Assessing Convergence of Markov Chain Monte Carlo Algorithms

Stephen P. Brooks and Gareth O. Roberts *

Abstract

We motivate the use of convergence diagnostic techniques for Markov Chain Monte Carlo algorithms and review various methods proposed in the MCMC literature. A common notation is established and each method is discussed with particular emphasis on implementational issues and possible extensions. The methods are compared in terms of their interpretability and applicability and recommendations are provided for particular classes of problems.

1 Introduction

There are many important implementational issues associated with MCMC methods. These include (amongst others) the choice of sampler, the number of independent replications to be run, the choice of starting values and both estimation and efficiency problems. In practice, we use ergodic averages over realisations of a Markov chain to estimate functionals of interest. In order to reduce the possibility of bias caused by the effect of starting values, iterates within an initial transient phase or *burn in* period are usually discarded. One of the most difficult implementational problems is that of determining the length of the required burn in, since rates of convergence of algorithms on different target distributions vary considerably.

Ideally, we would like to analytically compute or estimate the Markov chain convergence rate and then take sufficient iterations to satisfy any prescribed accuracy criteria. However, this is not possible in general (Tierney, 1994). In fact for Markov chains it is extremely difficult to prove even the existence of a geometric rate of convergence to stationarity (Roberts and Tweedie, 1994) and there are many commonly used algorithms which frequently fail to converge geometrically quickly at all.

Roberts and Polson (1994), Chan (1993) and Schervish and Carlin (1992) provide qualitative geometric convergence results for quite general classes of target densities. However, although these results have theoretical import in ensuring the existence of

*School of Mathematics, University of Bristol, University Walk, Bristol, BS8 1TW, and Statistical Laboratory, University of Cambridge, England. The Authors gratefully acknowledge the help of the joint Statistical Laboratory and MRC Biostatistics unit MCMC discussion group, and in particular the helpful and insightful comments and suggestions offered by Wally Gilks, Sujit Sahu and David Spiegelhalter. We are also grateful to the EPSRC (formerly the SERC) for funding the research of the first author.

central limit theorems for instance, they do not offer explicit bounds on the rate of convergence which could be used to determine MCMC sample lengths.

Certain specialised techniques exist to *bound* convergence rates (and hence required burn in lengths) for Markov chains. See for example Lawler and Sokal (1988), Sinclair and Jerrum (1988), Diaconis and Stroock (1991), and Meyn and Tweedie (1993). However, in general it is difficult to apply these results effectively in the context of MCMC. Notable exceptions include Applegate *et al* (1990), Amit and Grenander (1991), Amit (1991), Rosenthal (1995a) and Polson (1996).

We argue that because of the fact that theoretical convergence times vary widely (especially in high dimensional problems), and in the absence of any general techniques for *a priori* prediction of run lengths, it is necessary to carry out some form of statistical analysis in order to assess convergence. We call such procedures convergence diagnostics. There exist various informal methods for the diagnosis of convergence for example Gelfand's *thick pen technique* (Gelfand *et al*, 1990) , and the use of quantile plots and autocorrelation plots as suggested in Gelfand and Smith (1990). These methods are extremely easy to implement, and can provide a feel for the behaviour of the Markov chain.

However, many more elaborate methods have been proposed in the literature. All of these methods can in some way claim to be more reliable assessors of convergence. However to what extent do they really help? Some of these methods involve considerable extra implementational complication and computational expense. Is this worthwhile? Which convergence diagnostic procedures can be recommended in certain contexts, and for particular classes of problems? This paper will attempt to consider these issues.

As with all statistical procedures, any convergence diagnostic technique cannot be guaranteed to successfully diagnose convergence. In particular, for slowly mixing Markov chains, convergence diagnostics are likely to be unreliable since their conclusions will be based upon output from only a small region of the state space. These ideas are formalised by Asmussen *et al* (1992). Therefore it is important to emphasise that any convergence diagnostic procedure should not be unilaterally relied upon. Moreover, the use of additional methods for increasing understanding of the target distribution are strongly recommended. There are a number of useful techniques for *a priori* exploration of the target distribution in order to suggest an appropriate Markov chain for simulation and an appropriate collection of starting values. These methods include Simulated Tempering (see Geyer and Thompson, 1995), Simulated Annealing (see for example Jennison, 1993 and Applegate *et al*, 1990 in this context) and other mode hunting methods such as that proposed by Gelman and Rubin (1992).

There are also a number of methods which, whilst not actually diagnosing convergence, attempt to "measure" the performance of any particular sampler. These can be used either instead of or in addition to the diagnostic methods to be discussed in this paper. Such methods include that described by Brooks (1997) which can be applied to problems where the stationary distribution of the sampler is log-concave. In this case, Brooks (1997) shows how an upper bound can be put on the proportion of the parameter space covered by a particular output sequence. This provides a measure as to how well the chain has explored the parameter space, and an indication as to well inference based upon the sampler output may reflect characteristics of the stationary

distribution.

Before we introduce and discuss the various diagnostic methods, we shall briefly describe the two most common MCMC algorithms, namely the Gibbs sampler and the Metropolis Hastings algorithm. We shall also describe the notation that we will adopt throughout the paper.

We let $\boldsymbol{X}^t \in E \subseteq \mathbb{R}^p$ denote the (p-dimensional) state of an arbitrary chain at time t. Often, we will refer to several chains run simultaneously, in which case we let $\boldsymbol{X}_i^t$ denote the state of chain i at time t, with $X_{i(j)}^t$ denoting the jth element of the vector $\boldsymbol{X}_i^t$. It will also be necessary, on occasion, to refer not to the state of a Markov chain directly, but some function of $\boldsymbol{X}$. Such functions will generally be denoted by $\theta_i^t = \theta(\boldsymbol{X}_i^t)$.

1.1 The Metropolis Hastings Algorithm

To construct a Markov chain $\{\boldsymbol{X}^0, ..., \boldsymbol{X}^t, ...\}$ with state space E, and equilibrium distribution π, the Metropolis Hastings algorithm obtains state $\boldsymbol{X}^{t+1}$ from state $\boldsymbol{X}^t$ as follows. Let $\mathcal{P}$ be an arbitrary Markov transition kernel density (normally either with respect to Lebesgue measure, counting measure or a combination of the two), and let $\alpha(\boldsymbol{x}, \boldsymbol{y})$ to be a function taking values in $[0, 1]$. Then, given that the chain is in state $\boldsymbol{X}^t = \boldsymbol{x}$ at time t, the Metropolis Hastings algorithm generates a candidate value $\boldsymbol{y}$ for the next state $\boldsymbol{X}^{t+1}$, from the distribution $\mathcal{P}(\boldsymbol{x}, \cdot)$. With probability $\alpha(\boldsymbol{x}, \boldsymbol{y})$ this candidate is accepted as the new state so that $\boldsymbol{X}^{t+1} = \boldsymbol{y}$, otherwise it is rejected and the chain remains at $\boldsymbol{X}^{t+1} = \boldsymbol{x}$.

In this way we obtain a Markov chain with transition kernel

$$\mathcal{K}_H(\boldsymbol{x}^t, \boldsymbol{x}^{t+1}) = \begin{cases} \mathcal{P}(\boldsymbol{x}^t, \boldsymbol{x}^{t+1})\alpha(\boldsymbol{x}^t, \boldsymbol{x}^{t+1}) & \boldsymbol{x}^{t+1} \neq \boldsymbol{x}^t \\ 1 - \int_D \mathcal{P}(\boldsymbol{x}^t, \boldsymbol{y})\alpha(\boldsymbol{x}^t, \boldsymbol{y})d\boldsymbol{y} & \boldsymbol{x}^{t+1} = \boldsymbol{x}^t \end{cases}$$

Peskun (1973) shows that the optimal form for the acceptance function, in the sense that it rejects suitable candidates less often than other forms, is given by

$$\alpha(\boldsymbol{x}^t, \boldsymbol{x}^{t+1}) = \min\left(\frac{\pi(\boldsymbol{x}^{t+1})\mathcal{P}(\boldsymbol{x}^{t+1}, \boldsymbol{x}^t)}{\pi(\boldsymbol{x}^t)\mathcal{P}(\boldsymbol{x}^t, \boldsymbol{x}^{t+1})}, \quad 1\right) \tag{1}$$

Note that though the choice of transition kernel $\mathcal{P}$, is arbitrary, it must satisfy certain conditions in order to guarantee convergence, see for example Roberts and Smith (1994).

1.2 The Gibbs Sampler

The Gibbs sampler splits the p-dimensional parameter space into $k \leq p$ blocks. Let $\pi(\boldsymbol{x}) = \pi(\boldsymbol{x}_{(1)}, ..., \boldsymbol{x}_{(k)})$, $\boldsymbol{x} \in \mathbb{R}^p$ denote the joint density of $\boldsymbol{x}$ and $\pi(\boldsymbol{x}_{(i)}|\boldsymbol{x}_{(-i)})$ denote the full conditional density for the i^{th} element, $\boldsymbol{x}_{(i)}$, of $\boldsymbol{x}$, given each of the other $k-1$ components, where $\boldsymbol{x}_{(-i)} = \{\boldsymbol{x}_{(j)} : j \neq i\}$ for $i = 1, ..., k$. Then the Gibbs sampler, as proposed by Geman and Geman (1984), proceeds as follows.

We begin by picking an arbitrary starting point $\boldsymbol{x}^0$ and then successively sample from the full conditional $\pi(\boldsymbol{x}_{(i)}|\boldsymbol{x}_{(-i)})$, by sampling $\boldsymbol{x}_{(i)}^1$ from $\pi(\boldsymbol{x}_{(i)} \mid \boldsymbol{x}_{(j)}^1 \; j <$

i, $\boldsymbol{x}^0_{(j)}$ $j > i$), $i = 1, ..., k$, to generate the next state $\boldsymbol{x}^1$. This process is then repeated to produce a sequence $\boldsymbol{x}^0, \boldsymbol{x}^1, ..., \boldsymbol{x}^t, ...$ of states from a Markov chain with transition probability

$$\mathcal{K}_G(\boldsymbol{x}, \boldsymbol{y}) = \prod_{l=1}^{k} \pi(\boldsymbol{y}_{(l)} \mid \boldsymbol{y}_{(j)} \text{ for } j < l, \text{ and } \boldsymbol{x}_{(j)} \text{ for } j > l)$$

and which has stationary distribution, π.

Other algorithms commonly used usually consist of modifications or hybrids of these two methods.

1.3 What is a convergence diagnostic?

A convergence diagnostic procedure is a method for assessing how long to run a Markov chain in order to obtain observations from (or approximately from) the stationary distribution of the Markov chain.

Some of the techniques that we shall discuss in this paper, are rather informal statistical methods based on monitoring selected output from the Markov chain itself. Such procedures can often be implemented with little extra computational expense. We broadly term such approaches under the heading *output analysis*, and these are discussed in Section 2.

Other methods try to exploit aspects of the theoretical properties of the sampler, sometimes requiring the monitoring of more complicated functionals of the Markov chain, or even additional simulation. These methods, though still inherently statistical, are typically more complicated to implement than straightforward output analyses and we discuss such approaches in Section 3.

Probabilistic constructions for Markov chains can sometimes suggest alternative ways to implement Markov chains, which are better geared towards obtaining a stationary sample. We discuss exact and approximate techniques based around these ideas in Section 4.

Finally, no convergence diagnostic techniques are fool proof and *a priori* run length determination techniques are rarely available in many practical problems. However notable exceptions do exist, and we discuss analytic estimates for convergence rates and running times of Markov chains in Section 5, together with statistical methods for estimating Markov chain convergence rates.

However, before we discuss the various different approaches to convergence assessment outlined above we first take a brief look at how their performance might be assessed, in order to compare them and to provide general guidelines as to which techniques might be most suitable in different situations.

1.4 Assessing Diagnostics

Various different criteria can be used to compare different diagnostic methods. Some criteria can be classified as being purely implementational, such as their computational expense and the ease with which the method may be implemented, whereas others, such as the number of replications needed and the class of samplers to which the method can be applied, are more problem specific.

MCMC methodology relies on the fact that as $t \to \infty$, the t-step transition kernel of the sampler converges towards the target density. Thus, it may seem sensible to perform one long run of a single chain, ignoring an initial transient phase, and to form a sample from the remaining observations, see Geyer (1992). When convergence times are known, this single run implementation is inarguably the correct way to implement the algorithm.

The situation is a little less clear when known convergence times are not available. One plausible option is to take several shorter runs of a number of independent chains and form a sample from these observations, ignoring the initial transient phase in each, as advocated by Gelman and Rubin (1992). One might argue that the observations sampled from the end of a long run might be less biased than those from several shorter runs. However, running several chains in parallel, from a wide range of initial starts, can provide useful information about how well the chains are *mixing*, ie; to what extent the output from the individual chains are indistinguishable. Thus diagnostics based upon the output from several replications of the chain may have access to extra information which may lead to more reliable or more accurate assessment of convergence.

An alternative to running multiple replications, which retains many of the superior diagnostic properties of multiple start algorithms, is to use regenerative methods to *restart* the chain at appropriate *regeneration times*. Implementational details of an approach of this type, involving interspersing the Gibbs sampler with independence Metropolis Hastings steps, is described in Mykland *et al* (1995). The effect of using such an approach is that the regenerations allow a single long chain to be split to produce independent segment independent segments. Thus, a multi-replication diagnostic procedure can be adapted for use in this situation. The drawback with this is that it is not possible to monitor the replications as the sampler proceeds, since the replications appear sequentially. Thus, this method of splitting a single chain is only useful if retrospective analysis is acceptable. Another advantage of the use of multiple runs is that, in certain applications at least, the implementation can take advantage of parallel computing technology, whilst the computation of regenerations can be both difficult and computationally expensive.

The choice of best convergence diagnostic is problem specific. In fact some very reasonable types of diagnostic can only be applied to certain types of algorithm. (For instance many of the proposed methods work only for the Gibbs sampler.) Therefore, the range of samplers to which a particular diagnostic method may be applied, is also an important criterion to be considered when choosing a diagnostic.

Similarly, in any particular application, we may be interested in making inference about only a small number of scalar functions of the parameters. In this case, assessment of the convergence of the chain should perhaps be limited to procedures based on monitoring only those functionals of interest. Such approaches are usually easy to implement, but inference about convergence of the full multidimensional distribution can be unreliable. If we are interested in more general properties of π, such as distributions of a number of functionals of interest, or measures of dependence between functionals, then it may be more reliable and more efficient to assess convergence of the joint distribution. Note that whilst convergence of the marginal distribution of a collection of functionals of interest is a necessary condition for overall convergence, it

is not generally sufficient. Therefore diagnostics that try to assess convergence of the whole distribution might generally be preferable if they are computationally feasible.

A closely related issue is that of computational expense. Broadly speaking, full distributional diagnostics are the more computationally expensive, and often require problem specific computer code for their implementation. A diagnostic which is based solely upon the output from the sampler will generally be easier to implement than one which requires knowledge of the kernels for example, since the latter would require re-coding for each problem, being "problem specific" in that respect. Thus, methods which require no knowledge of the mechanism driving the chain and which are based solely upon an analysis of the output are generally somewhat easier to implement.

Finally, one of the most important criteria is that of interpretability. A diagnostic which produces a definitive solution will generally be preferred to one which requires subjective interpretation and/or experience on the part of the user.

With these criteria in mind we will now discuss and compare the various diagnostic methods proposed in the literature, beginning with the simplest procedures for diagnosing convergence of univariate functionals based upon the output from a single chain; considering a collection of multidimensional diagnostics which attempt to assess the convergence of a number of chains run in parallel, and finally considering the more probabilistic methods of convergence assessment.

2 Output Analysis

2.1 Variance Ratio Methods

The method of Gelman and Rubin (1992) consists of analysing m independent sequences to form a distributional estimate for what is known about some random variable, given the observations simulated so far. It provides a basis for an estimate of how close the process is to convergence and, in particular, how much we might expect the estimate to improve with more simulations. The method proceeds as follows. We begin by independently simulating $m \geq 2$ sequences of length $2n$, each beginning at different starting points which are overdispersed with respect to the stationary distribution. We discard the first n iterations and retain only the last n. Then, for any scalar functional of interest, $\theta(\boldsymbol{x})$, we calculate B/n, the variance between the m sequence means, which we denote by $\bar{\theta}_i^{\cdot}$. Thus, we define

$$\frac{B}{n} = \frac{1}{m-1} \sum_{i=1}^{m} (\bar{\theta}_i^{\cdot} - \bar{\theta}_{\cdot}^{\cdot})^2,$$

where

$$\bar{\theta}_i^{\cdot} = \frac{1}{n} \sum_{t=n+1}^{2n} \theta_i^t, \qquad \bar{\theta}_{\cdot}^{\cdot} = \frac{1}{m} \sum_{i=1}^{m} \bar{\theta}_i^{\cdot}$$

and $\theta_i^t = \theta(\boldsymbol{x}_i^t)$ is the t^{th} observation of θ from chain i. We then calculate W, the mean of the m within-sequence variances, s_i^2, each of which is based upon $n-1$ degrees of freedom. Thus, W is given by

$$W = \frac{1}{m} \sum_{i=1}^{m} s_i^2,$$

where

$$s_i^2 = \frac{1}{n-1} \sum_{t=n+1}^{2n} (\theta_i^t - \bar{\theta}_i^{\cdot})^2.$$

We allow for the variability of both $\hat{\mu}$ and $\hat{\sigma}^2$, by using an approximating t-distribution with mean $\hat{\mu}$, variance

$$\widehat{V} = \frac{n-1}{n} W + \left(1 + \frac{1}{m}\right) \frac{B}{n} \tag{2}$$

and degrees of freedom estimated by

$$\widehat{d} = \frac{2\widehat{V}^2}{\widehat{Var}(\widehat{V})},$$

see Gelman and Rubin (1992).

Now, rather than testing the (generally false) hypothesis that the MCMC algorithm has converged, Gelman and Rubin (1992) propose monitoring convergence by estimating the factor by which the estimated scale of the posterior distribution for θ will shrink as $n \to \infty$. This is given by

$$\widehat{R}_c = \frac{d+3}{d+1} \frac{\widehat{V}}{W},$$

see Brooks and Gelman (1997).

Gelman and Rubin (1992) suggest that the value of $\widehat{R}_c$, which is called the *potential scale reduction factor* (PSRF), can be interpreted as a convergence diagnostic as follows. If $\widehat{R}_c$ is large, this suggests that either the estimate of the variance, $\hat{\sigma}^2$, can be further decreased by more simulations, or that further simulation will increase W, since the simulated sequences have not yet made a full tour of the target distribution. Alternatively, if the PSRF is close to 1, they argue that we can conclude that each of the m sets of n simulated observations is close to the target distribution.

This original method has subsequently been generalised by Brooks and Gelman (1997). First of all, they suggest alternative implementations of this univariate diagnostic, by proposing the use of moments other than those of second order. They suggest the calculation of

$$\widehat{R}_s = \frac{\frac{1}{mn-1} \sum_{i=1}^{m} \sum_{t=n+1}^{2n} |\bar{\theta}_i^t - \bar{\theta}_{\cdot}^{\cdot}|^s}{\frac{1}{m(n-1)} \sum_{i=1}^{m} \sum_{t=n+1}^{2n} |\theta_i^t - \bar{\theta}_i^{\cdot}|^s},$$

for $s = 2, 3, 4, \ldots$. Clearly, in the case where $s = 2$, this is directly comparable to the original method. Brooks and Gelman (1997) also suggest an alternative based upon empirical interval lengths.

They construct an interval-based $\widehat{R}$ measure as follows. ¿From each individual chain, take the empirical $100(1-\alpha)\%$ interval, i.e., the $100\frac{\alpha}{2}\%$ and the $100(1-\frac{\alpha}{2})\%$ points of the n simulation draws, thus forming m within-sequence interval length estimates. Then, from the entire set of observations, gained from all chains, calculate

the empirical $100(1-\alpha)\%$ interval, gaining a total-sequence interval length estimate. Finally, evaluate $\widehat{R}$ defined as

$$\widehat{R}_{\text{interval}} = \frac{\text{length of total-sequence interval}}{\text{average length of the within-sequence intervals}}. \tag{3}$$

This method is considerably simpler than the original method. It is very easy to implement and does not even require the existence of second moments, let alone normality. $\widehat{R}_{\text{interval}}$ is still a PSRF, but based upon empirical interval lengths as a measure of information, rather than variance estimates.

An alternative method is also provided by Brooks and Gelman (1997), which generalises the original method to consider more than one parameter simultaneously. Let $\boldsymbol{\theta}$ denote a vector of parameters, then Brooks and Gelman (1997) estimate the posterior variance-covariance matrix by

$$\widehat{\boldsymbol{V}} = \frac{n-1}{n}\boldsymbol{W} + \left(1 + \frac{1}{m}\right)\boldsymbol{B}/n,$$

where

$$\boldsymbol{W} = \frac{1}{m(n-1)} \sum_{j=1}^{m} \sum_{t=n+1}^{2n} (\boldsymbol{\theta}_j^t - \bar{\boldsymbol{\theta}}_j^{\cdot})(\boldsymbol{\theta}_j^t - \bar{\boldsymbol{\theta}}_j^{\cdot})'$$

and

$$\boldsymbol{B}/n = \frac{1}{m-1} \sum_{j=1}^{m} (\bar{\boldsymbol{\theta}}_j^{\cdot} - \bar{\boldsymbol{\theta}}_{\cdot}^{\cdot})(\bar{\boldsymbol{\theta}}_j^{\cdot} - \bar{\boldsymbol{\theta}}_{\cdot}^{\cdot})',$$

denote the (p-dimensional) within and between-sequence covariance matrix estimates of the p-variate functional $\boldsymbol{\theta}$, respectively.

They then show that if λ_1 is the largest eigenvalue of the symmetric and positive definite matrix $\boldsymbol{W}^{-1}\boldsymbol{B}/n$, then

$$\widehat{R}^p = \frac{n-1}{n} + \left(\frac{m+1}{m}\right)\lambda_1$$

bounds above all of the $\widehat{R}_c$ values associated with any of the univariate elements of $\boldsymbol{\theta}$. They call $\widehat{R}^p$ the Multivariate PSRF (MPSRF). The advantage of the MPSRF is inherent in the fact that it reliably summarises each of the univariate measures in a single value, with only a nominal increase in computational expense. However, the authors suggest combining a number of alternative approaches. For example, for high dimensional problems, the MPSRF may be calculated for all parameters, and the PSRF's for each of the parameters of interest.

2.2 Spectral Methods

Functionals of positive-recurrent Markov chains are special cases of stationary time series, so it is natural to use time series methods to try to assess convergence. Here, we discuss two diagnostic methods which use standard techniques from Spectral Analysis in order to gain variance estimates via the spectral density, $S(\omega)$.

2.2.1 Geweke's Spectral Density Diagnostic

Suppose that we wish to estimate the expected value of some functional, $\theta(\boldsymbol{X})$. If we let $\theta^t = \theta(\boldsymbol{X}^{(t+n_0)})$, for $t = 1, ..., n$ then, given the sequence $\{\theta^t\}$, Geweke (1992) suggests that if the chain has converged by time n_0, then we should accept a test of equal location for two subsequences $\{\theta^t : t = 1, ..., n_A\}$ and $\{\theta^t : t = n^*, ..., n\}$, where $1 < n_A < n^* < n$ and $n_B = n - n^* + 1$. Define

$$\bar{\theta}_A = \frac{1}{n_A}\sum_{t=1}^{n_A}\theta^t \quad \text{and} \quad \bar{\theta}_B = \frac{1}{n_B}\sum_{t=n^*}^{n}\theta^t,$$

and let $\widehat{S}_\theta^A(0)$ and $\widehat{S}_\theta^B(0)$ denote consistent spectral sensitivity estimates for $\{\ \theta^t\ :\ t = 1, ..., n_A\ \}$ and $\{\ \theta^t :\ t = n^*, ..., n\}$ respectively, see Ripley (1987) for example. Then, if the ratios n_A/n and n_B/n are fixed, with

$$\frac{(n_A + n_B)}{n} < 1$$

and if the sequence $\{\theta^t\}$ is stationary, then

$$Z_n = \frac{(\bar{\theta}_A - \bar{\theta}_B)}{\sqrt{\frac{1}{n_A}\widehat{S}_\theta^A(0) + \frac{1}{n_B}\widehat{S}_\theta^B(0)}} \xrightarrow{d} N(0,1) \quad \text{as } n \to \infty. \tag{4}$$

We can use this result to test the null hypothesis of equal location, which, if it is rejected, indicates that the chain has not converged by time n_0.

Geweke (1992) suggests taking $n_A = n/10$ and $n_B = n/2$, arguing that these choices meet the assumptions underlying (4) whilst attempting to provide diagnostic power against the possibility that the $\{\theta^t\}$ process was not fully converged early on.

This method could be generalised to make use of multiple replications by using standard Analysis of Variance techniques to test for differences between replications. It might also be generalised to consider the full joint density rather than the marginals by taking the log posterior density as the scalar functional of interest.

In practice, it should be noted that this diagnostic attempts to verify a necessary, but not sufficient, condition for convergence. Thus, it can only inform the user when convergence has not been achieved, and not when it has. Experience with this diagnostic also highlights its sensitivity to the specification of the spectral window. The choice is arbitrary, and there seems to be no general guidelines available. For further discussion of the problems associated with spectral sensitivity estimation, see Ripley (1987).

2.2.2 Heidelberger and Welch's Convergence Diagnostic

The method of Heidelberger and Welch (1983) confines itself to univariate observations and only single replications. It combines earlier methodology, by Schruben (1982) and Schruben *et al* (1983), on the application of statistics distributed as Brownian Bridges to the detection of transient phases, with an earlier run length control procedure described by Heidelberger and Welch (1981a,b).

We begin by assuming that, in the steady state, we have a covariance (or weakly) stationary process. This is a reasonable assumption since we are interested in sequences generated from Markov chains, and which therefore have full stationarity.

In order to test the null hypothesis of stationarity, we first suppose that we have a sequence $\{X^t : t = 1, ..., n\}$ from a covariance stationary process with unknown spectral density, $S(\omega)$. Then, for $n \geq 1$, we let

$$Y_0 = 0, \quad Y_n = \sum_{t=1}^{n} X^t \quad \text{and} \quad \bar{X} = \frac{1}{n} Y_n,$$

and define

$$\widehat{B_n}(s) = \frac{(Y_{[ns]} - [ns]\bar{X})}{(n\widehat{S}(0))^{\frac{1}{2}}} \qquad 0 \leq s \leq 1,$$

where $[a]$ denotes the greatest integer less than or equal to a, and $\widehat{S}(0)$ is an estimate of the spectral density formed from only the second half of the currently considered data, in order to avoid possible over-estimation in the presence of an initial transient.

Then, for large n, $\widehat{B_n} = \{\widehat{B_n}(s) : 0 \leq s \leq 1\}$ will be distributed approximately as a Brownian Bridge, and we can use various statistics such as the Cramer-von Mises Statistic (von Mises, 1931), the Kolmogorov-Smirnov statistic (Kolmogorov, 1933), or Schruben's statistic (Schruben *et al*, 1983) to test the null hypothesis of stationarity. Heidelberger and Welch (1983) suggest an iterative procedure, based upon these statistics, in order to estimate the length of the burn-in.

This diagnostic uses only single replications and considers only univariate marginals, one at a time. The method may be generalised by making use of the multi-dimensional and multi-sample analogues of the Kolmogorov-Smirnov statistic, for example. See Conover (1965, 1967), Chorneyko and Zing (1982), Fasano and Franceschini (1987) and Ahmad (1976), for example.

Returning to the proposed diagnostic, it should be noted that the stationarity tests will have very little power to detect an initial transient when the run length is shorter than the length of that transient, ie; when the whole sample sequence is within the transient phase. It might also be noted that the sequential use of hypothesis tests should not be performed without some correction factor being combined with the test statistic to allow for the fact that multiple tests are being performed, see Brittain (1987). Finally, we note that, as with Geweke's method, we need to estimate the spectral density at frequency zero. As before, the diagnostic is highly dependent upon this estimate. However, Heidelberger and Welch (1981a) suggest a suitable estimator for use with this problem.

2.3 Yu and Mykland's CUSUM Method

The CUSUM method is a graphical method which monitors convergence via a CUSUM plot based upon the sampler output. It was proposed by Yu and Mykland (1997) and further discussed in Yu (1995a) and Brooks (1996). It can be applied to any sampler and can be implemented by generic problem-independent code.

Given the output $\{\boldsymbol{x}^1, ..., \boldsymbol{x}^n\}$, we begin by discarding the initial n_0 iterations, which we believe to correspond to the burn-in period. We then construct CUSUM path plots for some scalar summary statistic θ, as follows:

1. Calculate

$$\widehat{\mu} = (n - n_0)^{-1} \sum_{t=n_0+1}^{n} \theta(\boldsymbol{x}^t).$$

2. Calculate the CUSUM or partial sum

$$\widehat{S}_T = \sum_{t=n_0+1}^{T} [\theta(\boldsymbol{x}^t) - \widehat{\mu}] \quad \text{for } T = n_0 + 1, ..., n.$$

3. Plot $\{\widehat{S}_T\}$ against T for $T = n_0 + 1, ..., n$, connecting successive points by line segments.

Yu and Mykland (1997) argue that the speed with which the chain is mixing is indicated by the smoothness of the resulting CUSUM plot, so that a smooth plot indicates slow mixing, whilst a "hairy" plot indicates a fast mixing rate for θ. They justify their argument by referring to the theoretical work of Lin (1992, p323) which examines the behaviour of partial sums of ϕ-mixing sequences.

In practice, the assessment of smoothness based upon a single plot is, at best, subjective. In order to improve the method, in terms of its subjectivity, Yu and Mykland (1997) suggest adding a "benchmark" CUSUM plot, based upon a sequence of *iid* normal variates with mean and variance matched to estimated moments of the θ sequence. The authors argue that this plot approximates, to the second order, the "ideal" CUSUM path for an *iid* sequence from the target distribution. Thus, a favourable comparison of the two plots, in terms of smoothness and size of excursion, indicates that the chain is mixing well in terms of the statistic, θ. Brooks (1996) proposes an alternative implementation in an attempt to make this diagnostic more objective.

Let us first obtain a mathematical definition of "hairiness" in the sense discussed by Yu and Mykland (1997). A smooth plot is formed from line segments with the same or similar slope, whilst a perfectly "hairy" plot will consist of line segments which alternately have positive and negative slope, so that each point corresponds to a local optimum. Thus, if we count the number of such points, we get an index of "hairiness" associated with the plot.

If we define

$$d_T = \begin{cases} 1 & \text{if } S_{T-1} > S_T \text{ and } S_T < S_{T+1} \\ & \text{or } S_{T-1} < S_T \text{ and } S_T > S_{T+1} \\ 0 & \text{else} \end{cases}, \tag{5}$$

for all $T = n_0 + 1, ..., n - 1$. Then,

$$D_{n_0,n} = \frac{1}{n - n_0} \sum_{T=n0+1}^{n-1} d_T$$

takes values between 0 and 1, where a value of 0 indicates a totally smooth plot and a value of 1, indicates maximum "hairiness".

Brooks (1996) shows that we can treat $D_{n_0,n}$ as a binomial outcome with mean $\frac{1}{2}$ and variance $\frac{1}{4(n-n_0)}$. The Law of Large Numbers ensures that, for large $n - n_0$, the $D_{n_0,n}$ will be approximately normal, so that we can detect a lack of convergence if the $D_{n_0,n}$ statistic lies beyond the bounds

$$\frac{1}{2} \pm Z_{\alpha/2}\sqrt{\frac{1}{4(n-n_0)}}. \tag{6}$$

The original method, whilst quite simple and problem independent, is rather subjective in its graphical interpretation. By introducing the D sequence, it is possible to make the method more quantitative, though this is at the expense of an increase in computation.

Thus, there are various methods which use only the output from one or more replications of the Markov chain in order to diagnose convergence. These are entirely problem-independent methods and are therefore very simple to use. However, a number of problem-specific methods have also been suggested, which use additional information in the form of the transition kernel of the chain, in order to diagnose convergence. We shall discuss these in the next section.

3 Empirical Kernel Based Methods

3.1 Weighting Methods

Here we discuss two very similar methods for detecting convergence by calculating weighting functions based upon conditional density estimates. The first was proposed by Ritter and Tanner (1992) and attempts to detect convergence of the full joint distribution by monitoring importance weights, calculated from replications of a Gibbs sampler. It was one of the first to try and assess convergence of the joint distribution and is based upon the method of Importance Sampling, see Ripley (1987) and Geweke (1989).

If $\mathcal{K}(\cdot,\cdot)$ denotes a Gibbs transition density, then we can denote the density of $\boldsymbol{X}^t$ by $p^t(\boldsymbol{x})$, where

$$p^t(\boldsymbol{x}) = \int_E \mathcal{K}(\boldsymbol{y},\boldsymbol{x})p^{t-1}(d\boldsymbol{y}), \qquad p^1(\boldsymbol{x}) = \int_E \mathcal{K}(\boldsymbol{y},\boldsymbol{x})\pi_0(d\boldsymbol{y}), \tag{7}$$

and $\pi_0(\cdot)$ denotes the starting distribution for the chain, commonly taken to be a point mass at some point $\boldsymbol{x}^0$, say.

If we denote the non-normalised stationary distribution by $\tilde{\pi}$ then, at convergence, the importance weights

$$w^t(\boldsymbol{x}) = \frac{\tilde{\pi}(\boldsymbol{x})}{p^t(\boldsymbol{x})} \tag{8}$$

should be roughly constant. Of course, $p^t(\boldsymbol{x})$ is unavailable explicitly, but we can gain an unbiased Monte Carlo approximation in the form of the Rao-Blackwellised

estimator

$$\hat{p}^t(\boldsymbol{x}) = \frac{1}{m}\sum_{j=1}^{m} \mathcal{K}(\boldsymbol{x}_j^{t-1}, \boldsymbol{x}), \tag{9}$$

where $\{\boldsymbol{x}_j^{t-1} : j = 1, ..., m\}$ is a sample from p^{t-1}. Ritter and Tanner (1992) suggest sampling from p^{t-1} directly, by drawing m observations at the $(t-1)^{th}$ iteration, ie; from m parallel replications. This can be very computationally expensive. Alternatively, we may reduce this expense by using the last m states of the chain, $\{\boldsymbol{x}^{t-m}, ..., \boldsymbol{x}^{t-1}\}$ since, at convergence, this will simply be a sample from π. This idea can also be extended to the Metropolis Hastings algorithm by a suitable re-definition of p^t (see Gelfand *et al*, 1992), and to hybrid algorithms such as Metropolis-within-Gibbs, see Müller (1991).

Ritter and Tanner (1992) propose using these importance weights as a convergence diagnostic by running multiple replications of the chain and calculating the weights at equally spaced iterations. Then, for each considered iteration, a histogram of the weights from the different replications is produced until the plots move towards a "spike distribution", indicating convergence.

There are two main problems with this method. Firstly, if the dimension of $\boldsymbol{x}$ is large then we need large m in order for $\hat{p}^t$ to be a good estimate of p^t. Coupled with the need to perform a fairly large number of replications, this means that the method may become computationally very expensive. The need for multiple replications stems from the fact that, since the normalising constant for π is not known, we will not know whether the $\hat{w}^t$'s, (as defined in (8) and substituting $\hat{p}^t$ for p^t), tend to the correct value as $t \to \infty$. It is quite possible that the $\hat{w}^t$'s may remain constant, whilst a mass of π remains unexplored. However, if we use multiple replications and check that they all converge to the same value, then this provides some reassurance that the entire mass of π has been explored. Alternatively, in the special case where we can partition $\boldsymbol{x}$ into $(\boldsymbol{x}_{(1)}, \boldsymbol{x}_{(2)})$ and where the conditionals $p(\boldsymbol{x}_{(1)}|\boldsymbol{x}_{(2)})$ and $p(\boldsymbol{x}_{(2)}|\boldsymbol{x}_{(1)})$ are known standard densities, we may use the weights

$$\hat{w}^t(\boldsymbol{x}_{(1)}, \boldsymbol{x}_{(2)}) = \frac{p(\boldsymbol{x}_{(1)}|\boldsymbol{x}_{(2)})\hat{p}^t(\boldsymbol{x}_{(2)})}{p(\boldsymbol{x}_{(2)}|\boldsymbol{x}_{(1)})\hat{p}^t(\boldsymbol{x}_{(1)})}, \tag{10}$$

where

$$\hat{p}^t(\boldsymbol{x}_{(1)}) = \frac{1}{m}\sum_{j=1}^{m} p(\boldsymbol{x}_{(1)}|\boldsymbol{x}_{j(2)}^t), \tag{11}$$

and monitor the sequence $\{\hat{w}^t\}$ until it reaches stationarity with mean 1, since both the numerator and denominator are estimates of $p^t(\boldsymbol{x})$, see Gelfand (1992). This eliminates the need for multiple replications and greatly reduces computational expense.

A similar method to this, called the *Gibbs sampler difference convergence criterion*, was proposed by Zellner and Min (1995) who define their weight function by

$$\hat{w}^t = \hat{p}^t(\boldsymbol{x}_{(1)})p(\boldsymbol{x}_{(2)}|\boldsymbol{x}_{(1)}) - \hat{p}^t(\boldsymbol{x}_{(2)})p(\boldsymbol{x}_{(1)}|\boldsymbol{x}_{(2)}), \tag{12}$$

where each of the densities are normalised, and the $\hat{p}^t$ are as defined in (11). Zellner and Min (1995) argue that if the $\hat{w}^t$ are small, then this indicates that the estimated marginals have converged to the correct values.

Note the clear similarities between this weight function and that proposed by Ritter and Tanner (1992), except here we look at differences rather than a ratio. In fact Zellner and Min (1995) go on to discuss an alternative approach which actually looks at ratios rather than differences, and is directly comparable to the extension of Ritter and Tanner's diagnostic given in (10). However, Zellner and Min (1995) go further than Ritter and Tanner (1992) by attempting to formalise their method within a Bayesian framework, and obtain a posterior density for w, the asymptotic mean of the $\hat{w}^t$, so as to calculate posterior odds relating to the hypotheses

$$H_0 : w = 0 \qquad \text{vs} \qquad H_1 : w \neq 0. \tag{13}$$

Clearly, one of the main drawbacks with this method, as with that of Ritter and Tanner (1992), is the assumption that we are able to derive explicit conditional posterior densities. This is a strong assumption and restricts the method to only a small class of problems. The method is also restricted to the Gibbs sampler, with no obvious extension to other MCMC algorithms. Also, as with that of Ritter and Tanner, the estimates of the conditionals may be expensive to calculate.

3.2 Normed Distance Criteria

We shall now discuss four methods which use the transition kernel of the sampler to assess the convergence of the joint density, by using the output from multiple replications.

3.2.1 Liu, Liu and Rubin's L^2 Convergence Diagnostic

Liu *et al* (1993) define a scalar global control variable based upon iterates from multiple replications of a Markov chain, reducing the assessment of the convergence of a multi-dimensional Gibbs Sampling procedure to that of a one-dimensional random variable. This control variable incorporates information on the overall convergence of the chain to the target distribution π, and so is more comprehensive than one based solely upon subcomponents. Liu *et al* (1993) suggest that the method of Gelman and Rubin (1992) can be used to judge the convergence of this control variable and thus convergence of the chain to its stationary distribution.

Liu *et al* (1993) construct this control variable as follows. At iteration t, we construct a set of $m(m-1)$ values of the variable U, each using two independent parallel chains, i and j, given by

$$U_{ij}^t = \frac{\pi(\boldsymbol{x}_j^t)}{\pi(\boldsymbol{x}_i^t)} \frac{\mathcal{K}(\boldsymbol{x}_j^{t-1}, \boldsymbol{x}_i^t)}{\mathcal{K}(\boldsymbol{x}_j^{t-1}, \boldsymbol{x}_j^t)} \qquad i \neq j, \quad i, j = 1, ..., m.$$

Then,

$$U^t = \frac{1}{m(m-1)} \sum_{i \neq j} U_{ij}^t \tag{14}$$

is an unbiased estimator of the L^2 distance between $p^t(\cdot)$ and $\pi(\cdot)$.

Given the global control variable, U^t, Liu *et al* (1993) propose two ways to utilise it as a convergence diagnostic. The first is to separate the chains into $\frac{m}{2}$ independent pairs and construct the $\{U^t\}$ sequence for each pair. They then suggest using the method of Gelman and Rubin (1992) to diagnose the convergence of these $\frac{m}{2}$ paired sequences individually, and comparing the results. The second method is to produce the sample cumulative density function of U, based upon iterations $t = n+1, ..., 2n$ and use Smirnov-statistic based tests to assess convergence in a manner similar to that proposed by Heidelberger and Welch (1983).

One drawback of methods based upon this global control variable is that its variance can be extremely large, making the estimate of the L^2 distance in (14) unreliable. This problem can be overcome by constructing U from the $\log U_{ij}$, but then we lose the direct interpretability which makes this method attractive. Alternatively, we can improve the estimate by taking a large value of m, but this is computationally expensive. A second drawback which is encountered when we use the Gelman and Rubin-style implementation is in the justification of the assumption of normality for the U^t, though some of the extensions proposed by Brooks and Gelman (1997) may overcome this problem.

This method also has problems with interpretability and is problem specific, requiring different code to be written for each problem in order to produce the required output.

3.2.2 Roberts' L^2 Convergence Diagnostic

This diagnostic uses a density ratio statistic to assess the L^2 convergence of the distribution of the chain as a whole, using samples from several independent replications of a Markov chain with finite Hilbert-Schmidt norm. Thus, the method is essentially restricted to the analysis of the output from a reversible Gibbs sampler, see Roberts (1994).

Here we run m replications of the chain from starting points $\boldsymbol{x}_i^0 \quad i = 1, ..., m$, where each starting point produces a sample path $\{\boldsymbol{x}_i^t : t = 0, 1, 2, ...\}$, and define our scalar control variable by

$$\chi_{ij}^t = \frac{\mathcal{K}(\boldsymbol{x}_i^0, \boldsymbol{x}_j^{2t-1})}{\pi(\boldsymbol{x}_j^{2t-1})},$$

where $\mathcal{K}$ is the transition density for the chain with equilibrium distribution, π. Roberts suggests the following diagnostic procedure for the case where multiple replications are possible. We define

$$D_t = \frac{1}{m}\sum_{i=1}^{m} \chi_{ii}^t \quad \text{and} \quad I_t = \frac{1}{m(m-1)}\sum_{i \neq j} \chi_{ij}^t,$$

where the dependence term, D_t, measuring the dependence of the current value of the chain upon the starting point, gives us an indication as to how well the chain has mixed, and I_t denotes an interaction term. Roberts (1994) then suggests that we might monitor D_t and I_t until they both reach stationarity and have similar location.

We have the same drawback with this method as with that of Liu *et al*, in that the variance of the χ_{ij}^t can be very large. As before, this can be reduced by considering

the $\log \chi_{ij}^t$ which, in this case, retains an appealing interpretation in terms of the Kullback-Leibler distance. Another drawback is associated with the interpretation of the diagnostic. The output from the diagnostic is in the form of two sequences with different variances, but which have the same asymptotic location. This complicates the decision as to whether or not convergence has been achieved. However, the I_t sequence is less variable than the D_t sequence thus, in practice, it may be easier to interpret convergence if we define

$$I_t(i) = \frac{1}{m-1} \sum_{j \neq i} \chi_{ij}^t,$$

and monitor the $I_t(i)$ sequences for $i = 1, ..., m$.

3.2.3 Yu's L^1 Diagnostic

The method proposed by Yu (1995b), attempts to diagnose convergence of the full density by monitoring a plot based upon the output of a single replication of the chain.

If $\{x^t : t = 0, 1, 2, ...\}$ is the output from a p-dimensional Markov chain with stationary density $\pi(\boldsymbol{x}) = c\tilde{\pi}(\boldsymbol{x})$, where $\tilde{\pi}$ is known and c is the unknown normalisation constant for π, then Yu suggests the following procedure.

First, we define a kernel estimator of π, with bandwidth, b_n, by

$$\widehat{\pi}_n(\boldsymbol{x}) = \frac{1}{n} \sum_{t=1}^{n} h_{b_n}(\boldsymbol{x} - \boldsymbol{x}^t), \text{ where } h_\sigma(\boldsymbol{x}) = \frac{1}{\sigma^p} K\left(\frac{|\boldsymbol{x}|}{\sigma}\right),$$

$K(\cdot)$ is some one-dimensional bounded symmetric kernel, such that $\int_{\mathbb{R}^p} K(|\boldsymbol{x}|) d\boldsymbol{x} = 1$, and $|\cdot|$ denotes the Euclidean norm in $\mathbb{R}^p$.

Then, we define an estimate of the normalisation constant by

$$\widehat{c}_\sigma = \frac{1}{n(n-1)} \sum_{t=1}^{n} \sum_{\tau \neq t} \frac{h_\sigma(\boldsymbol{x}^t - \boldsymbol{x}^\tau)}{\tilde{\pi}(\boldsymbol{x}^\tau)},$$

and choose some subset, A, of the support of π, and an increment value, t_0. Beginning at time $t = t_0$, we estimate the optimal bandwidth, b_t, for the kernel density estimate, by the method described by Silverman (1986), and based upon the data, $\boldsymbol{x}^0, ..., \boldsymbol{x}^t$. Given b_t, we estimate the L^1 distance over A between the kernel density estimator $\widehat{\pi}_t$ and π, by

$$\widehat{D}_t(A) = \int_A |\widehat{\pi}_t(\boldsymbol{x}) - \widehat{c}_{b_t} \tilde{\pi}(\boldsymbol{x})| \, d\boldsymbol{x}.$$

This process is repeated for times $t = t_0, 2t_0, 3t_0, ...$, until a plot of the $\widehat{D}_t(A)$ values settles below some critical value, which Yu suggests should be taken as 0.3.

Like the other kernel-based methods, the method of Yu (1995b) is both problem specific (in that the form of $\tilde{\pi}$, is required) and computationally expensive. The output is rather difficult to interpret, though this can be improved by running several independent chains and plotting the resultant $\widehat{D}_t(A)$ sequences together. However,

the critical value of 0.3, is rather arbitrary and may be better suited to some problems than others.

It should also be noted that the choice of the set A is of vital importance since, as Yu points out, the diagnostic plot may falsely diagnose convergence if both the sample path and the set A omit the same mode(s) of π. However, Yu's method has an advantage over those of Liu *et al* and Roberts in that the $\widehat{D}_t(A)$ statistic stabilises sample-wise, rather than in expectation, ie; it is the expectation of the U_{ij}^t and χ_{ij}^t statistics that provide estimates of the respective L^2 distances, whereas $\widehat{D}^t$ is a direct estimate of the corresponding L^1 distance. In the next section, we provide another method for diagnosing convergence, by estimating L^1 distances between transition densities of different chains. Like that of Liu, this method also has the advantage of providing a statistic which stabilises sample-wise.

3.2.4 The Total Variation Diagnostic

Brooks *et al* (1997) suggest a new approach to diagnosing convergence which attempts to obtain an upper bound to the L^1 distance between full-dimensional kernel estimates from different chains, using ideas similar to those of rejection sampling, (see Smith and Gelfand, 1992).

Brooks *et al* (1997) suggest that we run m independent chains and split each chain into blocks of n_0 observations, with $\boldsymbol{x}_i^t$ denoting the state of chain i at time t. Then for the l^{th} block of the i^{th} chain we can define

$$K_{il}(\boldsymbol{x}) = \sum_{t=(l-1)n_0+1}^{ln_0} \frac{\mathcal{K}(\boldsymbol{x}_i^t, \boldsymbol{x})}{n_0} \tag{15}$$

where $\mathcal{K}(\boldsymbol{x}, \boldsymbol{y})$ is the one-step transition kernel for the chain moving from state $\boldsymbol{x}$ to $\boldsymbol{y}$. Thus, $K_{il}(\boldsymbol{x})$ is a Rao-Blackwellised estimate of the density of $\boldsymbol{X}$ in block l of chain i. Finally, we form a mean between chain distance by

$$B_l = \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i} \hat{r}_{ij}(l) \tag{16}$$

where $\hat{r}_{ij}(l)$ is given by

$$\hat{r}_{ij}(l) = 1 - \min\left(1, \frac{K_{il}(\boldsymbol{x})}{K_{jl}(\boldsymbol{x})}\right)$$

for some $\boldsymbol{x}$ sampled from the density $K_{jl}(\cdot)$, and estimates the L^1 distance between $K_{il}(\cdot)$ and $K_{jl}(\cdot)$, see Brooks *et al* (1997).

Brooks *et al* (1997) suggest monitoring the B_l values over blocks $l = 1, 2, ...$ to assess convergence. They highlight the interpretability of the diagnostic suggesting that convergence is indicated by a characteristic jump point in the B_l value, which is a physical manifestation of the "cut-off phenomenon" of Diaconis (1988, p91). The diagnostic can also be applied to a wide variety of samplers, though they discuss only the Metropolis Hastings algorithm and the Gibbs sampler.

One advantage of this method over the alternative methods for diagnosing convergence of the joint density, is that it is easily adapted to measure convergence of

particular multivariate conditional densities. This may be particularly useful in high dimensional problems, where assessing convergence of the full joint density may be prohibitively expensive. In this case, the L^1 diagnostic can be used to assess convergence of the joint density of a subset of the parameters, namely those that are of principal interest.

In terms of computational expense, this method compares favourably with the methods of Liu *et al* (1993) and Roberts (1994). However, for most problems, the evaluation of the diagnostic sequence typically increases the execution time for the sampler by an order of magnitude. One might argue that the added computational expense is acceptable in order to gain a reliable and interpretable diagnostic. It is really a matter of how much we are willing to "pay" in order to obtain such a diagnostic.

4 Regeneration and coupling methods

Here, we consider methods for diagnosing convergence motivated by the two important probabilistic constructions: coupling times and regeneration times.

First consider the following theoretical construction that motivates the use of coupling diagnostics. Let $\mathbf{X}_1$ and $\mathbf{X}_2$ be discrete time processes, marginally having distributions which are realisations from the Markov chain of interest, started at different starting distributions, say $\mathbf{X}_i^0 \sim \mu_i^0, \quad i = 1, 2$. We do not prescribe any particular dependence structure between the two chains, except to stipulate that the joint process $(\mathbf{X}_1, \mathbf{X}_2)$ be Markov. Indeed $\mathbf{X}_1$ and $\mathbf{X}_2$ will commonly be constructed in a highly dependent way.

Such a construction is called a *coupling* of $\mathbf{X}_1$ and $\mathbf{X}_2$, and if we let $\tau = \inf\{t : \mathbf{X}_1^t = \mathbf{X}_2^t\}$, τ is known as the coupling time of the construction. If μ_i^0 is taken to be the stationary distribution π, then the celebrated *coupling inequality* (see for example Lindvall, 1992 p.12) ensures that

$$\sup_A \left(\mathbb{P}[\mathbf{X}_2^t \in A] - \pi(A) \right) \leq \mathbb{P}[\tau > t] \ .$$

Therefore if the probability that coupling has occurred by time t is high, then the distribution of $\mathbf{X}_2^t$ is a good approximation to π. This construction motivates the following diagnostic introduced in Johnson (1996) and extended in Reutter and Johnson (1995).

4.1 The Johnson diagnostic

The Johnson diagnostic was first proposed for the Gibbs sampler, though refinements due to Reutter and Johnson (1995) extend its natural application to more general Hastings-Metropolis situation.

Let $\boldsymbol{U} = \{u_{(j)}^t : \ j = 1, ..., k, \ t = 1, 2, ...\}$ be an array of independent standard uniform random variables, and let $\boldsymbol{x}^t = \{\boldsymbol{x}_{(j)}^t : j = 1, ..., k\}$ denote the t^{th} sample vector in a sequence of draws from the Gibbs sampler. Finally, let $F(\cdot|\cdot)$ denote the conditional cumulative distribution associated with π, with corresponding quantile function, $F^{-1}(\cdot|\cdot)$. Then, the coupled Gibbs sampler algorithm can be defined as follows.

STEP 1. CHOOSE m STARTING VECTORS, $\boldsymbol{x}_i^0 \quad i = 1, ..., m$ FROM A DISTRIBUTION OVERDISPERSED WITH RESPECT TO π, AND SET $t = 1$.

STEP 2. FOR $j = 1, ..., k, \quad i = 1, ..., m$ AND CONTINUOUS $x_{(j)}^t$, SET

$$x_{i(j)}^t = F^{-1}(u_{(j)}^t | x_{i(1)}^t, ..., x_{i(j-1)}^t, x_{i(j+1)}^{t-1}, ..., x_{i(k)}^{t-1}),$$

WHERE $x_{i(j)}^t$ IS THE j^{th} COMPONENT OF THE VECTOR $\boldsymbol{x}_i^t$, THE STATE OF CHAIN i AT ITERATION t.

STEP 3. IF $|\boldsymbol{x}_i^t - \boldsymbol{x}_l^t| \leq \epsilon \quad \forall\, i \neq l,\ i, l \in \{1, ..., m\}$ AND SOME $\epsilon > 0$, THEN THE PATHS HAVE CONVERGED AFTER t ITERATIONS, AND WE STOP. OTHERWISE RETURN TO STEP 2 AND INCREASE t BY 1.

Variations on this theme involve situations where the inversion in Step 2 is not needed because of the existence of fully conjugate prior distributions. In this case, we can simply make the random draws for each chain by feeding the same seed to the random number generator. However, if this is not the case, then the implementation of this diagnostic can become rather complicated and computationally expensive.

The diagnostic produces appealing and easily interpretable graphical output, and in many situations can be surprisingly easy to implement, despite the need to produce problem specific computer code for its implementation. However since the coupling inequality does *not* state that $_2 \sim \pi$ (which is false in general), in order to attempt to justify this diagnostic using the coupling inequality, the procedure described in (1)-(3) needs to be carried out many times, and even then uncertainty remains about how well we can find an over-dispersed starting distribution that approximates π. The overall procedure may therefore very computationally expensive.

4.2 The Propp and Wilson construction

A more sophisticated attempt to employ the coupling inequality is due to Propp and Wilson (1996). Suppose that the state space for the Markov chain E say, is finite, and that we are able to carry out the following 'backwards' simulation. For a fixed $N \in \mathcal{N}$, start a Markov chain sample path from each point in E, at time $-N$. Call these chains $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{|E|}\}$ say. The $|E|$ chains can be constructed in a dependent way, though once coupled, we stipulate that they remain together, that is for $N \leq s < t \leq 0$,

$$\boldsymbol{X}_i^s = \boldsymbol{X}_j^s \implies \boldsymbol{X}_i^t = \boldsymbol{X}_j^t \ .$$

We examine the set $\{\boldsymbol{X}_i^0,\ i \in E\}$. It is possible that all $|E|$ chains have coupled by time 0, so that $\{\boldsymbol{X}_i^0,\ i \in E\}$ consists of a unique point. In this case, there is no dependence of $\boldsymbol{X}_i^0$ on its initial value i, and it follows that the distribution of $\boldsymbol{X}_i^0$ is π. If $\{\boldsymbol{X}_i^0;\ i \in E\}$ is not a singleton, we repeat the simulation , this time started at time $-2N$, though using the same sample paths on the time period $\{-N, \ldots, 0\}$ as for the first simulation. Now we continue to double the time for the simulation until we reach a situation where $\{\boldsymbol{X}_i^0;\ i \in E\}$ is a singleton.

This procedure may seem rather complicated, though its subtle differences from the procedure suggested in Johnson (1996) ensure that in this case, as soon as $\{\boldsymbol{X}_i^0;\ i \in E\}$ is a singleton, the distribution of that point is *exactly* the target distribution, π. Thus,

we can exactly simulate from π and hence the names often given to these simulation methods: *exact* or *perfect* simulation.

On the other hand, the simulation procedure is very cumbersome and for most practical problems, it is usually difficult if not impossible to simulate chains simultaneously from all possible starting points. The pure application of this idea is therefore (for the moment at least) restricted to a small class of problems where special structure of the state space needs only two processes to be constructed simultaneously, a 'top' and a 'bottom' process. Particularly innovative algorithms of this kind have appeared in the stochastic geometry literature, see Kendall (1997) and Møller (1997).

Strictly speaking, exact simulation is not a convergence diagnostic, it is a construction which, where possible, ensures draws from the target distribution. Despite its current practical limitations, the idea already promises to become an active research area, and undoubtedly practical simulation procedures of this kind (at least for some classes of problems) will eventually become available.

4.3 Methods for Regenerative Chains

Here, we consider the case where we have a Markov chain $\{\boldsymbol{X}^1, ..., \boldsymbol{X}^n\}$, which *regenerates* at times $\{\tau_t : t = 1, \ldots T\}$, and has stationary density π. A Markov chain X is said to regenerate at time τ if the Markov chain sample path before and after τ are independent conditioned on τ. For finite Markov chains, it is easy to spot regeneration times, for example just set τ_t to be the tth visit of the chain to the state i_0. For continuous state spaces, regeneration times are not in general naturally available and need to be constructed artificially.

Suppose that we wish to estimate $\mathbb{E}_\pi(\theta)$, based upon observations from the process, for some function θ. Then, the usual method is to form an ergodic estimate,

$$\frac{1}{N}\sum_{t=1}^{N}\theta(\boldsymbol{x}^t).$$

In this case, the Central Limit Theorem gives us that, if the $\boldsymbol{x}^t$ are *iid* random variables with finite variance, then

$$\frac{1}{\sqrt{N}}\sum_{t=1}^{N}\theta(\boldsymbol{x}^t) - \sqrt{N}\mathbb{E}_\pi\left[\theta(\boldsymbol{x}^t)\right] \xrightarrow{d} N(0, \sigma_\theta^2), \text{ as } N \to \infty.$$

However, since the $\boldsymbol{x}^t$ are realisations from a Markov chain, they will not be independent, nor will they necessarily have identical distributions. Thus, the traditional ergodic estimate of $\mathbb{E}_\pi(\theta)$ is biased and Central Limit Theorems need not apply.

However, since the process is regenerative, if we observe the process for a fixed number of tours T, say, let

$$n_t = \tau_t - \tau_{t-1}$$

and

$$S_t = \sum_{s=\tau_{t-1}+1}^{\tau_t}\theta(\boldsymbol{X}^s) \qquad t = 1, ..., T,$$

then $\sum_{t=1}^{T} S_t / \sum_{t=1}^{T} n_t$ *is* a consistent estimator for $\mathbb{E}_\pi(\theta)$, by the Law of Large Numbers and the Renewal Theorem.

In order to assess convergence of a regenerative chain run for T tours, Mykland *et al* (1995) suggest plotting τ_t/τ_T against t/T, which they refer to as a scaled regeneration quantile, or SRQ plot. Since the $\{\tau_t\}$ form a renewal process with constant probability of regeneration, the τ_t will be uniformly distributed so that, for large T, the SRQ plot should resemble a straight line.

Deviations from a straight line occur when relatively long tours are encountered. This is consistent with a highly positively skewed distribution for the length of the regeneration times, which is in turn indicative of a slowly mixing Markov chain. The deviation from a straight line can be quantified by fitting an intercept-slope model to the observed data and assessing the fit of the model.

The main problem with this method is in the introduction of regeneration points into the MCMC sampler. This is straightforward in the case of a discrete state space, since regenerations can occur each time the chain enters a particular pre-determined state (or set of states). In the general state space case, Mykland *et al* (1995) use Nummelin's Splitting technique (Nummelin, 1984) to introduce regenerations, and show how to introduce regenerations in a limited number of special cases. However Gilks *et al* (1996) show that the artificial construction of such regeneration times is effectively limited to small dimensional problems, so that these extensions have limited practical appeal.

As an alternative diagnostic, Robert (1996) proposes the following method. Given a renewal set (or *atom*), A, Robert shows that

$$\frac{\tau_T}{T} \stackrel{T\to\infty}{\longrightarrow} \mathbb{E}_\pi[\tau_2 - \tau_1] = \mu_A, \text{ say}$$

and, under suitable regularity conditions provided by Robert (1996),

$$\sqrt{T}\sum_{t=1}^{T}(S_t - n_t\mathbb{E}_\pi[\theta(\boldsymbol{x})]) \stackrel{d}{\to} N(0, \sigma_A^2).$$

Now, σ_A^2 can be estimated by

$$\widehat{\sigma}_A^2 = \sum_{t=1}^{T}\left(S_t - n_t\sum_{t=1}^{T}\frac{S_t}{N}\right)^2$$

and μ_A, by N/T. Thus, σ_θ^2 can be estimated by

$$\widehat{\sigma}_\theta^2(A) = \widehat{\sigma}_A^2/\widehat{\mu}_A,$$

for all atoms of the chain. Thus, Robert suggests stopping the simulation once the estimators $\widehat{\sigma}_\theta^2(A)$, stabilise to the same value for a collection of different atoms, $A \in \mathcal{A}$.

The difficulty with this method, as with that of Mykland *et al* (1995), is that it is often very difficult to obtain atoms for chains defined on a general state space. Thus, this method is of limited practical value in the general state space case.

However, for a discrete chain, as we explained earlier, atoms are easy to find, since the chain can be allowed to regenerate each time $\boldsymbol{X}^t = \boldsymbol{X}^*$, for some pre-determined

states, $\boldsymbol{X}^*$. Robert (1996) suggests taking the state which has greatest mass under π or, if the state space is large, a collection of states with "significant" mass under π, in order to ensure a reasonable regeneration rate. Thus, in the discrete case, it is easy to construct several atoms $A_1, ..., A_k$ and calculate $\widehat{\sigma}^2_\theta(A_i)$ for $i = 1, ..., k$, stopping when the sample paths of these statistics converge.

In summary, both regeneration-based methods are severely restricted in the range of problems to which they can be applied. However, for problems defined on a discrete state space, the methods are reasonably effective. The method of Mykland *et al* (1995) can be applied to problems defined on a more general state space, but this involves analytic work in defining a problem specific *split* for the problem at hand. Its implementation is also rather subjective in its interpretation. Robert (1996)'s method is more strictly limited to discrete problems, but is less subjective and requires no analytical work in its implementation.

5 Eigenvalue Bounds and Semi-Empirical Methods

Most (but not all) MCMC algorithms used in practice, converge to stationarity geometrically quickly. That is, there exists a measurable function $V \geq 1$, and a constant $\rho < 1$ such that

$$\|p^t(\boldsymbol{x}, \cdot) - \pi(\cdot)\| \leq V(\boldsymbol{x})\rho^n \tag{17}$$

see for example Roberts and Tweedie (1996). Furthermore, suitably well-behaved Markov chains exhibit an eigen-expansion for its transition probabilities, implying that

$$p^t(\boldsymbol{x}, A) - \pi(A) = a(\boldsymbol{x}, A)\lambda_1^n + O(\lambda_2^n) \tag{18}$$

where $|\lambda_2| < |\lambda_1| < 1$. See Roberts (1994) for a discussion of sufficient conditions for such an expansion. In this section we concentrate on methods which either estimate $a(\cdot)$ and λ_1 or analytically estimate $V(\cdot)$ and ρ.

It is clearly of interest to be able to estimate λ_1, but how useful knowledge about λ_1 is (and indeed how valid the procedure for estimating it) depends upon the existence of an expansion such as (18). Furthermore, we also need $|\lambda_2|$ to be significantly smaller than $|\lambda_1|$ so that its effect on the estimation procedure for λ_1 is negligible.

There are many situations for which MCMC algorithms reasonably satisfy these assumptions, for example Gibbs samplers on densities which are well approximated by Gaussians. However there are others for which (17) holds but (18) does not, for instance the independence sampler on a continuous space (see Smith and Tierney, 1996).

Two rather different procedures have been proposed for estimation of principle eigenvalues in the MCMC context on the basis of expansions of the form (18). A direct method based on (18) is proposed by Garren and Smith (1995). Whereas Raftery and Lewis (1992) propose a method for estimating a probability of interest $\pi(A)$ say, by assessing the convergence of a binary process derived from the original chain.

5.1 Garren and Smith's Convergence Rate Estimator

Suppose $\{\boldsymbol{X}^t\}$ is a sequence from a Markov chain with state space E, and stationary distribution $\pi(\cdot)$, and take some subset $A \subseteq E$. Garren and Smith (1995) define

$$Z^t = I_A(\boldsymbol{X}^t), \qquad \rho_t = \mathbb{E}(Z^t), \quad \text{and} \quad \rho = \lim_{t \longrightarrow \infty} \rho_t.$$

Then, Garren and Smith (1995) attempt to estimate ρ and assume that there exists an expression in terms of the eigenvalues of the transition kernel, of the form

$$\rho_t = \rho + a_2 \lambda_2^t + O(|\lambda_3|^t), \tag{19}$$

where $|\lambda_3| < |\lambda_2| < 1$ and $a_2 \in \mathbb{R}$.

The assumptions underlying (19) are satisfied if firstly, the transition kernel is self-adjoint on an appropriate space with finite Hilbert-Schmidt norm. This condition is not easily checked and, without it, there are many natural MCMC problems where the second eigenvalue is not bounded away from the first so that the estimating equation (19), becomes invalid. However, Roberts (1992) shows that the Hilbert Schmidt condition is satisfied by the *reversible* Gibbs sampler, where each iteration consists of first a *forward* and then a *reverse* sweep.

The second condition is that the two principal eigenvalues have unit multiplicity. For the principal eigenvalue, this is essentially an irreducibility condition and can often be easily checked in specific problems. However, the multiplicity of the second eigenvalue does not have such an intuitive probabilistic interpretation and is more difficult to check.

Given that (19) holds, Garren and Smith (1995) propose the following estimation scheme. Fix integers n_0 and n with $1 \leq n_0 < n$, and run m independent replications of the chain $\{\boldsymbol{X}_i^t : 0 \leq t \leq n\}$ $i = 1, ..., m$, each starting with the same initial distribution f_0, taken to be a point mass at some point $\boldsymbol{x}^0$. Then, define

$$\bar{Z}^t = \frac{1}{m} \sum_{i=1}^{m} I_A(\boldsymbol{X}_i^t),$$

where $\boldsymbol{X}_i^t$ is the t^{th} observation from chain i. Thus, $\bar{Z}^t$ is the sample frequency of the event $\{\boldsymbol{X}_t \in A\}$ over the m replications. Finally, we define estimators $\widehat{\rho}$, $\widehat{a}_2$ and $\widehat{\lambda}_2$ of ρ, a_2 and λ_2 respectively, which minimise

$$S(\rho, a_2, \lambda_2) = \sum_{t=n_0+1}^{n} (\bar{Z}^t - \rho - a_2 \lambda_2^t)^2.$$

Garren and Smith (1995) suggest a heuristic approach to assessing convergence, given these estimates. They note that as n increases the estimates for a_2 and λ_2 become unstable (with the standard errors increasing) and they suggest taking the largest value of n for which plots of both $\widehat{a}_2$ and $\widehat{\lambda}_2$ remain stable. A suitable value for n_0 may then be found by ignoring the $O(|\lambda_3|^t)$ term in (19), and imposing the condition that $|\rho_t - \rho| \leq \epsilon$, for some $\epsilon > 0$, then

$$n_0 = \frac{\log\left(\frac{\epsilon}{|\widehat{a}_2(n)|}\right)}{\log|\widehat{\lambda}_2(n)|}.$$

The main drawback to this diagnostic is the computational expense of the method, due to the problem of choosing a value of m which leads to good estimators of ρ, a_2 and λ_2. Garren and Smith (1995) suggest that m should be "large" in comparison to n, and Garren and Smith suggest taking values as large as 5000. Such large numbers of replications may cause serious problems for many practical applications and results in a considerable loss of computational efficiency. Note also that each of these replications have the same starting point and so they do little to help explore the different regions of the parameter space. Garren and Smith (1995) discuss the possibility of using multiple replications from different starting points, but show that this would require prohibitively long sample runs in order to gain reasonable estimators.

5.2 Raftery and Lewis' Convergence Rate Estimator

Raftery and Lewis (1992) consider the problem of calculating the number of iterations necessary to estimate a posterior quantile from a single run of a Markov chain. They propose a 2-state Markov chain model fitting procedure based upon a pilot analysis of output from the original chain.

Suppose that we wish to estimate a particular posterior quantile for some function θ of a parameter (or set of parameters) $\boldsymbol{X}$, ie; we wish to estimate from observational data the value of u such that

$$\mathbb{P}(\theta(\boldsymbol{X}) \leq u) = q$$

for some pre-prescribed q and so that, given our estimate $\widehat{u}$, $\mathbb{P}(\theta(\boldsymbol{X}) \leq \widehat{u})$ lies within $\pm r$ of the true value, say, with probability p.

Raftery and Lewis (1992) propose a method to calculate first of all the length of the "burn-in" period n_0, and secondly the number of further iterations, n, required to estimate the above probability to within the required accuracy. They suggest that we run the MCMC algorithm for an initial n_0 iterations, which we discard, and then a further n iterations which we thin by storing every s^{th}, and provide a method to determine the values of n_0, n and s. Here, we are interested only with the value of n_0 which estimates the number of iterations required before convergence is "achieved". This can be calculated as follows.

First, we calculate $\theta^t = \theta(\boldsymbol{X}^t)$ where $\boldsymbol{X}^t$ is the state of the chain at time t, and then form $Z^t = I_{(\theta^t \leq u)}$, where I is the indicator function. Then, Z^t is a binary process derived from a Markov chain by marginalisation and truncation but is not a Markov chain itself. We then take the sub-sequence $\{Z_s^t\}$ where $Z_s^t = Z^{1+(t-1)s}$.

The binary process $\{Z^t\}$ is not Markov in general. However for large enough $s \geq 1$, $\{Z_s^t\}$ will well approximate a Markov chain. The Raftery and Lewis procedure is based around this assumption, choosing s after observing a pilot sample of the data. In fact, they sequentially use BIC ratios to select s, though other ways of doing this are clearly possible.

Once s is established, the approximate transition matrix of $\{Z_s^t\}$ is approximated empirically by the obvious estimator $\widehat{\boldsymbol{P}}$, with i,jth element

$$\widehat{P}_{ij} = \frac{\#\{t : Z_s^t = j,\ Z_s^{t-1} = i\}}{\#\{t;\ Z_s^{t-1} = i\}} \qquad i,j = 0,1.$$

Since eigen-analysis of $\widehat{P}$ is straightforward, convergence diagnostic procedures both for the convergence in distribution of $\{Z_s^t\}$ and for convergence to a pre-specified accuracy of ergodic estimates are easily constructed.

The main strengths of this method lie in its ease of implementation and in its focus upon the accuracy of ergodic averages. However, the method relies upon three levels of approximation: the Markov approximation, the choice of s and the estimation of $\widehat{P_{ij}}$.

A further word of caution regarding this method is provided by Brooks and Roberts (1997) who further stress the fact that this method estimates the convergence rate of the chain only for the quantile of interest and does not provide any information as to the convergence rate of the chain as a whole. In fact they show that in the case of the Independence sampler the convergence rate estimate provided by Raftery and Lewis lies below the convergence rate of the full chain and that the choice of q is of critical importance to the estimate obtained. Thus, some care should be taken in selecting a suitable q value. In particular, the routine use of $q = 0.025$, suggested both in the literature and within the S-Plus and Fortran code that the original authors distribute, should not be adopted as a general rule, since it may lead to a strong underestimate of the true length of the burn-in period. When interested in a number of quantiles, it may be most sensible to run the procedure for a number of different u-values and take the largest of the resulting estimated burn- in lengths, but in the case where quantiles themselves are not of interest, this method should be used with caution.

5.3 Geometric Convergence Bounds

Because of the uncertainties associated with eigenvalue estimation outlined above, it is appealing to search for analytic inequalities of the form (17). However, for Markov chains which provably satisfy (17), it is rare for useful bounds on V and ρ to be attainable in even moderately complicated statistical models. The one technique which has made inroads into this problem is that of Rosenthal (1995b), which gives a flexible approach for computing ρ and V which can (in some cases at least) be applied to MCMC problems (see Rosenthal, 1995a,b,c).

This method is based upon the idea of running two chains independently, one of which is started from the stationary distribution π, and the other from some initial distribution π_0 and both with t-step transition kernel $\mathcal{P}^t(\cdot,\cdot)$. Let C be a small set, ie;

$$\pi(C) > 0 \quad \text{and} \quad \epsilon\nu(\cdot) \leq \mathcal{P}^t(\boldsymbol{x},\cdot) \quad \forall \boldsymbol{x} \in C,$$

and for some integer $t \geq 1$, constant $\epsilon > 0$ and probability measure ν on E. Then, each time that both chains are simultaneously in C, they have a given probability of coupling and Rosenthal (1995b) proves the following theorem.

Theorem 5.1 *Suppose we have two Markov chains, $\{\boldsymbol{X}^t\}$ and $\{\boldsymbol{Y}^t\}$ on state space E with transition kernel $\mathcal{P}$, which satisfy the minorisation condition*

$$\mathcal{P}^t(\boldsymbol{x},\cdot) \geq \epsilon\nu(\cdot) \quad \forall \boldsymbol{x} \in C$$

for some set C, positive integer t, $\epsilon > 0$ and probability measure $\nu(\cdot)$ on $\mathcal{E}$. Suppose also that there exists $\alpha > 1$ and a function $h : E \times E \to [1,\infty)$, such that

$$\mathbb{E}\left\{ h(\boldsymbol{X}^1, \boldsymbol{Y}^1) | \boldsymbol{X}^0 = \boldsymbol{x}, \boldsymbol{Y}^0 = \boldsymbol{y} \right\} \leq h(\boldsymbol{x},\boldsymbol{y})/\alpha \quad \forall (\boldsymbol{x},\boldsymbol{y}) \notin C \times C.$$

and that

$$A = \sup_{(\boldsymbol{x},\boldsymbol{y})\in C\times C} \mathbb{E}\left[h(\boldsymbol{X}^t,\boldsymbol{Y}^t)|\boldsymbol{X}^0=\boldsymbol{x},\boldsymbol{Y}^0=\boldsymbol{y}\right],$$

Then, if ν is the initial distribution of $\boldsymbol{X}^0$, π is the stationary distribution, and $j>0$,

$$|\mathcal{P}^k(\boldsymbol{x}^0,\cdot)-\pi(\cdot)| \le (1-\epsilon)^{[j/t]} + \alpha^{-k+jt-1}A^{j-1}\mathbb{E}_{\pi\times\nu}\left\{h(\boldsymbol{X}^0,\boldsymbol{Y}^0)\right\},$$

where $\mathbb{E}_{\pi\times\nu}\{h(\boldsymbol{X},\boldsymbol{Y})\}$ refers to the expectation with $\boldsymbol{X}$ having distribution ν and $\boldsymbol{Y}$ being distributed independently under π and $[i]$ denotes the smallest integer less than or equal to i.

In practice, this theorem is usually applied in terms of a drift condition on the original chain, rather than on the coupled chain, see Rosenthal (1995b, Theorem 12).

This method can require a substantial amount of analytical work in order to calculate A for example, or to estimate (bound above) the expectation $\mathbb{E}_{\pi\times\pi^0}\{h(\boldsymbol{X},\boldsymbol{Y})\}$. Often also, the bounds obtained are too poor to be of any practical value. However, this method has been applied to a number of statistical problems, see Rosenthal (1995a,b,c). . The approach remains the most promising avenue for analytic estimation of Markov chain running times.

An interesting idea to make these ideas more accessible and generally applicable is to use simulation techniques to calculate the constants, α, and ϵ, see Cowles and Rosenthal (1996). Although promising, this approach still remains rather computationally expensive, and the bounds obtained can still be very poor.

Similar, computable, bounds have also been derived by Meyn and Tweedie (1994). However, for complex target distributions, the approximations necessary in order for any of these bounds to be analytically tractable, lead to bounds too weak to be of much practical value.

It is possible to construct less general, but rather more accurate results applicable to certain classes of problems. only a small class of samplers. For example, Roberts and Sahu (1997) gives exact expressions for the rate of convergence of various types of Gibbs sampler on Gaussian target densities. For distributions close to Gaussian, corresponding approximation results can be obtained, see Roberts and Sahu (1996). Some related results appear in Amit (1991). Similarly, Polson (1996) discusses approaches to the estimation of the convergence rate in the case for log-concave target distributions on a lattice, using a particular form of the Metropolis algorithm. Finally, Liu (1994) provides a method for exact calculation of the rate of convergence for the independence sampler. If the importance weight, $w(\boldsymbol{x})$, has a finite maximum w^* for some $\boldsymbol{x}=\boldsymbol{x}^*\in E$, then the convergence rate of the chain is given by $\rho = 1-1/w^*$. Thus, the independence sampler converges geometrically to π, whenever $w^*<\infty$. This result is further discussed in Mengersen and Tweedie (1995) and extended in Smith and Tierney (1996), and is particularly easy to use. However, no similar results exist for more general samplers such as the Gibbs sampler or the Metropolis Hastings algorithm.

6 Discussion

6.1 Comparing the Methods

Tables 1 and 2 summarise the diagnostic methods discussed in this paper. ¿From these tables it is clear that there is a general trend to these methods in that, in order to utilise output from multiple replications and assess convergence of the joint density, the diagnostic necessarily becomes either less generally applicable, computationally more expensive to use, or less easily interpretable. In essence, we need to "pay" in order to get a better diagnostic. However, since there is an upper bound on the utility (or reliability) of any particular diagnostic, the most computationally expensive may not be practically viable under virtually any circumstances.

Each of the diagnostic methods has its own drawbacks and there is no one globally "best" diagnostic method which should be used for all problems. Of course, there is no reason why a number of techniques cannot be used simultaneously, particularly those for which generic code can be written and, in practice, this is what many people do.

In the case of unimodal, roughly symmetric stationary densities, the method of Gelman and Rubin (1992) is recommended, so long as it is possible to gain a rough estimate of the range of the distribution, so that suitably over-dispersed starting values can be chosen.

If we are concerned about multi-modality, then it is advisable to supplement the use of Gelman and Rubin's diagnostic with a complementary technique such as that of Johnson (1996), and/or a separate mode-hunting algorithm so that suitable starting values may be obtained.

For a more rigorous test, in small dimensional problems, the L^2 diagnostic of Roberts (1994) should work well, but it has the drawback that it can be difficult to implement and is problem specific. The L^2 diagnostic cannot generally be recommended in problems of greater than 15 dimensions.

For problems with discrete (or partly discrete) state spaces, the easiest, and most non-parametric method, is that of Johnson (1996) and, where regenerations are possible, Robert's regeneration-based method may also be useful. Also, if only quantiles are required, then the method of Raftery and Lewis (1992) is quick and very easy to use, though the results should be interpreted with care.

Most of the diagnostic methods that we have discussed have looked at convergence of the chain itself, and do not try to provide a model for convergence. Notable exceptions to this are the methods of Raftery and Lewis (1992) and Garren and Smith (1995). The method of Garren and Smith (1995) might be best used when we are interested in the chain itself, rather than the stationary distribution, otherwise the method seems to be addressing the problem rather indirectly.

Finally, a good compromise for the marginal versus full-dimensional trade-off of rigour versus interpretability is provided by the L^1 diagnostic of Brooks *et al* (1997), which makes use of multiple replications, is reliable and interpretable, but still reasonably computationally inexpensive.

Thus, there are now a large number of diagnostic methods available for diagnosing convergence of an MCMC algorithm to stationarity. Each has its own strengths and weaknesses. The more rigorous and reliable methods, tend to be the most compu-

Diagnostic Procedure	**Full Joint Density?**	**Multiple Chains?**	**Applicable Samplers?**	**Computational Expense**	**Interpretability**	**Comments and Restrictions**
Raftery and Lewis	No	No	Any	Cheap	Easy	Dependent upon arbitrary constant values
Garren and Smith	Yes	Yes**	Any, under certain conditions	Expensive	Subjective, requires experience	May be difficult to prove compliance with conditions for some samplers
Gelman and Rubin	No	Yes	Any	Cheap	Subjective	Widely used in practice
Geweke	No*	No*	Any	Cheap	Easy	Dependent upon arbitrary Spectral Window width
Heidelberger and Welch	No*	No*	Any	Fairly Cheap	Easy	Low power of test and Spectral Window problem
Yu and Mykland	No	No	Any	Cheap	Subjective, but quantifiable	D-statistic is simple and effective

Table 1: Summary of entirely output-based diagnostic methods. $*$ - Can be generalised. $**$ - Identical starting points

Diagnostic Procedure	**Full Joint Density?**	**Multiple Chains?**	**Applicable Samplers?**	**Computational Expense**	**Interpretability**	**Comments and Restrictions**
Ritter and Tanner	Yes	Yes	GS, MH and hybrids	Can be expensive	Subjective	Unreliable unless parameter space can be suitably partitioned
Zellner and Min	Yes	No	GS only	Could be expensive	Easy	Limited to problems where space can be partitioned
Liu, Liu and Rubin	Yes	Yes	GS only	Can be expensive	Subjective	Assumes normality of the global control variable
Roberts	Yes	Yes	Reversible GS	Can be expensive	Subjective, requires experience	Difficult to check Hilbert Schmidt condition
Brooks, Dellaportas and Roberts	Yes	Yes	Any	Can be expensive	Easy	Choice of n_0 value may be difficult.
Yu	Yes	No	Any	Expensive	Subjective	Easily "tricked" by multimodal densities Converges sample-wise
Johnson	Yes	Yes	Coupled GS only	Cheap if conditionals are standard	Subjective choice of ϵ, but generally easy	Dependent upon number of replications for certain classes of problems
Propp and Wilson	No	Yes	Any	Can be expensive	Easy	Finite state space restriction, but good where ordering is possible. Impractical.
Mykland, Tierney and Yu	Yes	No	Regenerative	Cheap	Subjective	Limited applicability, and useful only for discrete state spaces
Robert	Yes	No	Regenerative	Cheap	Easy	More rigorous than Mykland *et al*, but limited to discrete state spaces

Table 2: Summary of problem-specific diagnostic methods.

tationally expensive, and so the least well used, whilst those for which generic code is easily available are by far the most popular. Whilst the existence of these codes are extremely useful, they are open to naive misuse and mis-interpretation. It is essential to have some understanding of the basic concepts behind any diagnostic that you choose to use, and to understand its limitations in terms of the types of problems for which it may be applied. If in doubt, it is worth taking the time to use more than one diagnostic on a particular problem and this should avoid being mis-led by an inappropriately applied method.

7 Availability of Computer Code

Computer code for a few of these diagnostics are now publicly available. Fortran code to implement Raftery and Lewis' method on page 24, is available via Statlib. Send the one-line message "send gibbsit from general" to statlib@temper.stat.cmu.edu. S-Plus code for Gelman and Rubin's method on page 6, may also be obtained via statlib by sending the message "send itsim from S" to the same address. Fortran code to implement Geweke's method on page 9, may be obtained by sending an e-mail message to the author, geweke@atlas.socsci.umn.edu. Finally a suite of S-Plus routines written implement diagnostic routines on output from the BUGS package (Bayesian analysis Using the Gibbs sampler, see Gilks *et al* (1992)) is also publicly available, and distributed with the package.

References

Ahmad, R. (1976), On the Multivariate K-Sample Problem and the Generalization of the Kolmogorov-Smirnov Test. *Annals of the Institute of Statistical Mathematics* **28**, 259–265

Amit, Y. (1991), On Rates of Convergence of Stochastic Relaxation for Gaussian and Non-Gaussian Distributions. *Journal of Multivariate Analysis* **38**, 82–99

Amit, Y. and U. Grenander (1991), Comparing Sweep Strategies for Stochastic Relaxation. *Journal of Multivariate Analysis* **37**, 197–222

Applegate, D., R. Kannan and N. G. Polson (1990), Random Polynomial Time Algorithms for Sampling from Joint Distributions. Technical report, Carnegie-Mellon University, Tech. Rep. No 500

Asmussen, S., P. W. Glynn and H. Thorisson (1992), Stationarity Detection in the Initial Transient Problem. *ACM Transactions on Modelling and Computer Simulation* **2**, 130–157

Brittain, E. H. (1987), P-Values for the Multi-Sample Kolmogorov-Smirnov Test using the Expanded Bonferroni Approximation. *Communications in Statistics - Theory and Methods* **16**(3), 821–835

Brooks, S. P. (1996), Quantitative Convergence Diagnosis for MCMC via CUSUMS. Technical report, University of Bristol.

Brooks, S. P. (1997), MCMC Convergence Diagnosis via Multivariate Bounds on Log-Concave Densities. Technical report, University of Bristol.

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* To Appear.

Brooks, S. P. and A. Gelman (1997), Alternative Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics* To Appear.

Brooks, S. P. and G. O. Roberts (1997), On Quantile Estimation and MCMC Convergence. Technical report, University of Bristol

Chan, K. S. (1993), Asymptotic Behaviour of the Gibbs Sampler. *Journal of the American Statistical Association* **88**, 320–328

Chorneyko, I. Z. and L. L. K. Zing (1982), K-Sample Analogues of Two-Sample Kolmogorov-Smirnov Statistics. *Selecta Statistica Canadiana* **6**, 37–62

Conover, W. J. (1965), Several K-Sample Kolmogorov-Smirnov Tests. *Annals of Mathematical Statistics* **36**, 1019–1026

Conover, W. J. (1967), A k-Sample Extension of the One-Sided Two-Sample Smirnov Test Statistic. *Annals of Mathematical Statistics* **38**, 1726–1730

Cowles, M. K. and J. S. Rosenthal (1996), A Simulation Approach to Convergence Rates for Markov Chain Monte Carlo. Technical report, Harvard School of Public Health

Diaconis, P. (1988), *Group Representations in Probability and Statistics*, vol. 11 of *Lecture Notes - Monograph Series*. Institute of Mathematical Statistics

Diaconis, P. and D. Stroock (1991), Geometric Bounds for Eigenvalues of Markov Chains. *Annals of Applied Probability* **1**, 36–61

Fasano, G. and A. Franceschini (1987), A Multidimensional Version of the Kolmogorov-Smirnov Test. *Monthly Notices of the Royal Astronomical Society* **225**, 155–170

Garren, S. and R. L. Smith (1995), Estimating the Second Largest Eigenvalue of a Markov Transition Matrix. Technical report, University of Cambridge

Gelfand, A. E. (1992), Discussion of Gelman and Rubin (1992). *Statistical Science* **7**, 486–487

Gelfand, A. E., S. E. Hills, A. Racine-Poon and A. F. M. Smith (1990), Illustration of Bayesian Inference in Normal data Models using Gibbs Sampling. *Journal of the American Statistical Association* **85**, 972–985

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. and D. Rubin (1992), Inference from Iterative Simulation using Multiple Sequences. *Statistical Science* **7**, 457–511

Geman, S. and D. Geman (1984), Stochastic Relaxation , Gibbs Distributions and the Bayesian Restoration of Images. *IEEE Transactions on pattern analysis and machine intelligence* **6**, 721–741

Geweke, J. (1989), Bayesian Inference in Econometric Models using Monte-Carlo Integration. *Econometrica* **57**, 1317–1339

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. (1992), Practical Markov Chain Monte Carlo. *Statistical Science* **7**, 473–511

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

Gilks, W. R., G. O. Roberts and S. K. Sahu (1996), Adaptive Markov Chain Monte Carlo. Technical report, MRC Biostatistics Unit, Cambridge

Gilks, W. R., D. Thomas and D. J. Spiegelhalter (1992), Software for the Gibbs Sampler. *Computing Science and Statistics* **24**, 439–448

Heidelberger, P. and P. D. Welch (1981a), Adaptive Spectral Methods for Simulation Output Analysis. *IBM Journal of Research and Development* **25**, 860–876

Heidelberger, P. and P. D. Welch (1981b), A Spectral Method for Confidence Interval Generation and Run Length Control in Simulations. *Communications of the Association Computing Machinery* **24**, 233–245

Heidelberger, P. and P. D. Welch (1983), Simulation Run Length Control in the Presence of an Initial Transient. *Operations Research* **31**, 1109–1144

Jennison, C. (1993), Discussion to Smith and Roberts (1993). *Journal of the Royal Statistical Society, Series B* **55**, 54–56

Johnson, V. E. (1996), Studying Convergence of Markov chain Monte Carlo Algorithms using Coupled Sample Paths. *Journal of the American Statistical Association* **91**, 154–166

Kendall, W. S. (1997), Perfect simulation for the area interaction process. In C. C. Heyde and L. Acardi (eds.), *Probability Perspective*, World Scientific Press

Kolmogorov, A. N. (1933), Sulla Determiazione Empirica delle Leggi di Probabilita. *Giornal Istitutto Italia Attuari* **4**, 1–11

Lawler, G. F. and A. D. Sokal (1988), Bounds on the L^2 spectrum for Markov Chains and their Applications. *Transactions of the American Mathematical Society* **309**, 557–580

Lin, Z. Y. (1992), On the Increments of Partial Sums of a ϕ-mixing Sequence. *Theory of Probability and its Applications* **36**, 316–326

Lindvall, T. (1992), *Lectures on the Coupling Method*. Wiley

Liu, C., J. Liu and D. B. Rubin (1993), A Control Variable for Assessment the Convergence of the Gibbs Sampler. In *Proceedings of the Statistical Computing Section of the American Statistical Association*, pp. 74–78

Liu, J. S. (1994), Metropolized Independent Sampling with Comparisons to Rejection Sampling and Importance Sampling. Technical report, Harvard University

Müller, P. (1991), A Generic Approach to Posterior Integration and Gibbs Sampling. Technical report, Dept. of Statistics, Purdue University

Mengersen, K. L. and R. L. Tweedie (1995), Rates of Convergence of the Hastings and Metropolis Algorithms. *Annals of Statistics* **24**, 101–121

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

Meyn, S. P. and R. L. Tweedie (1994), Computable Bounds for geometric Convergence Rates of Markov Chains. *Annals of Applied Probability* **4**, 981–1011

Mises von, R. (1931), *Wahrscheinlichkeitschrechnung*. Deutiche, Vienna

Møller, J. (1997), Markov chain Monte Carlo and spatial point processes. In W. S. Kendall and M. N. M. vanLieshout (eds.), *Stochastic geometry, likelihood and computation*, Chapman and Hall

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

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

Peskun, P. H. (1973), Optimum Monte Carlo Sampling using Markov Chains. *Biometrika* **60**, 607–612

Polson, N. G. (1996), Convergence of Markov Chain Monte Carlo Algorithms. In J. M. Bernardo, A. F. M. Smith, A. P. Dawid and J. O. Berger (eds.), *Bayesian Statistics 5*, Oxford University Press

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*, Oxford University Press

Reutter, A. and V. Johnson (1995), General Strategies for Assessing Convergence of MCMC Algorithms Using Coupled Sample Paths. Technical report, Duke University

Ripley, B. D. (1987), *Stochastic Simulation*. John Wiley and Sons

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

Robert, C. P. (1996), Convergence Assessments for Markov Chain Monte Carlo Methods. *Statistical Science* **10**, 231–253

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*, Oxford University Press

Roberts, G. O. (1994), Methods for Estimating L^2 Convergence of Markov Chain Monte Carlo. In D. Berry, K. Chaloner and J. Geweke (eds.), *Bayesian Statistics and Econometrics: Essays in Honor of Arnold Zellner*, North Holland: Amsterdam

Roberts, G. O. and N. G. Polson (1994), On the Geometric Convergence of the Gibbs Sampler. *Journal of the Royal Statistical Society, Series B* **56**, 377–384

Roberts, G. O. and S. K. Sahu (1996), Rate of Convergence of the Gibbs Sampler by Gaussian Approximation. Technical report, University of Cambridge

Roberts, G. O. and S. K. Sahu (1997), Updating Schemes, Covariance Structure, Blocking and Parametrisation for the Gibbs Sampler. *Journal of the Royal Statistical Society, Series B* **59**, 291–318

Roberts, G. O. and A. F. M. Smith (1994), Simple Conditions for the Convergence of the Gibbs Sampler and Metropolis Hastings Algorithms. *Stochastic Processes and Applications* **49**, 207–216

Roberts, G. O. and R. L. Tweedie (1996), Geometric Convergence and Central Limit Theorems, for Multidimensional Hastings and Metropolis Algorithms. *Biometrika* **83**, 95–110

Rosenthal, J. (1995a), Rates of Convergence for Gibbs Sampling for Variance Component Models. *Annals of Statistics* **23**, 740–761

Rosenthal, J. S. (1995b), Minorization Conditions and Convergence Rates for Markov Chain Monte Carlo. *Journal of the American Statistical Association* **90**, 558–566

Rosenthal, J. S. (1995c), Rates of Convergence for Data Augmentation on Finite Sample Spaces. *Annals of Applied Probability* **3**, 319–339

Schervish, M. J. and B. P. Carlin (1992), On the Convergence of Successive Substitution Sampling. *Journal of Computational and Graphical statistics* **1**, 111–127

Schruben, L., H. Singh and L. Tierney (1983), Optimal Tests for Initialization Bias in Simulation Output. *Operations Research* **31**, 1167–1178

Schruben, L. W. (1982), Detecting Initialization Bias in Simulation Output. *Operations Research* **30**, 569–590

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall

Sinclair, A. J. and M. R. Jerrum (1988), Conductance and the Rapid Mixing Property for Markov Chains: The Approximation of the Permanent Resolved. In *Proceedings of the 20th annual ACM symposium on the Theory of Computing*

Smith, A. F. M. and A. E. Gelfand (1992), Bayesian Statistics Without Tears: A Sampling-Resampling Perspective. *The American Statistician* **46**, 84–88

Smith, R. L. and L. Tierney (1996), Exact Transition Probabilities for the Independence Metropolis Sampler. Technical report, Statistical Laboratory, Cambridge

Tierney, L. (1994), Markov Chains for Exploring Posterior Distributions. *Annals of Statistics* **22**, 1701–1762

Yu, B. (1995a), Discussion to Besag *et al* (1995). *Statistical Science* **10**, 3–66

Yu, B. (1995b), Estimating L^1 Error of Kernel Estimator: Monitoring Convergence of Markov Samplers. Technical report, Dept. of Statistics, University of California, Berkeley

Yu, B. and P. Mykland (1997), Looking at Markov Sampler through Cusum Path Plots: A Simple Diagnostic Idea. *Statistics and Computing* To Appear.

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