COVER SHEET

PHASE RANDOMISATION : A CONVERGENCE DIAGNOSTIC TEST FOR MCMC

Darfiana Nur †, Kerrie L. Mengersen ‡ and Rodney C. Wolff ‡

University of Newcastle †, *Queensland University of Technology* ‡

Summary

Most MCMC users address the convergence problem by applying diagnostic tools to the output produced by running their samplers. Potentially useful diagnostics may be borrowed from diverse areas such as time series. One such method is phase randomisation. The aim of this paper is to describe this method in the context of MCMC, summarise its characteristics, and contrast its performance with those of the more common diagnostic tests for MCMC. It is observed that the new tool contributes information about third and higher order cumulant behaviour which is important in characterising certain forms of nonlinearity and nonstationarity.

Keywords: convergence diagnostics; higher cumulants; Markov Chain Monte Carlo; non-linear time series; stationarity ; surrogate series.

1 Introduction

Markov Chain Monte Carlo (MCMC) methods have revolutionised Bayesian statistics, enabling evaluation of complex distributions and thus facilitating careful modelling in a very wide range of disciplines. An important consideration in the implementation of these methods, however, is whether the chain converges in some formal sense to the target distribution and, if so, how quickly. Detailed convergence properties have been established through Markov chain theory in specific or general setups, with often stringent conditions on the sampler which are difficult to verify in practice (Mengersen & Tweedie (1996), Robert, Ryden & Titterington (1998)). On a more practical basis, convergence may also be assessed through a suite of diagnostics borrowed from diverse areas such as time series, exploratory data analysis, coupling theory and other probabilistic tools.

Cowles & Carlin (1996) presented a comprehensive review of thirteen convergence diagnostics for MCMC. They compared their performance in two simple models and concluded that all the methods can fail to detect the sorts of convergence they were designed to identify. Mengersen & Robert (1999) reviewed

MCMC convergence diagnostics based on three categories : exploration (or 'burn-in', time to reach the stationarity distribution in an average sense); stationarity (adequate movement around the target distribution) and estimation (adequate length of sampling for estimation, appealing to Central Limit Theorem properties). Those diagnostic tests that are comparable with the method proposed in this paper include the *diag* command from Bayesian Inference Using Gibbs Sampling (BUGS) (Spiegelhalter *et al.* (1994)), the tests proposed by Geweke (1992), Raftery & Lewis (1992) and Heidelberger & Welch (1983), and the usual autocorrelation test. The last four tests can be found in the Convergence Output and Diagnostics Analysis (CODA) software (Best, Cowles & Vines, 1995).

The convergence diagnostics that are comparable with the method proposed in this paper are briefly described here, in order to facilitate later discussion. A spectral analysis approach is adopted by Heidelberger & Welch (1983) who use the Crámer-von-Mises statistic to test the null hypothesis that the sampled values form a stationary process. In contrast, the method of Raftery and Lewis (1996) applies to a single chain to detect convergence to the stationary distribution. They reduce the chain to a two-state Markov chain which can be analysed explicitly, although there is some discussion about the appropriateness of this approach (see Mengersen & Robert). Geweke (1992) proposes a different convergence diagnostic based on comparing spectral density estimates of the first and last parts of the MCMC chain.

There remains a concern, however, that although the MCMC may satisfy first order stationarity, it may exhibit nonstationary behaviour for higher moments. This might be excarbated if the model is complex in terms of hierarchical structure or non-standard distributions, or the required expectations are based on these higher order moments. Alternatively, such behaviour might indicate insufficient "burn-in" of the MCMC algorithm or poor mixing of the MCMC chain. Detection of nonlinearity and/or nonstationarity in the higher cumulants is important for the estimation phase, which assumes that the CLT is satisfied.

In this paper we consider the output of a MCMC simulation as a time series for which nonlinear and dynamical properties may be investigated using classical and modern analytical methods. A recently developed device used in analysis of dynamical systems output is that of *phase randomisation,* which consists of taking the Fourier transform of a given series, replacing the phase with a value sampled uniformly on $(0, 2\pi)$ and back-transforming to render a so-called *surrogate series.* This procedure, variously known as a method of surrogate data (Theiler *et al.* (1992)), phase scrambling (Davison & Hinkley (1997)), or Fourier bootstrap (Braun

& Kulperger (1997)), is commonly used in non-linear time series analysis (Theiler *et al.* (1992), Timmer (1998)).

Since the randomisation technique in phase randomisation leaves amplitudes unaltered, second order structure is preserved in the surrogate series. Due to the requirement of symmetry amongst the array of angles used in the Fast Fourier Transform, it is important to determine which other features of the original distribution, if any, are preserved under phase randomisation. It is shown in Nur, Wolff & Mengersen (2001) that the higher cumulant estimates of the surrogate series can be used to detect the stationarity of a time series graphically. If the smoothing densities of higher cumulant estimates of surrogates are unimodal around zero then stationarity can be concluded. In the present paper we take this a step further and borrowing from Robert & Casella (1999), we evaluate the effectiveness of traditional tests of normality, namely Kolmogorov-Smirnov and Shapiro & Wilk, in supporting the assertion of convergence. The asymptotic normality of the third cumulant of the surrogate series has been established by Nur (2003) under certain conditions using Edgeworth expansion. Using this theoretical result, which is further described in Section 4, the third cumulant estimates might be constructed as a statistical test to assess the convergence of MCMC.

The organisation of the paper is as follows. Section 2 presents a summary of phase randomisation in time series and characteristics of the resultant bootstrap chains in a time series context. Its application in MCMC is then proposed. In Section 3, we apply the phase randomisation method and other convergence diagnostic tests to well-known MCMC examples, namely the surgical example described in BUGS (Spiegelhalter *et al.* (1994)) and another example using the Metropolis-Hastings algorithm. Formal tests for convergence are presented in Section 4 and finally, conclusions are drawn in Section 5.

2 Phase Randomisation

2.1 What is phase randomisation?

Suppose $\{X_t\}$ is a time series and $\boldsymbol{X}= (X_1, X_2, ..., X_N)$ is a data set. Let

$$\mathrm{E}(X_t) = \mu, \quad \gamma_k = \mathrm{E}\left((X_t - \mu)(X_{t+k} - \mu)\right)$$

be the expectation and the autocovariances of $\{X_t\}$ respectively. The method of surrogate data generates fictitious data $\boldsymbol{Y}= (Y_1, ..., Y_N)$, such that for $N = 2m+1$

$$Y_t = \bar{X} + \sqrt{\frac{2\pi}{N}} \sum_{j=1}^{m} 2\sqrt{I(\boldsymbol{X}, w_j)} \cos(w_t j + \theta_j), \ (t = 1, ..., N), \tag{1}$$

where I denotes the modified periodogram, w_j are the angular frequencies such that $w_j = 2\pi j/N$ and $\theta_1, ..., \theta_m$ are iid $U(0, 2\pi)$. If $N = 2(m + 1)$, then

$$Y_t = \bar{X} + \sqrt{\frac{2\pi}{N}} \sum_{j=1}^{m} 2\sqrt{I(\boldsymbol{X}, w_j)} \cos(w_t j + \theta_j) + \sqrt{\frac{2\pi}{N} I(\boldsymbol{X}, w_{m+1})} \cos(\pi t + \theta_{m+1}),$$

where $\theta_1, ..., \theta_m$ are independent of θ_{m+1}, which is equal to 0 or π with probability 0.5 each.

By construction, the surrogate data Y preserve the observed sample mean and periodogram, that is

$$\bar{Y} = \bar{X}, \quad I(\boldsymbol{Y}, w_j) = I(\boldsymbol{X}, w_j), \ \ (j = 1, ..., N) \text{ with probability 1.}$$

As a consequence, $\boldsymbol{Y}$ preserves the sample circular auto-covariances, that is

$$\frac{1}{N} \sum_{t=1}^{N} (Y_t - \bar{Y})(Y_{t+k} - \bar{Y}) = \frac{1}{N} \sum_{t=1}^{N} (X_t - \bar{X})(X_{t+k} - \bar{X})$$

where $Y_{t+N} = Y_t$ and $X_{t+N} = X_t$ for all $t \geq 0$.

Let $U = U(\boldsymbol{X})$ denote the sample mean $\bar{X}$ and the periodogram values $I(\boldsymbol{X}, w_j), j = 1, ..., N$. Assume that $U = u$ is fixed and $\boldsymbol{Y}$ is generated according to (1). Assume that $N = 2m + 1$ is odd and let $p \geq q$ be two positive integers. Then

$$\mathrm{E}(Y_t) = \bar{X},$$

$$\mathrm{cov}(Y_p, Y_q) = \mathrm{E}\left((Y_p - \bar{X})(Y_q - \bar{X})\right) - \bar{X}^2,$$

since $\mathrm{E}\left(cos(w_t j + \theta_j)\right) = 0$ and $\mathrm{var}\left(cos(w_t j + \theta_j)\right) = \frac{1}{2}$.

Continuing for higher cumulants, we have the properties that $C(Y_p, Y_q, Y_r) = \mathrm{E}\left((Y_p - \bar{X})(Y_q - \bar{X})(Y_r - \bar{X})\right) = 0$; odd cumulants of the surrogate series are zero and even cumulants of the surrogate series are non-zero. This implies that the surrogate series have a highly symmetric joint distribution.

The algorithm for generating the bootstrap surrogates is given below (Theiler et al., 1992).

Rescaling Surrogate algorithm

(i) Input the original data into an array $x(t),\ t = 1, 2, ..., N$.

(ii) Let $y_t = \Phi^{-1}\left(r_t/(n+1)\right)$ where r_t is the rank of x_t among the original series $x_1, ..., x_N$ and Φ denotes the standard normal cumulative distribution function.

(iii) Compute the Discrete Fourier Transform: $z(t) = \mathrm{DFT}(y(t))$. Note that $z(t)$ has real and imaginary components.

(iv) Randomize the phases: $z'(t) = z(t)\exp\left(i\phi(t)\right)$, where $\phi(t)$ is uniformly distributed between 0 and 2π.

(v) Symmetrize the phases such that

$$\mathrm{Re}\left(z''(t)\right) = \frac{1}{2}\mathrm{Re}\left(z'(t) + z'(N+1-t)\right)$$

$$\mathrm{Im}\left(z''(t)\right) = \frac{1}{2}\mathrm{Im}\left(z'(t) - z'(N+1-t)\right)$$

where Re and Im are the real and imaginary parts of a complex number respectively.

(vi) Invert the DFT: $y'(t) = \mathrm{DFT}^{-1}(z''(t))$.

(vii) Set the surrogate series $X_t^* = x_{(r_t')}$, where r_t' is the rank of y_t' among $y_1', ..., y_N'$.

2.2 Characteristics of phase randomisation in time series

For the purpose of testing for stationarity in a time series, the above phase randomisation procedure was reviewed and modified, and applied to a wide range of time series models by Nur *et al* (2001). The models include linear stationary, linear non-stationary, non-linear stationary and non-linear non-stationary processes for which the characterization of stationary or non-stationary is known theoretically from the ergodicity condition. They computed the estimates of higher cumulants of original and surrogate series respectively. One thousand surrogate series were constructed for each time series. Estimates of the functions in Table 1 were calculated for each original and surrogate series.

TABLE 1

Higher cumulants functions and estimates at lags $k = 1, ..., 20$

Mathematical forms	Estimates
$\mathrm{E}((X_t - \mu)^r), \quad r = 2, ..., 7$	$\frac{1}{N}\sum_{t=1}^{N}(X_t - \bar{X})^r$
$\mathrm{E}\Big(\prod_{j=1}^{r}(X_{t+k+j} - \mu)\Big), \ r = 2, ..., 6$	$\frac{1}{N-k-r}\sum_{t=1}^{N-k-r}\Big(\prod_{j=1}^{r}(X_{t+k+j} - \bar{X})\Big)$
$\mathrm{E}((X_t - \mu)^r(X_{t+k} - \mu)^r), \quad r = 1, 2.$	$\frac{1}{N-k-r}\sum_{t=1}^{N-k-r}\Big((X_t - \bar{X})^r(X_{t+k} - \bar{X})^r\Big)$

The empirical density estimates of the higher cumulants estimates of the surrogates were then visually assessed and stationarity was characterised as follows. For each higher cumulant in Table 1:

(a) if the empirical densities of the cumulant estimates from the surrogates are unimodal around zero with a small variance then the original process is linear stationary or weakly nonlinear stationary,

(b) if the empirical densities of the cumulant estimates from the surrogates are unimodal around zero with a large variance then the original process is strongly nonlinear stationary,

(c) if the empirical densities of the cumulant estimates are multimodal or unimodal nonzero with a long tail then the original process is nonstationary.

2.3 Phase Randomisation in MCMC

We aim now to make use of phase randomisation applied to time series as discussed above by treating an MCMC sequence as a time series and creating a large number of replicate chains from the one MCMC chain in order to evaluate linearity and stationarity properties of the original chain. Using the results above, we expect that under stationarity, the behaviour of higher cumulant estimates of the surrogates of the MCMC will be asymptotically normal as shown in Table 2.

TABLE 2

Conclusion on the behaviour of higher cumulants for stationary and non-stationary processes using the Rescaling method

Process	Original	Surrogate
Stationary	Higher moments are non-zero	Higher moments are non-zero
	Higher cumulants around zero or small	Higher cumulants are around zero
	Cross cumulants are small	Cross cumulants are around zero (odd) or quite small (even)
Nonstationary	Higher moments are non-zero	Higher moments are non-zero
	Higher cumulants are very large	Higher cumulants are quite large
	Cross cumulants are large	Cross cumulants are quite large

Following the empirical results in Nur *et al* (2001) and theoretical results in Nur (2003), we choose the third cumulant estimates as the statistic for MCMC convergence. We summarise the steps for assessing the convergence of MCMC as follows.

(i) Generate an original MCMC chain according to the chosen sampler (Gibbs, Metropolis-Hastings or others).

(ii) Generate the surrogates of the original chain by applying the Rescaling algorithm described in Section 2.1.

(iii) Plot the histogram or empirical densities of the third cumulant estimates of surrogates in step (ii).

(iv) Use the result in 2.2 or Table 2 to visually assess convergence based on these densities. For example, if the histogram is unimodal around zero either with small or quite large variance, then the chain converges. The small variance reflects the linearity and stationarity of the original chain whereas the large variance reflects the nonlinearity and stationarity of original chain. If the histogram is unimodal nonzero with a long tail or multimodal then we can conclude that the chain has not converged yet.

(v) Apply a Kolmogorov-Smirnov or Shapiro & Wilk test for normality in supporting the assertion of convergence.

In the following section, we apply the above steps to two examples and compare the inferences about convergence that can be made using phase randomisation with those that are made using existing comparable MCMC diagnostics.

3 Applications

Two examples of MCMC chains arising from Bayesian analysis are presented in this section. The first one are generated by a Gibbs algorithm which samples directly from the posterior conditional distributions themselves and is thus expected to converge quickly. In fact, the example exhibits a very stable process. The second series is generated by a Metropolis-Hastings algorithm which samples from a proposal distribution and then accepts or rejects the proposed variable with a defined probability. Although theoretically it is known to converge uniformly or geometrically in special cases (Mengersen & Tweedie, 1996), it may be more awkward to visually assess convergence of the chain in practice. See Besag *et al* (1995) for details of these algorithms.

For comparison in the two examples, the diagnostics of Geweke (1992), Raftery & Lewis (1996), Heidelberger & Welch (1983), and simple autocorrelation were calculated using CODA. Default values were used for Geweke, Heidelberger & Welch and autocorrelation, whereas for Raftery & Lewis the precision and probability were set to 0.02 and 0.90 respectively to reflect the number of observations.

3.1 Surgical example

This example considers mortality rates in 12 hospitals performing cardiac surgery in babies. The data are displayed in Spiegelhalter *et al.* (1994). It is of interest to estimate the true underlying mortality rate in each hospital. The random effects model proposed by Spiegelhalter *et al.* expresses the number of deaths, r_i, for hospital i with true failure probability p_i, as with priors $r_i \stackrel{d}{=} \text{Bi}(p_i, n_i)$, $(i = 1, ..., 12)$; $\log(p_i) = b_i$, where $b_i \stackrel{d}{=} \text{N}(\mu, 1/\tau^2)$; $\mu \stackrel{d}{=} \text{N}(0, 10^{-6})$; $\tau \stackrel{d}{=} \text{Ga}(0.001, 0.001)$.

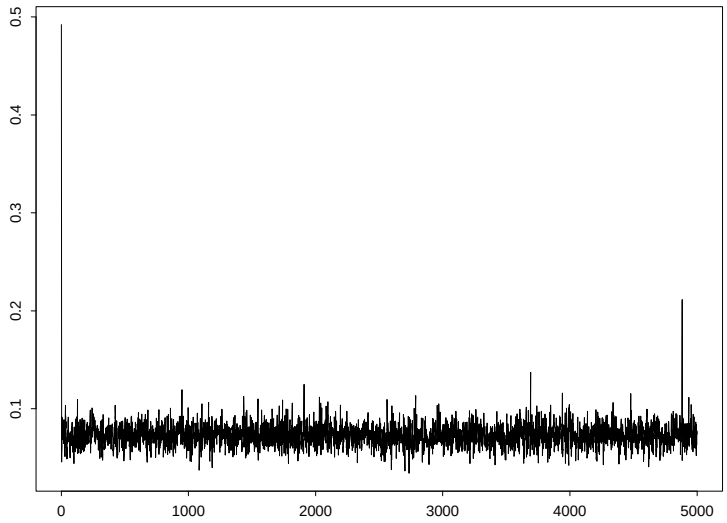

Figure 1. Surgical example: traceplot of parameter μ, 5000 iterations

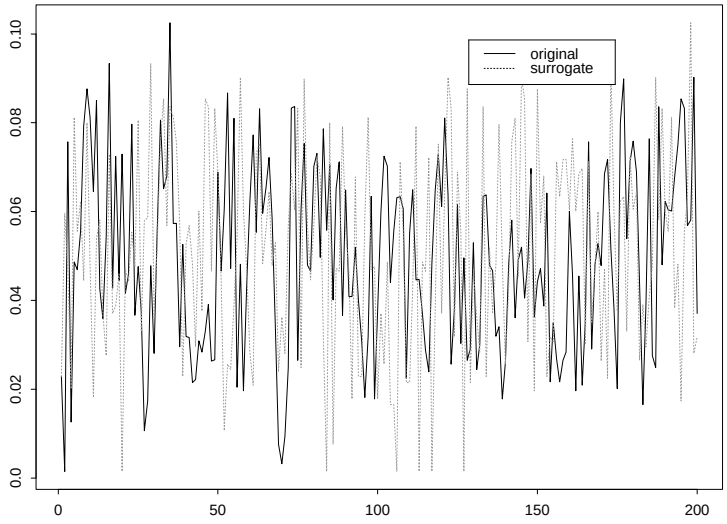

Figure 2. Surgical example: traceplot of the first 200 iterations of the chain in Figure 1

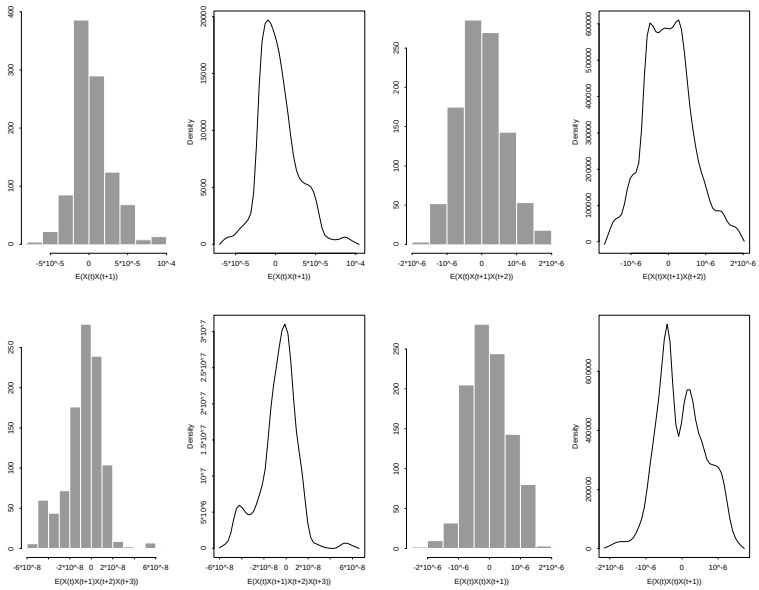

Figure 3. Surgical example: Histogram and smoothing density for the second, third, fourth and cross cumulants of the distribution represented in Figure 2, obtained by phase randomisation

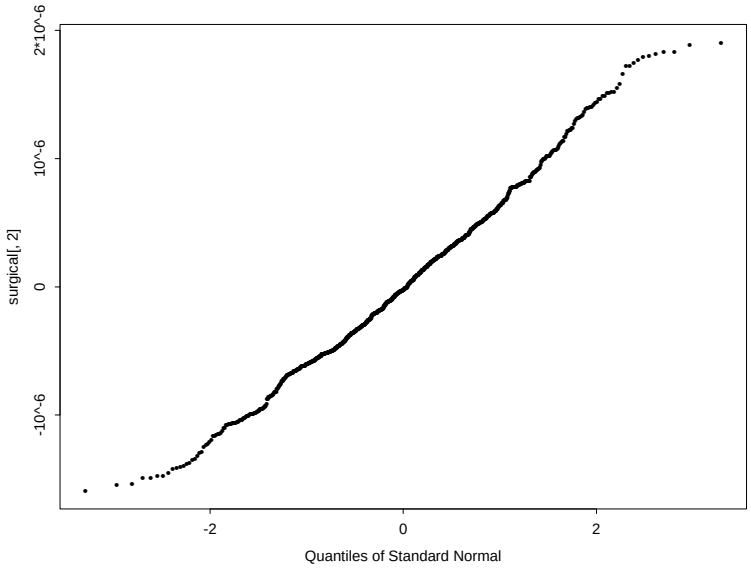

Figure 4. Surgical example: Normal QQ-plot for the third cumulant in Figure 3

For our purposes we focus on the overall mean mortality rate μ. Given the dataset, the form of the model and the Gibbs algorithm, we expect this MCMC chain to be very stable and indeed this is confirmed by the traceplot in Figure 1. Hence we consider whether we can assert convergence based on a very short run of, say, 200 cycles.

Figure 2 exhibits a time plot of the first 200 cycles. By comparison with the processes considered in Nur *et al* (2001), the MCMC run in Figure 2 appears to be similar to AR stationary or weakly bilinear stationary. The smoothing densities of the first four cumulants estimates in Figure 3 tend to be unimodal around zero with small variance. Normality assesment of the third cumulant estimates is presented visually in Figure 4 by the normal QQ-plot.

Our conclusion is that the MCMC series in Figure 2 is stationary. This is in accord with the *diag* assessments from BUGS, Raftery & Lewis, and Heidelberger & Welch, as well as the autocorrelation tests. However, Geweke's test is failed in this example, possibly because of not enough observations in the first 200 cycles. In fact, if Geweke's test is run on the first 1000 cycles, it is passed.

3.2 Metropolis-Hastings algorithm

In Robert & Casella (1999), a collection of Metropolis-Hastings algorithms is reviewed together with some examples. Two of the algorithms are the Random-Walk and Independent Metropolis-Hastings (Robert & Casella, page 245). The Random-Walk algorithm does not enjoy uniform ergodicity properties although it is possible to have a geometric ergodicity (Mengersen & Tweedie, 1996). The Independent algorithm converges if the tails of the proposal distribution are heavier than those of the target distribution. The form of the target distribution itself is also important, as discussed below.

The example considered here is taken from Mengersen & Tweedie (1996) in order to evaluate the performance of phase randomisation in assessing convergence where the theoretical convergence properties are known. Consider Random-Walk and Independent algorithms for the generation of two functions:

(a) $f \propto e^{-\frac{1}{2}x^2}$ (standard normal);

(b) $f \propto (1+ \mid x \mid)^{-3}$

For the purposes of illustration, proposal distributions of g the $\mathrm{N}(x,1)$ pdf with initial $x = 3$ and g the $\mathrm{N}(1,1)$ pdf were adopted for the Random-Walk

and Independent algorithms respectively. Applying the theorem of Mengersen & Tweedie (1996), it can be shown that the Random-Walk algorithm for case (a) is geometrically ergodic to the average, whereas for case (b) it is not. For the Independent algorithm, neither (a) nor (b) are expected to converge to its stationary distribution as the tail of proposal distribution is not heavier than the target distribution.

We also consider two types of sampling practices. First, as in the previous examples, we explore convergence of the above chains using all consecutive sampled values. For this purpose we take an entire slice of the chain at an early stage of iteration, as described below. Second, we consider the practice of "thinning", that is, taking a systematic sample of a larger section of the chain. The commonly acknowledged purpose of this practice is to reduce autocorrelation; however this is at the cost of increasing the variance of the estimates. Thus, it is of interest to consider the comparative impact on convergence by this practice, as assessed through our proposed diagnostics.

For case (a), for each algorithm we run 1000 iterations or longer, apply the phase randomisation method, compile the empirical densities, make corresponding assertions about convergence and compare to other diagnostics in CODA. Figure 5 shows the traceplot of 1000 iterations of the Random-Walk algorithm. Convergence assessment is undertaken separately for cycles 100-300 and a subchain of size 300 comprising every 8th cycle of the chain. Figure 6 exhibits the histogram and QQ-plot of the third cumulant using phase randomisation for cycles 100-300 and the subchain. Tables 3 presents the results of other diagnostic tests over these ranges.

In assessing convergence of cycles 100-300, the histogram in Figure 6(i) is almost unimodal but with a slight tail for the third cumulant, indicating that the chain almost converges but more runs than just these 300 are needed as confirmation. This is supported by the variable conclusions reached by the Geweke, Heidebelger & Welch, Raftery & Lewis and autocorrelation tests as summarised in Table 3. Over the subchain of systematically sampled cycles, Figure 6(ii) shows that the histogram of the third-cumulant is almost unimodal and symmetric, with an improved normality approximation as shown in QQ-plot compared to that in (i). Other diagnostic tests for the subchain are similar to the results of cycles 100-300 as shown in Table 3.

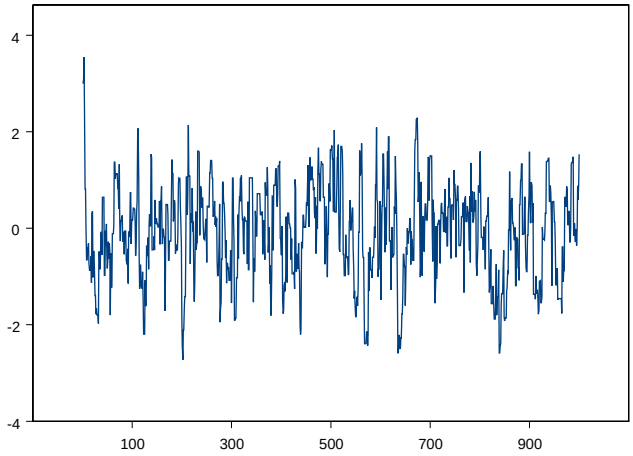

Figure 5. Traceplot of parameter μ, 1000 iterations using Random-Walk algorithm for case (a)

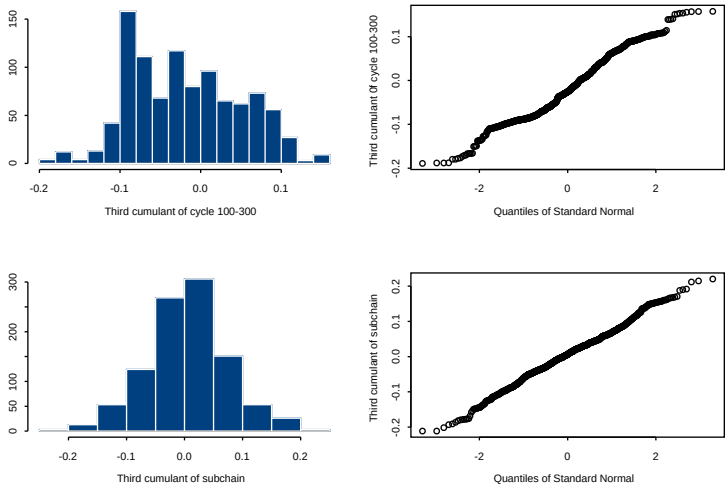

Figure 6. Histograms and QQ-plots for the third cumulant of (i) cycles 100-300 (top) and (ii) the subchain (below) of Figure 5

TABLE 3

Results of diagnostic tests for cycles 100-300 and subchain of the series in Figure 5

Tests	*100-300 cycles*	*Subchain*
Geweke	0.189, passed	0.448, passed
Raftery&Lewis	3.98, failed	8.05, failed
Heidelberger&Welch	0.203, failed	0.0592, failed
Autocorrelations	Failed	Conditionally passed
Phase randomisation	Failed	Passed

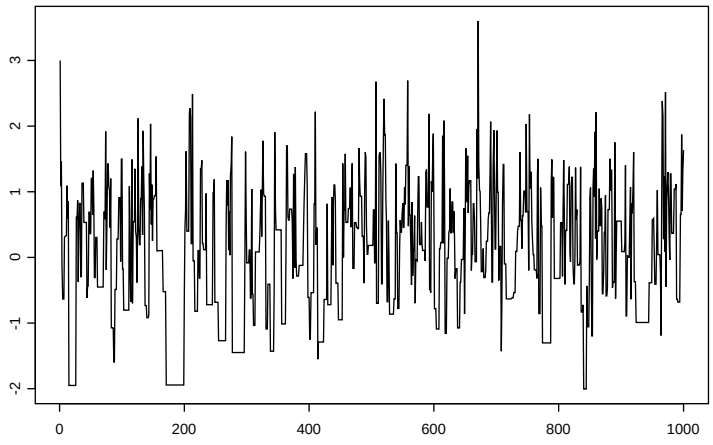

Figure 7. Traceplot of parameter μ, 1000 iterations using Independent algorithm for case (a)

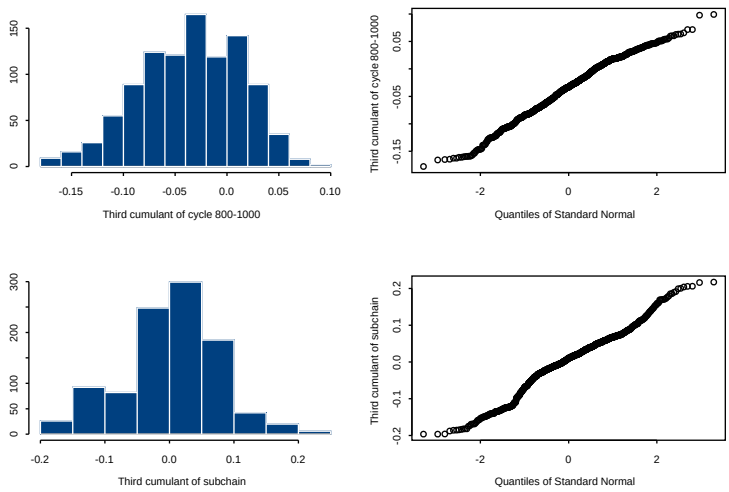

Figure 8. Histogram and QQ-plots for the third cumulants of (i) cycles 800-1000 and (ii) subchain of Figure 7

TABLE 4

Results of diagnostic tests for cycles 800-1000 and the subchain of the series in Figure 7

Tests	*800-1000 cycles*	*Subchain*
Geweke	2.44, failed	0.0713, passed
Raftery&Lewis	8.97, failed	66.1, failed
Heidelberger&Welch	0.226, failed	0.0536, passed
Autocorrelations	High, failed	Low, passed
Phase randomisation	Bimodal, failed	Unimodal but skew, failed

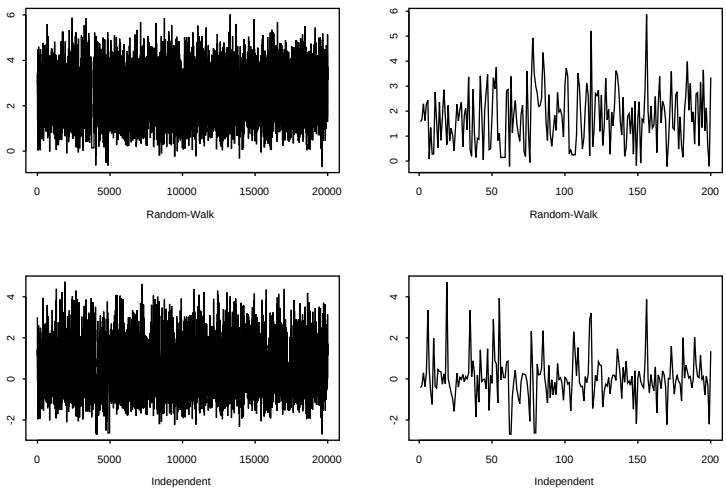

Figure 9. Traceplot of parameter μ for case (b) of (i) the first 20000 iterations and (ii) the subchains using Random-Walk algorithm (top) and Independent algorithm (below)

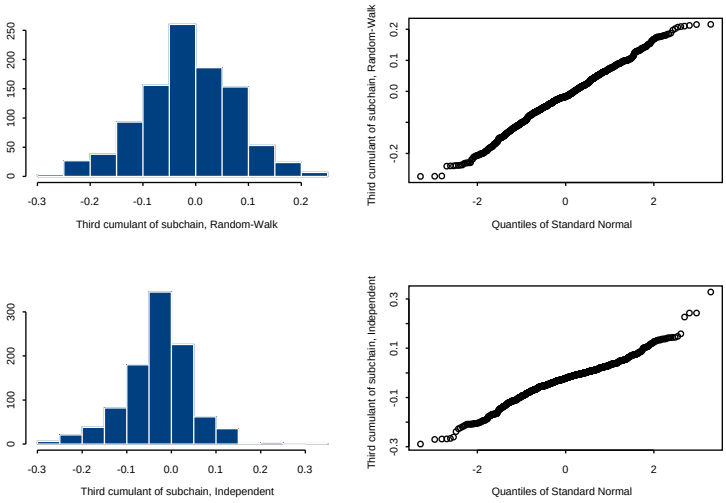

Figure 10. Histogram and QQ-plots for the third cumulants of the subchain depicted in Figure 9 by Random-Walk (top) and Independent (below) respectively

TABLE 5

Results of diagnostic tests for the subchain of the series in Figure 16 by Random-Walk and Independent algorithms

Tests	*Random-Walk*	*Independent*
Geweke	-2.16, failed	1.09, passed
Raftery&Lewis	38.5, failed	116, failed
Heidelberger&Welch	0.45, failed	0.0803, failed
Autocorrelations	High, Failed	Low, passed
Phase randomisation	Unimodal, conditionally passed	Unimodal but skew, failed

Estimation of case (a) using the Independent Metropolis-Hastings algorithm is depicted in Figure 7. The histogram and QQ-plots of the third cumulant estimates for (i) cycles 800-1000 and (ii) the subchain are shown Figure 8. The histogram of the third cumulant estimates for cycles 800-1000 shows that the chain is not stable, as also confirmed by the normality plots in (i). A similar conclusion is reached for the subchain, as supported by the QQ-plot in (ii). Table 4 presents other diagnostic tests for cycles 800-1000 and the subchain in which all tests failed to confirm the convergence for cycles 800-1000, but confirm the convergence for the subchain except for the phase randomisation and Raftery & Lewis test.

For case (b), in which we expect convergence in the long run for the Random-Walk algorithm but not convergence to its stationary distribution for the Independent algorithm, we run 20000 iterations and then apply phase randomisation for the subchain of every 50th iteration. The traceplots of the chains and subchain for each algorithm are shown in Figure 9. The histogram of the third cumulant estimates and corresponding QQ-plots of both algorithms are depicted in Figure 10. It is apparent that approximate normality is achieved by the Random-Walk algorithm, but not by the Independent algorithm. This is in line with other diagnostics as in Table 5. All tests except phase randomisation fail to confirm convergence in the long run for the Random-Walk algorithm, and all other tests except Geweke failed for the Independent algorithm of case (b).

4 Formal Tests for Convergence

As discussed in Section 1, Nur (2003) showed that the third cumulant estimates are asymptotically normally distributed under the stationarity condition. The proof follows by verifying conditions for the validity of formal Edgeworth expansions for the distribution of third cumulants of stationary linear and nonlinear processes that can be representated by a Volterra expansion. The conditions are: the existence of the higher moments and uniform bounded properties; the approximation of the random vectors by other random vectors; strong mixing; conditional Crámer condition; approximation which satisfies a Markov type condition and finally the existence of a variance-covariance matrix. This work is based on Götze and Hipp (1994) who presented verifiable conditions for the validity of formal Edgeworth expansions for the distribution of sums of random variables of more general processes which include linear processes as well as the nonlinear AR processes. The results apply to many statistics in nonlinear time series models.

In Section 3 we considered visual convergence assessment using the proposed phase randomisation method. By considering the results in Nur (2003), it is thus interesting to consider whether we might formalise the assessment of convergence by a test of normality of the third cumulant, for example through Kolmogorov-Smirnov or Shapiro & Wilk tests. We consider this approach for the examples in the previous section. This approach is also motivated by Robert *et al* (1998) in which they use Kolmogorov-Smirnov and Shapiro & Wilk tests as a diagnostic by appeal to the CLT.

For the surgical example of Section 3.1, the P-value of the Kolmogorov-Smirnov test applied to the data displayed in Figure 4 is 0.091. It is apparent that the QQ-plot and the Kolmogorov-Smirnov test favour the assertion of asymptotic normality. The Shapiro & Wilk test, on the other hand, is much more strict, giving $P = 0$. However, by taking longer chains within each surrogate, normality is more strongly verified. Indeed, when a chain length of 400 cycles was evaluated, the P-value from Shapiro & Wilk changes to 0.004. By evaluating a longer chain, the Shapiro & Wilk test is in favour of the assertion of asymptotic normality.

The Metropolis-Hastings algorithm seems to give rather different results which may be due to the stronger correlation between iterations compared to the Gibbs algorithm of previous examples. For the Random-Walk algorithm of case (a), the Kolmogorov-Smirnov test of normality for the third cumulant based on iterations 100-300 (Figure 6 (i)) gives a P-value of 0. However, the same test applied to the systematic subsampled chain (Figure 6 (ii)) produces a P-value of 0.0514.

Interestingly, after subsampling, normality is more strongly verified asymptotically.

For the Random-Walk algorithm of case (b) using systematic subsampling, the Kolmogorov-Smirnov test of the data in Figure 10 gives $P = 0.0037$, again a marginal improvement from $P = 0$ obtained without systematic subsampling. The P-values corresponding to the Independent algorithm for both cases (Figures 8 and 10) are zero, subsampling or not, leading to a strong conclusion of non-normality.

In light of the above results, this phase randomisation test might also be used to determine the thinning interval for MCMC chains. As the interval between the systematic subsample increases, the P-value corresponding to the Kolmogorov-Smirnov test of normality of the third cumulant also increases. For example, for the Random-Walk algorithm of case (a), the P-values associated with thinning intervals of 5 and 8 are 0.0144 and 0.0514, respectively. If the thinning interval is identified when normality is accepted by this test, in this example, we would take systematic subsampling of every 8th iteration. Using this approach, the appropriate thinning interval using the Random-Walk algorithm for case (a) and (b) respectively are 8 and 50. Using the Independent algorithm, no thinning interval can be found to satisfy the test, further confirming lack of convergence. Of course, this approach is computationally intensive and further research is being undertaken to reduce this effort. Moreover, these advantages must be evaluated against the known drawbacks of the practice, as discussed earlier.

5 Discussion

An interesting set of conclusions can be drawn from the three examples in Section 3. First, the phase randomisation performs at least as well as other diagnostics tests and is more informative about the behaviour of third and higher order cumulants which is important in characterising certain forms of nonlinearity and nonstationarity. Furthermore, under the stationary assumption, the third cumulant estimates obtained by phase randomisation are conjectured to be asymptotically normally distributed and can be visually and formally tested. It is thus a valuable addition to the diagnostic toolbox. Second, this new tool has similar features to the autocorrelation test, Heidelberger & Welch test and Raftery & Lewis test. These are quite sensitive to the measure of the dependence of the data, that is, they are concerned with convergence to iid sampling. On the other

hand, Geweke's test is concerned with convergence to the mean. Third, the phase randomisation diagnostic supports the result of Mengersen & Tweedie (1996) that the Metropolis-Hastings algorithm results in a Markov chain which is geometrically ergodic to the average when the target density is log-concave in the tails. The investigations of Section 4 have also suggested that phase randomisation may be used to identify appropriate subsampling strategic for Metropolis-Hastings chains. The disadvantages of this method were also identified in particular its computational burden.

Although it is ideal and somewhat expected that all of the various diagnostics are in agreement about convergence of a particular chain, disagreement is not uncommon in practice. This is supported by the studies of Cowles and Carlin (1996), who advocate the use of multiple diagnostics in any study of convergence. The intention is that this might reveal different features of the chain and overcome individual deficiencies in the individual diagnostic methods. Such disagreement among diagnostics was indeed observed in the examples of Section 3.3. In this light, the addition of a new diagnostic that reveals distinct features of a chain as described above maybe welcome.

Because the phase randomisation method has a strong connection with bootstrapping, then it would be possible to develop a diagnostic test for multiple chains as well as a single chain. Moreover, the phase randomisation method could suggest the number of iterations in the burn-period. Based on theoretical results of the asymptotic distribution of third cumulant in time series, one could construct a statistical test to obtain the number of iterations that should be run to ensure the convergence of the MCMC chain.

References

Besag, J., Green, P., Higdon, D., Mengersen, K.L. (1995). Bayesian computation and stochastic systems. *Statistical Science* **10**, 3-66.

Best, N., Cowles, M.K., Vines, L. (1995). CODA Manual Version 0.30. MRC Biostatistics Unit, Cambridge.

Braun, W. J. & Kulperger, R.J. (1997). Properties of a Fourier Bootstrap Method for Time Series. *Comm. Statist- Theory Meth.* **26,** 1329-1336.

Chan,K.S. (1997). On the Validity of the Method of Surrogate Data. *Field Institute Communications* **17** 77-97.

Cowles, M.K. & Carlin, B.P. (1996). Markov Chain Monte Carlo convergence diagnostics : a comparative study. *Journ. of Amer. Statist. Assoc.* **91,** 883-904.
Davison, A.C & Hinkley,D.V. (1997). *Bootstrap Methods and their Applications*, Cambridge University Press.
Gelfand, A.E., Hills, S.E., Racine-Poon, A. and Smith, A.F.M. (1990). Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling. *Journ. of Amer. Statist. Assoc.* **85,** 972-985.
Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics* **4** (J.M. Bernardo, J. Berger, A.P.Dawid, A.F.M. Smith, eds), Oxford Univ. Press, 169-193.
Götze, F. and Hipp, C. (1994). Asymptotic Distribution of Statistics in Time Series. *Annals of Statistics* **22**, 2062-2088.
Heidelberger, P. & Welch, P.D. (1983). A spectral method for confidence interval generation and run-length control in simulation. *Communication ACM* **24,** 233-245.
Mengersen, K.L. & Robert, C.P. (1999). MCMC Convergence Diagnostics: A Review, in Bernardo, J.M., Berger, J.O., David, A.P. and Smith, A.F.M.(editors) *Bayesian Statistics* Vol 6 (Oxford Univ. Press) 415-440.
Mengersen, K.L. & Tweedie, R.L. (1996). Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Stats.* **24,** 101-121.
Nur,D., Wolff,R.C. & Mengersen, K.L. (2001). Phase Randomisation : Numerical Study of Higher Cumulants Behaviour. *Computational Statistics and Data Analysis,* **37,** 487-513.
Nur,D.(2003). Asymptotic distribution of third cumulant in time series. Submitted to *Statistics*.
Raftery, A.E. & Lewis, S. (1992). Implementing MCMC. In *MCMC in Practice.* W.R.Gilks, S.T. Richardson, D.J.Spiegelhalter (Eds), Chapman and Hall London, 115-130.
Robert, C.P. & Casella, G. (1999). *Monte Carlo Statistical Methods.* Springer-Verlag, New York.
Robert, C.P., Ryden, T. & Titterington, D.M. (1998). Convergence controls for MCMC algorithms with applications to Hidden Markov Chains. Research report No 9805, Centre de Recherche en Economie et Statistique, INSEE, France.
Schittenkopf, C. & Deco, G. (1997). Testing Nonlinear Markovian Hypotheses in Dynamical Systems. *Physica D,* **104**,61-74.
Schreiber, T. & Schmitz, A. (1996). Improved Surrogate Data for Nonlinearity

Tests. *Physical Review Letters,* **77**,635-638.
Schreiber, T. (1998). Constrained Randomization of Time Series Data. *Physical Review Letters,* **80**,2105-2108.
Schreiber, T. (1999). Interdisciplinary Application of Nonlinear Time Series Methods. *Physics Reports,* **308**,1-64.
Shapiro, S. & Wilk, M. (1965). An analysis of variance test for normality. *Biometrika,* **52,** 591-611.
Spiegelhalter, D.J., Thomas, A., Best, N. G. & Gilks, W.R. (1994). BUGS : Bayesian Inference Using Gibbs Sampling, Version 0.4. Technical Report, MRC Biostatistic Unit, University of Cambridge.
Theiler, J., Eubank, S., Longtin, A., Galdrikian, B. & Farmer, J. D. (1992). Testing for Nonlinearity in Time Series: the Method of Surrogate Data. *Physica D,* **58**,77-94.
Theiler, J. & Prichard, D. (1996). Constrained Realizations Monte-Carlo Method for Hypothesis Testing. *Physica D,* **94**,221-235.
Timmer, J. (1998). Power of surrogate data testing with respect to non-stationarity. *Physical Review E,* **58**,5153-5156.