1} - y_j) + b_j(y_{j+1}^2 - y_j^2)/2 & t \in (y_m, \infty) \end{cases}$$

In the context of our diagnostic, if we wish to estimate $\pi_k(\theta)$ we have, from (3) and (2), that $y_i = \theta_k^{(i)}$, $t = \theta$, $a_i = p_k(\theta_k^{(i)})$ and $b_i = 0$. By Lemma 4.1 the normalisation constant for our density estimator is given by

$$P_1(\theta_k^{(m)}) = \sum_{j=1}^{m-1} a_j(\theta_k^{(j+1)} - \theta_k^{(j)}),$$

and we obtain our first estimator

$$\hat{\pi}_{k,1}(\theta) = p_1(\theta) / P_1(\theta_k^{(m)}).$$

4.2 Piecewise linear estimation

An alternative approach based upon estimating the gradient of π_k , may be obtained by using the gradients to π_k at each of the sample points to form a piecewise (rather than stepwise) linear approximation to π_k . Again, we arbitrarily set the approximating function to be zero at $\theta_k^{(1)}$ and then use the gradient estimate at that point to define an approximation over the range $[\theta_k^{(1)}, \theta_k^{(2)}]$. We can then repeat this procedure to obtain a sequence of lines defined between successive $\theta_k^{(i)}$.

This second approximation can be used to obtain an estimate of the corresponding distribution function, by analytically integrating the piecewise linear approximation. The resulting estimate of the distribution function is then normalised by setting its value at $\theta_k^{(m)}$ to be 1. We proceed as follows.

By Lemma 3.1, the \bar{U}_k provide an empirical estimate of the gradient of λ_k at points in the sample and, from (1), $\tilde{\pi}_k \bar{U}$ provides an empirical estimate of the gradient of $\tilde{\pi}_k$. Thus, we may construct an arbitrarily scaled approximation to π_k , denoted by p_k , as follows. First, set $p_k(\theta_k^{(1)}) = 0$ and then, for $i = 2, \dots, m$, set

$$p_k(\theta_k^{(i)}) = p_k(\theta_k^{(i-1)}) + (\theta_k^{(i)} - \theta_k^{(i-1)}) (\tilde{\pi}_k(\theta_k^{(i)}) \bar{U}(i) + \tilde{\pi}_k \theta_k^{(i-1)} \bar{U}(i-1)) / 2.$$

Thus, we define p_k at the points in the sample. We can then extrapolate between these points to produce a piecewise linear function defined over the entire range $[\theta_k^{(1)}, \theta_k^{(m)}]$, by setting

$$p_k(\theta) = p_k(\theta_k^{(i)}) + \frac{\theta - \theta_k^{(i)}}{\theta_k^{(i+1)} - \theta_k^{(i)}} \left(p_k(\theta_k^{(i+1)}) - p_k(\theta_k^{(i)}) \right), \quad \text{for } \theta \in [\theta_k^{(i)}, \theta_k^{(i+1)}]. \quad (4)$$

Having obtained this piecewise linear approximation to π_k , we can then estimate the corresponding distribution function, by integrating p_k within the range $[\theta_k^{(1)}, \theta_k^{(m)}]$. This estimate is normalised by dividing by the integral over the entire range so that the estimate of the distribution function becomes 1 at the last observed data point.

From (4), we have $y_i \equiv \theta_k^{(i)}$, $t \equiv \theta$,

$$a_i = p_k(\theta_k^{(i)}) - \frac{\theta_k^{(i)}}{\theta_k^{(i+1)} - \theta_k^{(i)}} \left(p_k(\theta_k^{(i+1)}) - p_k(\theta_k^{(i)}) \right) \quad \text{and} \quad b_i = \frac{p_k(\theta_k^{(i+1)}) - p_k(\theta_k^{(i)})}{\theta_k^{(i+1)} - \theta_k^{(i)}}.$$

Thus, Lemma 4.1 suggests that

$$\hat{\pi}_{k,2}(\theta) = p_1(\theta) / P_1(\theta_k^{(m)})$$

is a normalised estimator for the marginal density $\pi_k(\theta)$, where

$$P_1(\theta_k^{(m)}) = \sum_{j=1}^{m-1} a_j(\theta_k^{(j+1)} - \theta_k^{(j)}) + b_j((\theta_k^{(j+1)})^2 - (\theta_k^{(j)})^2) / 2.$$

By arbitrarily setting $p_k(\theta_k^{(1)}) = 0$, we make $p_1(t)$ negative whenever the gradient $\tilde{\pi}_k \bar{U}$ is less than zero. In this case we must re-normalise $p_2(t)$ by setting the right tail to be zero. If we let $C = a_{m-1} + b_{m-1}\theta_m$, then if $C < 0$, we set $a'_i = a_i + |C|$ for $i = 1, \dots, m-1$ and use a'_i instead of a_i .

4.3 Piecewise exponential estimation

An alternative method for estimating the density π_k is to use the $\bar{U}_k(i)$ to construct a piecewise exponential approximation to π_k . This can be done by forming a piecewise linear approximation to $\lambda_k(\theta_k)$, which is then exponentiated to form an approximation to π_k . We proceed as follows.

We begin by defining $\hat{\lambda}(\theta_k^{(i)})$ at each of the sample points, as in (2) and obtain a piecewise linear estimator $\hat{\lambda}_k(\theta)$ of $\lambda_k(\theta)$ over $[\theta_k^{(1)}, \theta_k^{(m)}]$, by setting

$$\hat{\lambda}_k(\theta) = \hat{\lambda}_k(\theta_k^{(i)}) + \frac{\theta - \theta_k^{(i)}}{\theta_k^{(i+1)} - \theta_k^{(i)}} \left(\hat{\lambda}_k(\theta_k^{(i+1)}) - \hat{\lambda}_k(\theta_k^{(i)}) \right), \quad \text{for } \theta \in [\theta_k^{(i)}, \theta_k^{(i+1)}]. \quad (5)$$

Having obtained this piecewise linear approximation to $\lambda_k(\theta_k)$, we may obtain an estimate to the density π_k by exponentiating to obtain $p_k(\theta) = \exp[\hat{\lambda}_k(\theta)]$ and integrating the piecewise exponential function $p_k(\theta)$ within the range $[\theta_k^{(1)}, \theta_k^{(m)}]$. As before, we then normalise by dividing by the integral over the entire range, which we obtain via the following lemma.

Lemma 4.2 *Given a piecewise exponential function,*

$$p_3(t) = \begin{cases} \exp(a_i + b_i t) & t \in (y_i, y_{i+1}], \\ 0 & t > y_m \text{ or } t < y_1 \end{cases}, \quad i = 1, \dots, m-1, \quad \text{then}$$

$$P_3(t) = \int_{-\infty}^t p_3(\tau) d\tau = \begin{cases} 0 & t \in (-\infty, y_1] \\ \frac{e^{a_i}(e^{b_i t} - e^{b_i y_i})}{b_i} + \sum_{j=1}^{i-1} \frac{e^{a_j}(e^{b_j y_{j+1}} - e^{b_j y_j})}{b_j} & t \in (y_i, y_{i+1}], \quad i = 1, \dots, m-1 \\ \sum_{j=1}^{m-1} \frac{e^{a_j}(e^{b_j y_{j+1}} - e^{b_j y_j})}{b_j} & t \in (y_m, \infty) \end{cases}.$$

As before, we have $y_i \equiv \theta_k^{(i)}$, $t \equiv \theta$ and, from (5), we have

$$a_i = \hat{\lambda}_k(\theta_k^{(i)}) - \frac{\theta_k^{(i)}}{\theta_k^{(i+1)} - \theta_k^{(i)}} \left(\hat{\lambda}_k(\theta_k^{(i+1)}) - \hat{\lambda}_k(\theta_k^{(i)}) \right) \quad (6)$$

and

$$b_i = \frac{\hat{\lambda}_k(\theta_k^{(i+1)}) - \hat{\lambda}_k(\theta_k^{(i)})}{\theta_k^{(i+1)} - \theta_k^{(i)}}. \quad (7)$$

Thus, the piecewise exponential estimator is given by

$$\hat{\pi}_{k,3}(\theta) = p_3(\theta) / P_3(\theta_k^{(m)}),$$

where

$$P_3(\theta_k^{(m)}) = \sum_{j=1}^{m-1} \frac{e^{a_j} (e^{b_j \theta_k^{(j+1)}} - e^{b_j \theta_k^{(j)}})}{b_j} \quad \text{by Lemma 4.2.} \quad (8)$$

4.4 Extending the piecewise exponential estimate

Suppose that π has bounded support, $[t_{min}, t_{max}]$ then we can extend Lemma 4.2 as follows.

Lemma 4.3 *Given a piecewise exponential function and $t \in (t_{min}, t_{max})$,*

$$p_4(t) = \begin{cases} 0 & t \in (-\infty, t_{min}) \\ \exp(a_1 + b_1 t) & t \in (t_{min}, y_1] \\ \exp(a_i + b_i t) & t \in (y_i, y_{i+1}], \quad i = 1, \dots, m-1 \\ \exp(a_{m-1} + b_{m-1} t) & t \in (y_m, t_{max}] \\ 0 & t \in (t_{max}, \infty) \end{cases}$$

then,

$$P_4(t) = \int_{-\infty}^t p_4(\tau) d\tau = \begin{cases} 0 & t \in (-\infty, t_{min}) \\ \frac{e^{a_1} (e^{b_1 t} - e^{b_1 t_{min}})}{b_1} & t \in (t_{min}, y_1] \\ \frac{e^{a_1} (e^{b_1 y_1} - e^{b_1 t_{min}})}{b_1} + \sum_{j=1}^{i-1} \frac{e^{a_j} (e^{b_j y_{j+1}} - e^{b_j y_j})}{b_j} + \frac{e^{a_i} (e^{b_i t} - e^{b_i y_i})}{b_i} & t \in (y_i, y_{i+1}], \quad i = 1, \dots, m-1 \\ \frac{e^{a_1} (e^{b_1 y_1} - e^{b_1 t_{min}})}{b_1} + \sum_{j=1}^{m-1} \frac{e^{a_j} (e^{b_j y_{j+1}} - e^{b_j y_j})}{b_j} + \frac{e^{a_{m-1}} (e^{b_{m-1} t} - e^{b_{m-1} y_m})}{b_{m-1}} & t \in (y_m, t_{max}] \\ \frac{e^{a_1} (e^{b_1 y_1} - e^{b_1 t_{min}})}{b_1} + \sum_{j=1}^{m-1} \frac{e^{a_j} (e^{b_j y_{j+1}} - e^{b_j y_j})}{b_j} + \frac{e^{a_{m-1}} (e^{b_{m-1} t_{max}} - e^{b_{m-1} y_m})}{b_{m-1}} & t \in (t_{max}, \infty) \end{cases}$$

Thus, with a_i and b_i as defined in (6) and (7) and with p_4 and P_4 as defined in Lemma 4.3, we can take

$$\hat{\pi}_{k,4}(\theta) = p_4(\theta) / P_4(t_{max}),$$

to be our estimator which is defined over the range $[t_{min}, t_{max}]$ rather than the smaller $[\theta_k^{(1)}, \theta_k^{(m)}]$ range common to the previous estimators. Of course, t_{min} and t_{max} need not be finite. However, in order to ensure that the distribution function remains finite, we can only extend t_{min} to $-\infty$ if $b_1 > 0$ else $\exp(b_1 t_{min})$ does not have a finite limit. Similarly, we can only extend the upper limit to ∞ if $b_{m-1} < 0$, else $\exp(b_{m-1} t_{max})$ does not have a finite limit.

The values of b_1 and b_{m-1} are entirely problem- and data-dependent therefore though in many cases it will be possible to extend the $\hat{\pi}_{k,4}(\theta)$ estimator to the whole real line, these conditions on b_1 and b_{m-1} will have to be checked.

4.5 Comparisons

Before considering the application of these results to convergence assessment, we examine how well these methods work as density estimators. As an illustration, we take three different densities: a $N(0, 1)$, an exponential(1), and an even mixture of $N(-3, 1)$ and $N(2, 1)$ densities. Each of these take quite different shapes and we test the performance of the four estimators. For each example, we simulated five sequences of 100 observations from the corresponding densities and density estimates were obtained for each sequence separately, the average is taken over the five sequences together with the 95% confidence interval, the results are given in Figure 2.

[Figure 2 about here.]

We can see from Figure 2 that Method 4 (the extended piecewise exponential estimator) clearly outperforms the remaining estimators, though Method 3 (the basic piecewise exponential estimator) also performs well. Closer inspection reveals that Method 4 is far better at estimating the tails of the distribution and so would generally be the preferred method when a choice is available.

As a further test, we used Method 4 to estimate the mixture density on the basis of 10, 20, 30, 50, 70, and 100 observations drawn directly from the mixture. Our results show that although for small samples the estimate is poor, the estimator rapidly improves with sample size and provides almost perfect performance with only 70 observations. This provides reassurance that large sample sizes are not required to obtain reasonable density estimates. This is an important consideration if they are to be used for convergence diagnosis as described in the next section.

Of course, these density estimates are of value in themselves and may be useful as a means for estimating marginal densities of interest from MCMC output, for example, where Rao-Blackwell estimators are not available. However, our focus in this paper is on their application to convergence assessment and we explain how they may be used for that purpose in the next section.

5 The path sampling diagnostic

Once we obtain our density estimate (whichever method we use), we might compare it with similar estimates from other chains in order to check that they are each sampling from the same (presumably the stationary) distribution. This may be done separately for all $k = 1, \dots, K$. The idea here is similar to that of Gelman and Rubin (1992a), in that if the differences between the replications are small, then it is reasonable to assume that they are all sampling from the same distribution and that this would be their stationary distribution.

Although it is clear from Section 4.5, that Method 4 gives the best estimate of the distribution function, the method fails when for example $b_1 < 0$ in Lemma 4.3, as discussed in Section 4.4. In this situation, we cannot extend the density to $-\infty$, and we expect that it is not unusual to encounter this type of situation in practice. Though the following results can be easily applied to any of the density estimation procedures, we shall focus only upon the standard piecewise exponential estimator $\hat{\pi}_{k,3}$ described in Section 4.3 for the remainder of the paper since this can be applied in any setting and Figure 2 suggests that it performs best amongst the remaining estimators.

In order to compare density estimates between replications, we require some measure of distance between the corresponding estimates. Two natural choices are the L^1 and L^2 distances between the densities. For any two density estimates, the L^1 distance between them can be obtained as follows.

Proposition 5.1 *Suppose that we have m observations from each of the two simulation schemes, and that we denote them by x_1, x_2, \dots, x_{n_x} and z_1, z_2, \dots, z_{n_z} ($m = n_x + n_z$) with corresponding density estimates $\hat{\pi}_x$ and $\hat{\pi}_z$ respectively. Let $y_i \in \{x_1, x_2, \dots, x_{n_x}, z_1, z_2, \dots, z_{n_z}\}$ such that $y_1 < y_2 < \dots < y_m$. Then the L^1 distance between the two corresponding piecewise exponential estimators is given by D_1 , defined as follows.*

$$D_1 = \int \left| \hat{\pi}_x(\theta) - \hat{\pi}_z(\theta) \right| d\theta = \left| P_x(y_1) - P_z(y_1) \right| + \sum_{i=2}^m \left| [P_x(y_i) - P_x(y_{i-1})] - [P_z(y_i) - P_z(y_{i-1})] \right| + \left| [1 - P_x(y_m)] - [1 - P_z(y_m)] \right|,$$

where $P_x(\cdot)$ and $P_z(\cdot)$ denote the normalised distribution functions obtained by dividing the distribution function estimates given in Lemma 4.2 by the normalisation constant in (8) based upon the x and z series respectively.

Similarly, a second measure between two density estimates can be obtained using L^2 distance, as follows.

Proposition 5.2 *Suppose that we have m observations from each of the two simulation schemes, and that we denote them by x_1, x_2, \dots, x_{n_x} and z_1, z_2, \dots, z_{n_z} ($m = n_x + n_z$) with corresponding density estimates $\hat{\pi}_x$ and $\hat{\pi}_z$ respectively. Let $y_i \in \{x_1, x_2, \dots, x_{n_x}, z_1, z_2, \dots, z_{n_z}\}$ such that $y_1 < y_2 < \dots < y_m$. Then the L^2 distance between the two corresponding piecewise exponential estimators is given by D_2 , defined as follows.*

$$\begin{aligned} D_2 &= \int [\hat{\pi}_x(\theta) - \hat{\pi}_z(\theta)]^2 d\theta = \sum_{i=1}^{m-1} \int_{y_i}^{y_{i+1}} \frac{1}{P_x^2} \exp[2(a_{x,i} + b_{x,i}t)] dt \\ &\quad + \int_{y_i}^{y_{i+1}} \frac{1}{P_z^2} \exp[2(a_{z,i} + b_{z,i}t)] dt - \frac{2}{P_x P_z} \int_{y_i}^{y_{i+1}} \exp[(a_{x,i} + a_{z,i}) + (b_{x,i} + b_{z,i})t] dt \\ &= \sum_{i=1}^{m-1} \left\{ \frac{1}{2b_{x,i} P_x^2} (\exp[2(a_{x,i} + b_{x,i}y_{i+1})] - \exp[2(a_{x,i} + b_{x,i}y_i)]) \right. \\ &\quad + \frac{1}{2b_{z,i} P_z^2} (\exp[2(a_{z,i} + b_{z,i}y_{i+1})] - \exp[2(a_{z,i} + b_{z,i}y_i)]) \\ &\quad \left. - \frac{2}{(b_{x,i} + b_{z,i}) P_x P_z} (\exp[a_{x,i} + a_{z,i} + (b_{x,i} + b_{z,i})y_{i+1}] - \exp[a_{x,i} + a_{z,i} + (b_{x,i} + b_{z,i})y_i]) \right\} \end{aligned}$$

where P_x and P_z denote the normalisation constants given in (8) based upon the x and z sequences respectively and the $a_{x,i}$ are the a_i given in (6) based upon the x sequence, with similar definitions for $a_{z,i}$, $b_{x,i}$ and $b_{z,i}$.

Each of these distances is calculated for a particular marginal density, $\pi_k(\theta)$. An overall measure of distance may be obtained by simply summing these distances (L^1 or L^2) over all parameters k to obtain what we shall refer to as the multivariate D_1 and D_2 plots respectively.

5.1 Toy Example

Here we again illustrate the path sampling diagnostic in a toy example, and compare its performance with existing diagnostics.

Let the target distribution be

$$\pi = N(0.6, 1)$$

We obtain 10000 samples each, from 3 skew-normal distributions which are very similar to each other and to the target distribution

$$SN(1.14, 1, -1) \tag{9}$$

$$SN(1.29, 1, -2) \tag{10}$$

$$SN(1.6, 1.3, -5) \tag{11}$$

The histograms in Figure 3 indicates the degree of skewness in each sample.

[Figure 3 about here.]

As expected, due to the similarity of the output, the Gelman and Rubin diagnostic in Figure 3(d) and other existing popular diagnostics such as Heidelberger and Welch (1983) and Geweke (1992), fail to detect a difference between the samples.

We can see that in Figure 3(b,c), the path sample density estimation has shown quite some differences between the three skew-normal samples, particularly in the tails of the density estimation. The path sampling density estimates are expected to be extremely effective in tail areas, as relatively

few points are needed to gain confidence in the estimations, thus we expect that path sampling diagnostic would be particularly useful for detecting chains which do not explore low density areas well. Discrepancies between density estimates and the raw histograms of output is another indication that individual chains have not converged.

For the L^1 and L^2 diagnostics, we calculated distances for all three combinations of chain pairs (chains (1, 2), (2, 3), and (1, 3)). This is done at each iteration t , using only the second half of the chains up to t . The final diagnostic values are calculated as the average over the three sets. We can add a 95% confidence band over these values as an indication of variations between the different pairwise comparisons. Finally, smoothing of the diagnostics may be desirable in some cases for clarity, here we averaged over the second half of the diagnostic values up to iteration t , to obtain smoothed plots in Figure 3(e,f). The plots indicate that although the outputs are very close to each other, the variation in the case of D_1 is quite large and it was not decreasing rapidly. Both diagnostics appear to suggest that we cannot be confident that convergence has been achieved. Thus the path sampler was able to use some information from the tails of the target distribution to help assess convergence.

6 Examples

In this section, we examine the application of our new diagnostic methods for the determination of MCMC burn-in in the context of, first of all, the analysis of an autoregressive times series and secondly, the analysis of censored survival data, and finally a bivariate Normal model with a non-identified parameter problem.

6.1 Autoregressive Time Series

Here, we fit an autoregressive model of order 3 to the dataset described and modelled by Huerta and West (1999). This series comprises $T = 540$ monthly observations of the Southern Oscillation Index (SOI) during 1950-1995 which is computed as the difference of the departure from the long-term monthly mean sea level pressures at Tahiti in the South Pacific and Darwin in Northern Australia.

The data comprises univariate observations x_1, \dots, x_T and the AR(3) model suggests that

$$x_t = \sum_{i=1}^3 a_i x_{t-i} + \epsilon_t \quad (12)$$

where $\epsilon_t \sim N(0, 1/\tau)$. Thus, our model admits four parameters, $\boldsymbol{\theta} = \{a_1, a_2, a_3, \tau\}$. We take a vague $\Gamma(0.001, 0.001)$ prior distribution for τ and independent $N(0, 1)$ prior distributions for the a_i . We use the usual approximation to the likelihood, taking

$$L(\mathbf{x}|\boldsymbol{\theta}) = \prod_{t=4}^{540} \sqrt{\frac{\tau}{2\pi}} \exp\left(-\frac{\tau}{2} \left[x_t - \sum_{i=1}^3 a_i x_{t-i}\right]^2\right).$$

It is then fairly simple to show that

$$U_j(\boldsymbol{\theta}) = \tau \sum_{t=4}^{540} \left(x_t - \sum_{i=1}^3 a_i x_{t-i}\right) x_{t-j} - a_j, \quad j = 1, 2, 3$$

and

$$U_4(\boldsymbol{\theta}) = \frac{267.501}{\tau} - 0.001 - \frac{1}{2} \sum_{t=4}^{540} \left(x_t - \sum_{i=1}^3 a_i x_{t-i}\right)^2.$$

To illustrate our diagnostic methods we ran 5 replications of an MCMC chain each of length 10000 iterations. The MCMC algorithm comprised block updates of the autoregressive parameters followed by a univariate update of the error variance, each using Gibbs updates. This should provide a fairly rapidly mixing chain. See Brooks *et al* (2003), for example.

The univariate score diagnostic plots for the four parameters are provided in Figure 4(a-d). Diagnostics were calculated over second half of the chains at regular intervals of 100 iterations to minimise

computational cost. We can clearly see from the individual plots of Figure 4(a-d) that the chains perform well and that convergence appears to have been achieved rapidly. The diagnostic value settles quickly to values around zero and the confidence bounds generally shrink as the simulation continues. The multivariate diagnostic (Figure 4e) also clearly indicates convergence, with the diagnostic plot generally lying mostly within the 95% upper confidence bound.

Figure 4(g,h) provide the multivariate path sampling plots, taken as the average over the four univariate path sampling diagnostic values for each parameter. A 95% confidence interval is provided to indicate the amount of variation between the parameters. We note that the D_1 diagnostic is scaled to be between zero and one by dividing by its maximum value of 2. D_2 is left unscaled.

As with the score statistic, the univariate path sampling diagnostics appeared to suggest that the samplers were performing well and moving fairly rapidly towards convergence. Perhaps the most interesting point to note here is the difference in scale between the plots for different parameters. As we might expect, the parameter about which we know most is τ and the y -scales for τ had a far smaller scale than the other plots. Similarly, the parameter about which we know least is α_3 which had the largest scale along the y -axis.

We note also a slight increase in many of the distance measures after about 3000 iterations, as was also indicated in the multivariate diagnostic plot of Brooks and Gelman (1998a) in Figure 4f. We note that this is consistent with an increase in error bounds for the corresponding score statistic plots at around the same point. Further investigation of the raw trace plots reveals that one of the chains moves further out into the tails of the posterior at this point thereby slightly altering the density estimate from that chain. As the other chains slowly explore the same tail, the uncertainties and corresponding distances decrease again. This highlights the sensitivity of both the diagnostic methods to even small differences between chains.

For comparison, we tested our sample outputs on well known diagnostic techniques. Heidelberger and Welch (1983) and Raftery and Lewis (1992) both diagnosed convergence for these outputs almost immediately after the start of the chains, while the multivariate method of Brooks and Gelman (1998a) shown in Figure 4f returned similar conclusions to our methods.

[Figure 4 about here.]

6.2 Censored Survival Analysis

Here, we revisit the Weibull example used in Brooks and Gelman (1998a) to demonstrate difficulties in assessing convergence. Grieve (1987) provides data that measure photocarcinogenicity or survival times for four groups of mice subjected to different treatments. The survival times are assumed to follow a Weibull distribution, so that the likelihood is given by

$$\prod_i \left(\rho e^{\beta' z_i} t_i^{\rho-1} \right)^{c_i} \exp(-e^{\beta' z_i} t_i^\rho),$$

where t_i denotes the failure or censor time of an individual, $\rho > 0$ is the shape parameter of the Weibull distribution, β is a vector of unknown parameters, the z_i denote covariate vectors assigning each observation to one particular treatment group, and the c_i denote indicator variables such that $c_i = 1$ if time t_i is uncensored and zero otherwise.

Thus, the model has 5 parameters, β_1, \dots, β_4 and ρ . Following Dellaportas and Smith (1993), we assume vague $N(\mu_i, \sigma_i^2)$ prior distributions for the β_i parameters and a similarly vague $\Gamma(\alpha, \gamma)$ prior distribution for ρ , and we use the Gibbs sampler to fit the above model to Grieve's data.

If we let $\theta_i = \beta_i$, $i = 1, \dots, 4$ and $\theta_5 = \rho$, then it is easy to show that, for $k = 1, \dots, 4$,

$$U_k(\theta) = \sum_{i=1}^n z_{ik} (c_i - \exp(\beta' z_i) t_i^\rho) - (\beta_k - \mu_k) / \sigma_k^2$$

and

$$U_5(\theta) = \left(\sum_i c_i + (\alpha - 1) - \gamma \rho \right) / \rho + \sum_{i=1}^n (c_i \log t_i - (\log t_i) \exp(\beta' z_i) t_i^\rho).$$

Figure 5(a, b) provides the multivariate score statistic diagnostic plots, calculated from the univariate score statistic using the approximation to χ^2 distribution, and the multivariate score statistic

U , respectively. We show error bounds based upon 5 replications each comprising 5000 iterations, using dispersed starting points. Individual univariate score plots suggested that the β_1 and β_4 parameters were the slowest to settle but that all chains were performing well beyond 2000 iterations. This conclusion is less easily drawn from the multivariate diagnostic plot which appears to provide a more conservative convergence assessment criterion. Both figures 5(a, b) suggests that approximate convergence may have been achieved after around 3000 iterations, but indicates that a longer run may be required to confirm this.

[Figure 5 about here.]

Figure 5(c, d) provides the corresponding multivariate path sampling diagnostics. Diagnostic values were calculated over the second half of each chain at an interval of 100 iterations to save computational cost. Both of these plots indicate the distance (both L^1 and L^2) between density estimates formed from the five independent replications rapidly decreases with time. However, there appeared to be a slight increase in some of the univariate diagnostic plots towards the end of the simulation which is most notable in D_2 diagnostics for β_3 and β_4 parameters. On closer inspection, it appears that one of the chains for β_3 did not spend enough time in the left tail of the posterior distribution, particularly in the last 1000 iterations, if a lack of convergence at this point was not detected, this would lead to an underestimation in posterior variance. Similarly, there was both slight overestimation in both tails for the β_4 parameter in one chain, and an underestimation in another. The later case may be related to the lack of convergence in β_3 . Thus, a longer run length might be desirable in order to gain greater confidence that these chains have indeed converged.

As a comparison, we ran the multivariate diagnostics of Brooks and Gelman (1998a) on the same sample output, the result is shown in Figure 5g, this diagnostic suggests that the chains may have converged after 2000 iterations but again, a slight deviation away from 1 is found after iteration 4000.

We ran the chains for a further 5000 iterations and computed the diagnostics for the second half of each chain, at an interval of 200 iterations. Univariate D_1 plots suggested that all parameters have converged, particularly after 8000 iterations. The multivariate D_1 plot (Figure 5(e,f)) again shows strong confidence in convergence after 8000 iterations.

In general, D_2 appears to be more conservative, most of the univariate D_2 diagnostics settle down fairly quickly after 2000 iterations, but hovers just above zero for a long time. We note however that the D_1 diagnostic plots can be scaled by dividing by 2 while the D_2 plots are left on the original scales, making it harder to interpret.

6.3 Bivariate Normal Model with a Non-Identified Parameter

We examine the bivariate normal example discussed in Brooks and Gelman (1998a). The distribution of data y depends upon two parameters θ and ϕ :

$$y_i \sim N(\theta_i + \phi_i, 1),$$

where θ and ϕ are not identified by the likelihood but are separated via their prior distribution $p(\theta, \phi)$.

We follow Brooks and Gelman (1998a) and consider only one single observation $y = 0$ and independent prior distributions,

$$p(\theta) \sim N(\mu_\theta, \sigma_\theta^2), \quad p(\phi) \sim N(\mu_\phi, \sigma_\phi^2).$$

The Gibbs sampler can be used to move through the posterior distribution, using the transformation of variables from (θ, ϕ) to (θ, η) where $\eta_i = \theta_i + \phi_i$ to speed convergence. We use the same 1000 iterations of Gibbs sampler output from five replications as those used in Brooks and Gelman (1998a), with $\mu_\theta = \mu_\phi = 50$ and $\sigma_\theta = \sigma_\phi = 10$. In their paper, Brooks and Gelman (1998a) showed that the method proposed by Gelman and Rubin (1992a) had failed to detect a lack of convergence in the η sequence.

Here we monitor the convergence of the θ and η sequences using the multivariate score function diagnostic, as well as the path sampling diagnostic. Thus we need

$$U_\theta(\theta) = -(\theta - 50)/100 + (\eta - \theta - 50)/100$$

and

$$U_{\eta}(\boldsymbol{\theta}) = -\eta - (\eta - \theta - 50)/100.$$

Figure 6a shows the score function diagnostics using multivariate score \boldsymbol{U} , calculated over the second half of the output for each of the five replicated chains. This diagnostic successfully detect a lack of convergence at 1000 iterations.

[Figure 6 about here.]

We also calculated the path sampling diagnostics for these outputs, diagnostics were calculated at every iteration, and results were then smoothed over the last half of the diagnostic values to make interpretation easier. Figure 6(b-e) show the univariate D_1 and D_2 diagnostics. The D_1 and D_2 diagnostic values move close to zero after 800 iterations, however, a lack of confidence in the convergence is shown by the fact that zero does not fall within the confidence bands.

Finally, we ran the MCMC sampler a further 5000 iterations until convergence has been achieved. Figure 6(f, g) show the marginal density estimation for the parameters θ and η , using the piecewise exponential estimator. Density estimates here are based on the combined final 5000 iterations of replicate chains.

7 Discussion

This paper has considered two new methods for convergence assessment, using the score function statistic and path sampling methods. Both are easily implemented in a generic fashion (using the target density function as specified up to a normalising constant) and do not require extensive computation beyond what was already needed for MCMC. Furthermore the output from both of these diagnostics are easily interpretable, particularly the score statistic diagnostic, with its natural scale. Of course, the derivatives required for the construction of the score statistics may not always be available. In this case, numerical techniques can be used to estimate the derivatives, facilitating the construction of a highly generic code which requires only the functional form of the target distribution and the sample values as input.

In developing the path sampling diagnostic, we show how accurate density estimates can be constructed even from small sets of sample output. Though we only discuss their applications for convergence diagnosis, these are potentially useful tools in their own right and provide an efficient mechanism for parametric density estimation from sample output in general when alternatives such as the Rao-Blackwell density estimation procedure (Gelfand and Smith 1990) are not available. The path sampling diagnostic also has potential in algorithms such as simulated tempering (Marinari and Parisi 1992; Geyer and Thompson 1995) that move through a generalised model space in such a way that it might be difficult to get overdispersed starting points.

No discussion of the issue of convergence assessment techniques could be complete without some more general discussion of the wider context of their use. One issue relating to convergence assessment that is rarely discussed in the literature is the fact that deciding to stop the simulation on the basis of an output-based diagnostic can induce a bias in the resulting estimates. Cowles *et al* (1999) illustrate this idea for a number of simple models and diagnostic techniques. A simple illustration of the general idea can be seen by observing that stationarity is less likely to be diagnosed on occasions when the sample path is out in the tails of the distribution, and so variances (for example) are likely to be underestimated when many of the standard convergence diagnostics are used. Of course, the effect of this bias can be minimised by using overdispersed starting points and generating large post-convergence samples. However, the existence of a bias in such simple cases raises the question of what may happen for more complicated problems where both the sampling algorithm and posterior surface may be less well understood.

Another issue, discussed by Brooks and Gelman (1998b), is that the question of convergence depends, in general, upon what the simulations will be used for. For example, when computing posterior intervals, there is a natural limit on the necessary precision of inferences (e.g., the 95% interval [3.5, 8.4] is as good, in practice, as [3.51345, 8.37802]). In contrast, when estimating functionals such as posterior expectations, the required precision of inferences must be given externally. Thus, no automatic convergence test could work in such a setting without some input as to the desired precision level.

8 Acknowledgements

The authors wish to thank an associate editor and three referees for their helpful and constructive comments. We are particularly indebted to one referee for suggesting the use of the $U^T I U$ statistic. The authors also acknowledge the financial support of the EPSRC for funding the research of the second author.

A Proof of Lemma 3.1

Let $\boldsymbol{\theta}_{(k)}$ denote the vector $\boldsymbol{\theta}$ with element θ_k removed and let $\pi(\boldsymbol{\theta}_{(k)}|\theta_k)$ denote the conditional distribution of $\boldsymbol{\theta}_{(k)}$ given θ_k . Then, clearly,

$$\int \dots \int \pi(\boldsymbol{\theta}_{(k)}|\theta_k) d\boldsymbol{\theta}_{(k)} = 1, \quad (13)$$

since $\pi(\boldsymbol{\theta}_{(k)}|\theta_k)$ is a density. Thus,

$$\begin{aligned} \int \dots \int \frac{\partial}{\partial \theta_k} \pi(\boldsymbol{\theta}_{(k)}|\theta_k) d\boldsymbol{\theta}_{(k)} &= \frac{\partial}{\partial \theta_k} \int \dots \int \pi(\boldsymbol{\theta}_{(k)}|\theta_k) d\boldsymbol{\theta}_{(k)} \quad \text{by regularity} \\ &= \frac{\partial}{\partial \theta_k} (1) \quad \text{by (13)} = 0. \end{aligned} \quad (14)$$

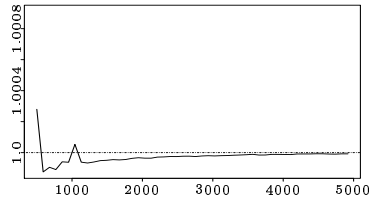
$$\begin{aligned} \text{Now } \mathbb{E}(U_k(\boldsymbol{\theta})) &= \int \dots \int \frac{\partial}{\partial \theta_k} \log(\pi(\boldsymbol{\theta})) \pi(\boldsymbol{\theta}_{(k)}|\theta_k) d\boldsymbol{\theta}_{(k)} \\ &= \int \dots \int \frac{\partial}{\partial \theta_k} \log(\pi(\boldsymbol{\theta}_{(k)}|\theta_k) \pi_k(\theta_k)) \pi(\boldsymbol{\theta}_{(k)}|\theta_k) d\boldsymbol{\theta}_{(k)} \quad \text{by Bayes' theorem} \\ &= \int \dots \int \frac{\partial}{\partial \theta_k} \log(\pi(\boldsymbol{\theta}_{(k)}|\theta_k)) \pi(\boldsymbol{\theta}_{(k)}|\theta_k) d\boldsymbol{\theta}_{(k)} + \int \dots \int \frac{\partial}{\partial \theta_k} \log(\pi_k(\theta_k)) \pi(\boldsymbol{\theta}_{(k)}|\theta_k) d\boldsymbol{\theta}_{(k)} \\ &= \int \dots \int \frac{1}{\pi(\boldsymbol{\theta}_{(k)}|\theta_k)} \frac{\partial}{\partial \theta_k} \pi(\boldsymbol{\theta}_{(k)}|\theta_k) \pi(\boldsymbol{\theta}_{(k)}|\theta_k) d\boldsymbol{\theta}_{(k)} + \frac{\partial}{\partial \theta_k} \log(\pi_k(\theta_k)) \int \dots \int \pi(\boldsymbol{\theta}_{(k)}|\theta_k) d\boldsymbol{\theta}_{(k)} \\ &= 0 + \frac{\partial}{\partial \theta_k} \log(\pi_k(\theta_k)) \cdot 1 \quad \text{by (13) and (14)} = \frac{\partial}{\partial \theta_k} \log(\pi_k(\theta_k)) \\ &= \frac{\partial}{\partial \theta_k} \log(\tilde{\pi}_k(\theta_k)) + \frac{\partial}{\partial \theta_k} \log c \quad \text{where } c \text{ denotes the normalisation constant for } \tilde{\pi} \\ &= \frac{\partial}{\partial \theta_k} \log(\tilde{\pi}_k(\theta_k)) = \frac{\partial}{\partial \theta_k} \lambda_k(\theta_k) \end{aligned}$$

□

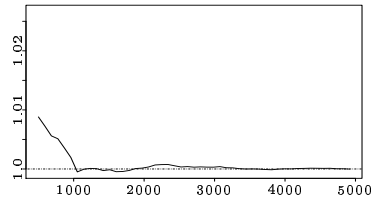
References

- Brooks, S. P. (1998a), Markov Chain Monte Carlo Method and its Application. *The Statistician* **47**, 69–100
- Brooks, S. P. (1998b), Quantitative Convergence Diagnosis for MCMC via CUSUMS. *Statistics and Computing* **8**, 267–274
- Brooks, S. P., P. Dellaportas and G. O. Roberts (1997), A Total Variation Method for Diagnosing Convergence of MCMC Algorithms. *Journal of Computational and Graphical Statistics* **6**, 251–265
- Brooks, S. P. and A. Gelman (1998a), Alternative Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics* **7**, 434–455
- Brooks, S. P. and A. Gelman (1998b), Alternative Methods for Monitoring Convergence of Iterative Simulations. In S. Weisburg (ed.), *Computing Science and Statistics 30*, pp. 30–36, Interface Foundation of North America Inc.

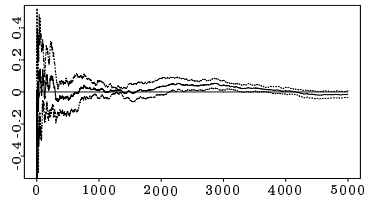
- Brooks, S. P., P. Giudici and G. O. Roberts (2003), Efficient Construction of Reversible Jump Proposal Distributions (with discussion). *Journal of the Royal Statistical Society, Series B* **65**, 3–55
- Cowles, M. K., G. Roberts and J. S. Rosenthal (1999), Possible Biases Induced by MCMC Convergence Diagnostics. *Journal of Statistical Computing and Simulation* **64**, 87–104
- Cowles, M. K. and J. S. Rosenthal (1998), A Simulation Approach to Convergence Rates for Markov Chain Monte Carlo algorithms. *Statistics and Computing* **8**, 115–124
- Cox, D. R. and D. V. Hinkley (1974), *Theoretical Statistics*. Chapman and Hall: London
- Dellaportas, P. and A. F. M. Smith (1993), Bayesian Inference for Generalized Linear and Proportional Hazards Models via Gibbs Sampling. *Applied Statistics* **42**, 443–460
- Dobson, A. J. (1990), *An Introduction to Generalised Linear Models*. Chapman and Hall
- Fosdick, L. D. (1959), Calculation of Order Parameters in a Binary Alloy by the Monte Carlo Method. *Physical Review* **116**, 565–573
- Gelfand, A. E. and A. F. M. Smith (1990), Sampling Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association* **85**, 398–409
- Gelman, A. (1992), Iterative and non-iterative simulation algorithms. *Computing Science and Statistics* **24**, 433–438
- Gelman, A. and X. L. Meng (1998), Simulating Normalizing Constants: From Importance Sampling to Bridge Sampling to Path Sampling. *Statistical Science* **13**, 163–185
- Gelman, A., G. O. Roberts and W. R. Gilks (1996), Efficient Metropolis Jumping Rules. In J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith (eds.), *Bayesian Statistics 5*, pp. 599–608, New York: Oxford University Press
- Gelman, A. and D. B. Rubin (1992a), Inference from Iterative Simulation using Multiple Sequences. *Statistical Science* **7**, 457–511
- Gelman, A. and D. B. Rubin (1992b), A Single Series from the Gibbs Sampler Provides a False Sense of Security. In J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith (eds.), *Bayesian Statistics 4*, pp. 625–631, New York: Oxford University Press
- Geweke, J. (1992), Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments. In J. M. Bernardo, A. F. M. Smith, A. P. Dawid and J. O. Berger (eds.), *Bayesian Statistics 4*, pp. 169–193, New York: Oxford University Press
- Geyer, C. J. and E. A. Thompson (1995), Annealing Markov Chain Monte Carlo with Applications to Ancestral Inference. *Journal of the American Statistical Association* **90**, 909–920
- Grieve, A. P. (1987), Applications of Bayesian Software: Two Examples. *Statistician* **36**, 283–288
- Heidelberger, P. and P. D. Welch (1983), Simulation Run Length Control in the Presence of an Initial Transient. *Operations Research* **31**, 1109–1144
- Huerta, G. and M. West (1999), Priors and Component Structures in Autoregressive Time Series Models. *Journal of the Royal Statistical Society, Series B* **61**, 881–900
- Kong, A. (1992), A Note on Importance Sampling using Standardised Weights. Technical report, Department of Statistics, University of Chicago
- Marinari, E. and G. Parisi (1992), Simulated Tempering: A New Monte Carlo Scheme. *Europhysics letters* **19**, 451–458
- Propp, J. G. and D. B. Wilson (1996), Exact Sampling with Coupled Markov Chains and Applications to Statistical Mechanics. *Random Structures and Algorithms* **9**, 223–252
- Raftery, A. E. and S. M. Lewis (1992), How Many Iterations in the Gibbs Sampler? In J. M. Bernardo, A. F. M. Smith, A. P. Dawid and J. O. Berger (eds.), *Bayesian Statistics 4*, pp. 763–774, Oxford University Press
- Roberts, G. O. (1992), Convergence Diagnostics of the Gibbs sampler. In J. M. Bernardo, A. F. M. Smith, A. P. Dawid and J. O. Berger (eds.), *Bayesian Statistics 4*, pp. 775–782, Oxford University Press



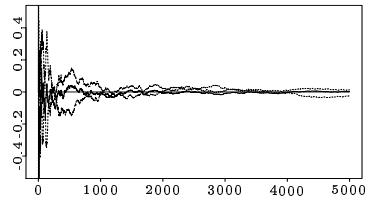
(a) Gelman and Rubin (1992a) diagnostic for x



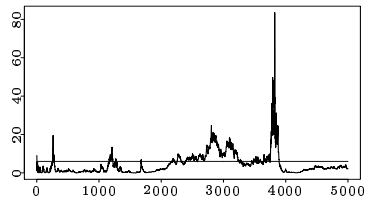
(b) Gelman and Rubin (1992a) diagnostic for y



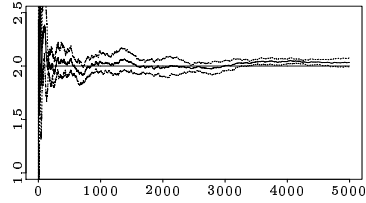
(c) Diagnostic based on univariate score U_x



(d) Diagnostic based on univariate score U_y



(e) Multivariate diagnostic based on univariate scores X^2



(f) Diagnostic based on multivariate score \mathbf{U} , averaged over 5 replications

Figure 1: Mixture of two bivariate Normal densities: (a,b) Gelman and Rubin (1992a) diagnostics; (c,e) univariate score function diagnostics U_k ; and (f) multivariate score function $U^T I U$ diagnostic. Solid lines indicates the diagnostic value and the dashed lines the 95% confidence bands.

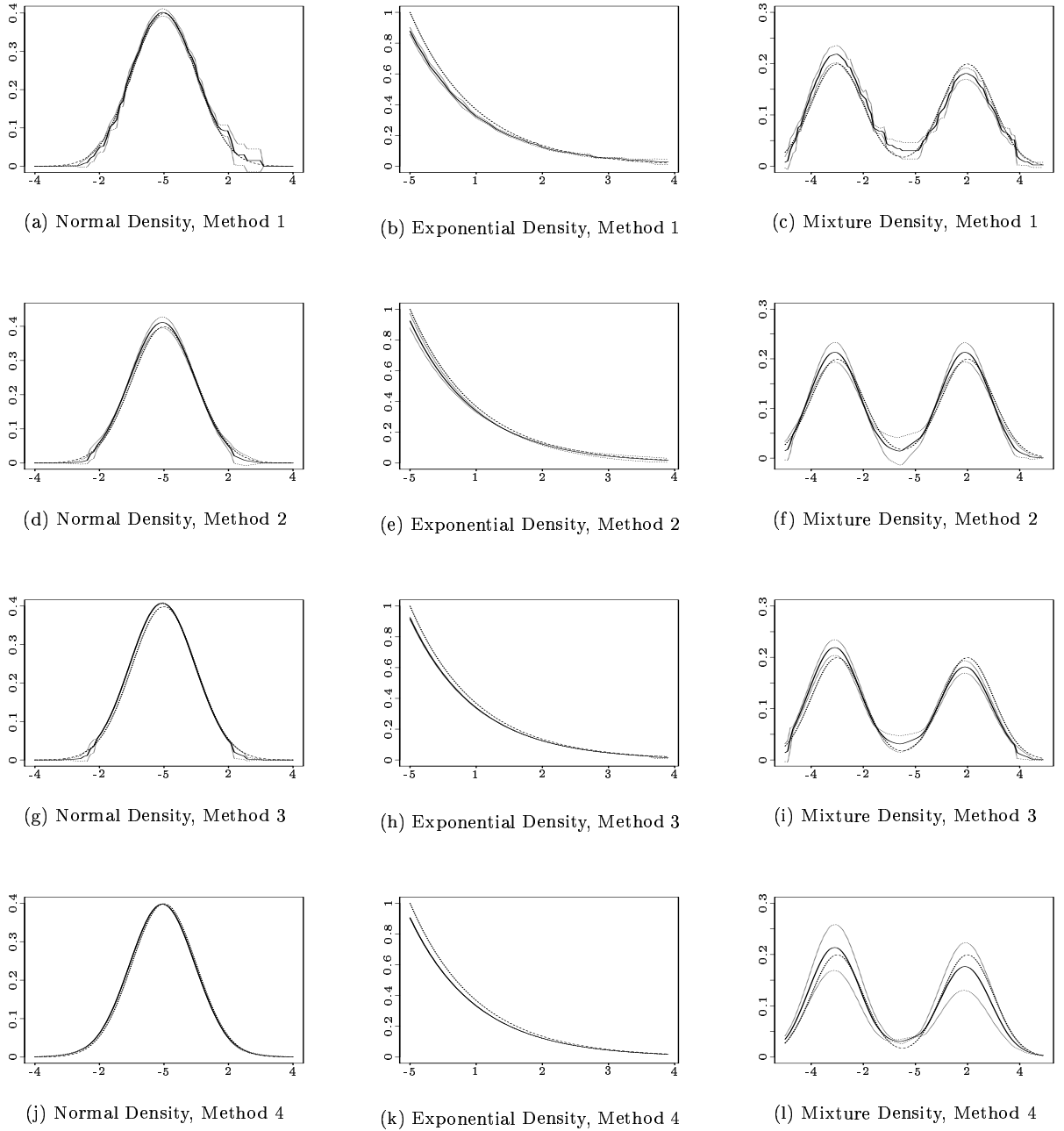


Figure 2: Density estimates using methods 1–4 on a $N(0, 1)$, exponential(1) and an even mixture of $N(-3, 1)$ and $N(2, 1)$ densities. For each graph, 5 simulated data sets were used for each density curve estimation, the mean curve for the 5 repetitions is represented by a solid line, and the corresponding 95% confidence intervals are plotted in dotted lines, with the true densities indicated by a dashed line.

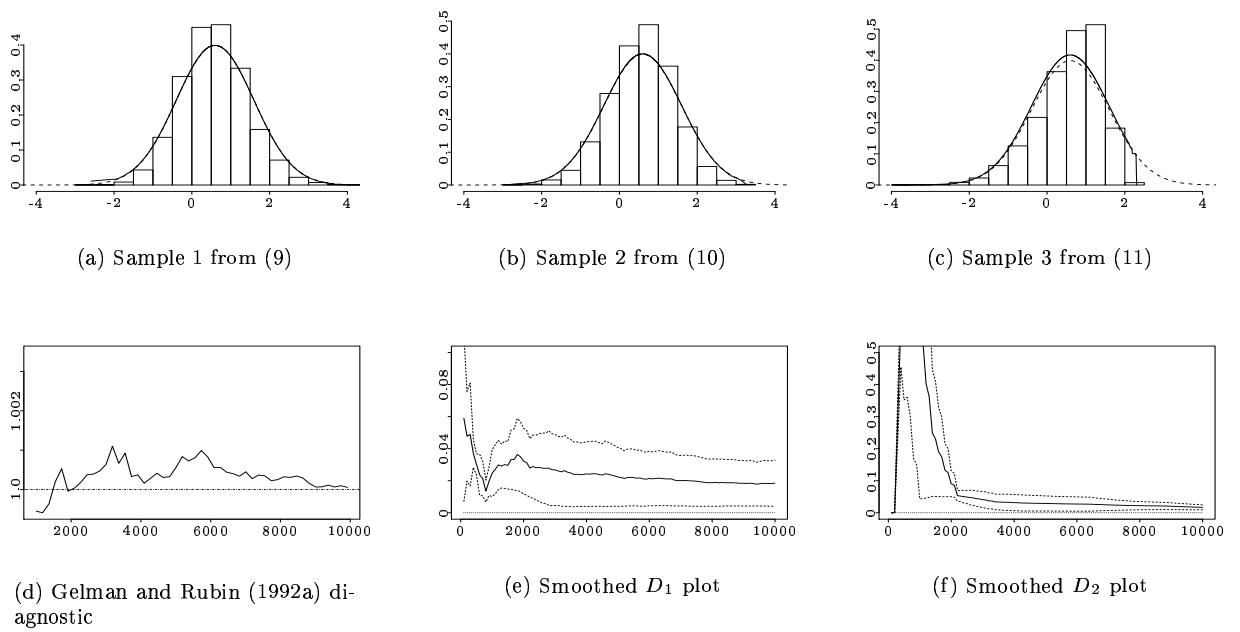


Figure 3: (a-c) Histograms of samples from the skew-normal distributions. Solid curve indicates the path sampling density estimates based on these samples using the piecewise exponential estimator of Section 4.3, and the dotted lines indicates the density curve of the target distribution $N(0.6, 1)$. (d-f) Gelman and Rubin diagnostic (1992a) output (d); L^1 and L^2 path sampling diagnostic (e,f) from lemma 4.2 based on Proposition 5.1 and Proposition 5.2. Solid lines indicate the respective L^1 and L^2 diagnostic values, and the dotted lines gives the point-wise 95% confidence interval.

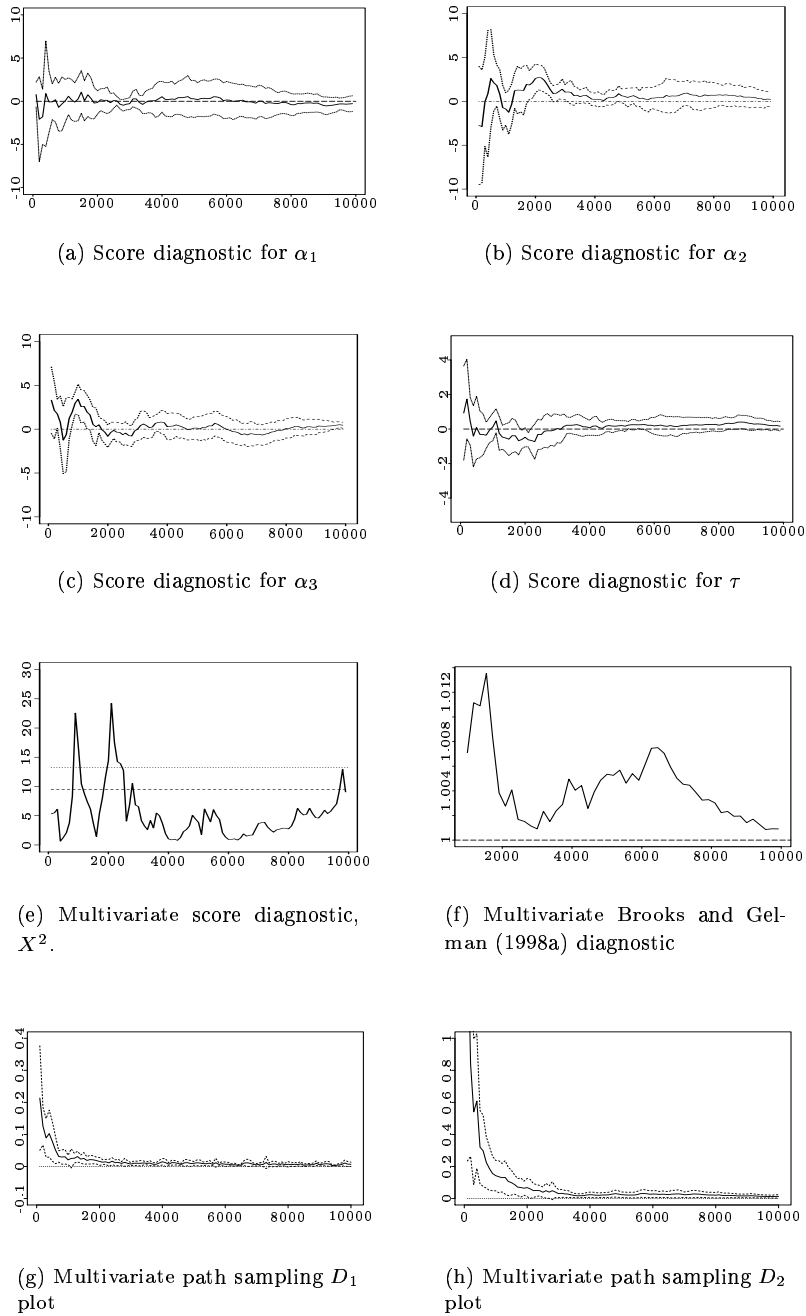
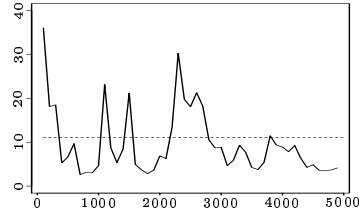
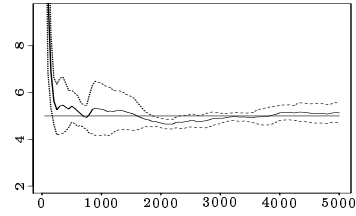


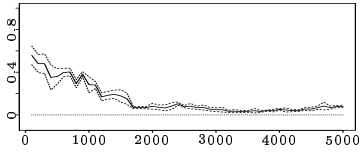
Figure 4: For the autoregressive example: univariate score function diagnostic for each of the four parameters (a-d); the corresponding multivariate diagnostic based on X^2 (e), solid line indicates the diagnostic value and the dashed lines the 95% and 99% (bottom and top respectively) confidence bands; multivariate diagnostic plot based on Brooks and Gelman (1998a) (f); multivariate path sampling plots based on D_1 and D_2 (g,h), with 95% confidence intervals.



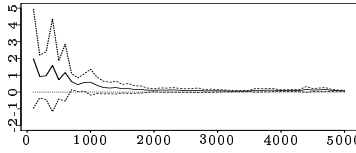
(a) Multivariate score diagnostic, X^2



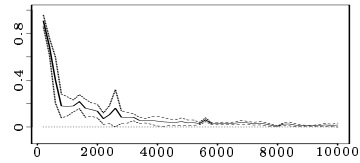
(b) Multivariate score diagnostic, \mathbf{U}



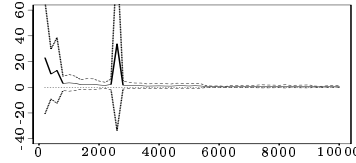
(c) Multivariate D_1 plot, 5000 iterations



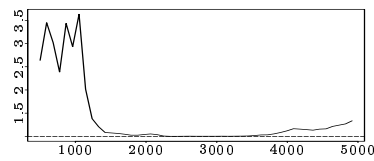
(d) Multivariate D_2 plot, 5000 iterations



(e) Multivariate D_1 plot, 10000 iterations

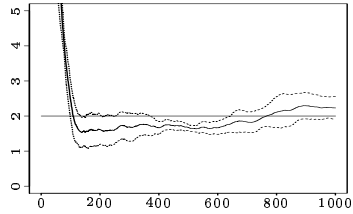


(f) Multivariate D_2 plot, 10000 iterations

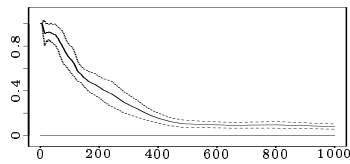


(g) Multivariate diagnostic from Brooks and Gelman (1998a), after 5000 iterations

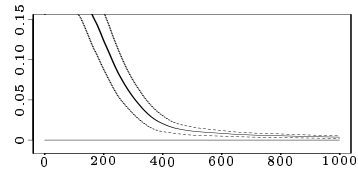
Figure 5: For the censored survival example: multivariate score function diagnostic plots (a, b) based on X^2 and \mathbf{U} ; multivariate path sampling diagnostic plots (c, d) using D_1 and D_2 for 5000 iterations; multivariate path sampling diagnostic plots (e, f) using D_1 and D_2 for 10000 iterations; and multivariate diagnostic from Brooks and Gelman (1998a). Solid line indicates the diagnostic value and the dashed lines the 95% confidence bands.



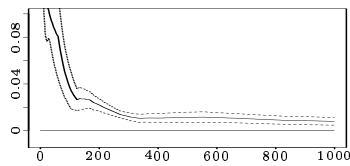
(a) Multivariate score diagnostic based on \mathbf{U}



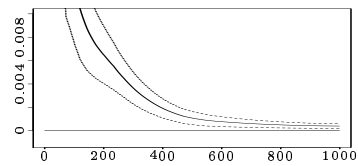
(b) D_1 plot for θ



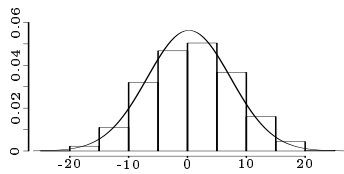
(c) D_2 plot for θ



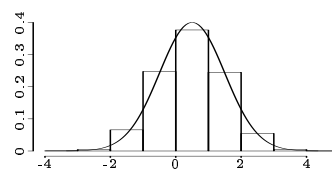
(d) D_1 plot for η



(e) D_2 plot for η



(f) Density estimation for θ



(g) Density estimation for η

Figure 6: For the bivariate Normal with non-identified parameter example: multivariate score function diagnostic plot using \mathbf{U} (a); univariate path sampling diagnostic plots (b-e) using D_1 and D_2 statistics for the parameters θ and η , solid line indicates the diagnostic value and the dashed lines the 95% confidence bands over 5 replications; and path sampling density estimates for θ and η (f, g).