*Biometrika* (2019), **106**, 2, *pp.* 321–337
*Printed in Great Britain*
doi: 10.1093/biomet/asz002
Advance Access publication 3 April 2019

# Multivariate output analysis for Markov chain Monte Carlo

BY DOOTIKA VATS

*Department of Statistics, University of Warwick, Coventry CV4 7AL, U.K.*
dootika.vats@gmail.com

JAMES M. FLEGAL

*Department of Statistics, University of California, Riverside, 900 University Ave, Riverside, California 92521, U.S.A.*
jflegal@ucr.edu

AND GALIN L. JONES

*School of Statistics, University of Minnesota, 224 Church St SE, Minneapolis, Minnesota 55455, U.S.A.*
galin@umn.edu

SUMMARY

Markov chain Monte Carlo produces a correlated sample which may be used for estimating expectations with respect to a target distribution. A fundamental question is: when should sampling stop so that we have good estimates of the desired quantities? The key to answering this question lies in assessing the Monte Carlo error through a multivariate Markov chain central limit theorem. The multivariate nature of this Monte Carlo error has been largely ignored in the literature. We present a multivariate framework for terminating a simulation in Markov chain Monte Carlo. We define a multivariate effective sample size, the estimation of which requires strongly consistent estimators of the covariance matrix in the Markov chain central limit theorem, a property we show for the multivariate batch means estimator. We then provide a lower bound on the number of minimum effective samples required for a desired level of precision. This lower bound does not depend on the underlying stochastic process and can be calculated a priori. This result is obtained by drawing a connection between terminating simulation via effective sample size and terminating simulation using a relative standard deviation fixed-volume sequential stopping rule, which we demonstrate is an asymptotically valid procedure. The finite-sample properties of the proposed method are demonstrated in a variety of examples.

*Some key words*: Covariance matrix estimation; Effective sample size; Markov chain Monte Carlo; Multivariate analysis.

## 1. INTRODUCTION

Markov chain Monte Carlo algorithms are often used to estimate expectations with respect to a probability distribution when obtaining independent samples is difficult. Typically, interest is in estimating a vector of quantities. However, analysis of Markov chain Monte Carlo

output routinely focuses on inference about complicated joint distributions only through their marginals, despite the fact that the assumption of independence across components holds only rarely in settings where Markov chain Monte Carlo is relevant. Thus standard univariate convergence diagnostics, sequential stopping rules for termination, effective sample size definitions, and confidence intervals all lead to an incomplete understanding of the estimation process. We overcome the drawbacks of univariate output analysis by developing a methodological framework for multivariate analysis of Markov chain Monte Carlo output.

Let $F$ be a distribution with support $\mathcal{X}$ and $g : \mathcal{X} \to \mathbb{R}^p$ be an $F$-integrable function such that $\theta = E_F(g)$ is of interest. If $\{X_t\}$ is an $F$-invariant Harris recurrent Markov chain, set $\{Y_t\} = \{g(X_t)\}$ and estimate $\theta$ with $\theta_n = n^{-1} \sum_{t=1}^{n} Y_t$ since $\theta_n \to \theta$, with probability 1, as $n \to \infty$. Finite sampling leads to an unknown Monte Carlo error, $\theta_n - \theta$, the estimation of which is essential to assessing the quality of estimation. If for some $\delta > 0$, $g$ has $2 + \delta$ moments under $F$ and $\{X_t\}$ is polynomially ergodic of order $m > (2 + \delta)/\delta$, an approximate sampling distribution for the Monte Carlo error is available via a Markov chain central limit theorem (Geyer, 2011; Jones, 2004). That is, there exists a $p \times p$ positive-definite matrix, $\Sigma$, such that as $n \to \infty$,

$$n^{1/2}(\theta_n - \theta) \xrightarrow{\mathrm{d}} N_p(0, \Sigma). \tag{1}$$

The central limit theorem describes asymptotic behaviour of the Monte Carlo error, and the strong law for $\theta_n$ ensures that large $n$ leads to a small Monte Carlo error. But how large is large enough? This question has not been adequately addressed in the literature since current approaches are based on the univariate central limit theorem,

$$n^{1/2}(\theta_{n,i} - \theta_i) \xrightarrow{\mathrm{d}} N(0, \sigma_i^2) \tag{2}$$

as $n \to \infty$, where $\theta_{n,i}$ and $\theta_i$ are the $i$th components of $\theta_n$ and $\theta$ respectively and $\sigma_i^2$ is the $i$th diagonal element of $\Sigma$. The output analysis tools of Atchadé (2011; 2016), Flegal & Jones (2010), Flegal & Gong (2015), Gelman & Rubin (1992), Gong & Flegal (2016), and Jones et al. (2006) are all based on the central limit theorem in (2). These methods ignore cross-correlation among components, leading to an inaccurate understanding of the estimation process. Consider the example of a two-component Markov chain $\{X_t^{(1)}, X_t^{(2)}\}$, where interest is in estimating the covariance between the two components, $\mathrm{cov}_F(X^{(1)}, X^{(2)}) = E_F(X^{(1)}X^{(2)}) - E_F(X^{(1)})E_F(X^{(2)})$. The Monte Carlo standard error for the sample covariance is obtained by first estimating the asymptotic covariance matrix in the multivariate central limit theorem for the vector $(X^{(1)}, X^{(2)}, X^{(1)}X^{(2)})$, followed by an application of the delta method. A univariate central limit theorem for $X^{(1)}X^{(2)}$ would completely ignore the sampling error for the mean estimates of $\mathrm{E}_F(X^{(1)})$ and $\mathrm{E}_F(X^{(2)})$.

Ignoring the joint sampling distribution of Monte Carlo estimators affects the performance of the stopping rules of Jones et al. (2006); where simulation is terminated when the size of the confidence interval constructed using (2) is smaller than some prespecified tolerance level. Such confidence intervals are constructed for all $p$ quantities in $\theta$, thus requiring a multiple testing correction. We propose the relative standard deviation fixed-volume sequential stopping rule, which differs from the Jones et al. (2006) procedure in two fundamental ways; firstly, it is motivated by the multivariate central limit theorem in (1) and not the univariate central limit theorem in (2); and secondly, it terminates simulation not by the absolute size of the confidence region, but by its size relative to the inherent variability in the problem. Specifically, simulation stops when the Monte Carlo standard error is small compared to the variability in the target distribution. Naturally, an estimate of the Monte Carlo standard error is required and, for now, we assume that $\Sigma$ in (1) can be estimated consistently. Let $\Lambda_n$ be the sample covariance matrix, det

denote determinant, $\alpha$ be the confidence level, and $\epsilon$ be the desired tolerance level. The relative standard deviation fixed-volume sequential stopping rule terminates the first time after some user-specified $n^* \geqslant 0$ iterations that

$$\{\text{volume of } 100(1-\alpha)\% \text{ confidence region}\}^{1/p} + n^{-1} < \epsilon \det(\Lambda_n)^{1/2p}. \tag{3}$$

The role of $n^*$ is to avoid premature termination due to early bad estimates; we will say more about how to choose $n^*$ in § 3.

Since the determinant of a covariance matrix is also known as the generalized variance (Wilks, 1932), an equivalent interpretation of (3) is that simulation is terminated when the generalized variance of the Monte Carlo error is small relative to the generalized variance of $g$ with respect to $F$; that is, a scaled estimate of $\det(\Sigma)$ is small compared to the estimate of $\det(\Lambda) = \det(\text{var}_F g)$. We call $\det(\Lambda)^{1/2p}$ the relative metric. The practitioner is free to choose from a large class of relative metrics and a different choice of the relative metric leads to a fundamentally different approach to termination. For a wide class of relative metrics, we show that if the estimator of $\Sigma$ is strongly consistent, the stopping rule in (3) is asymptotically valid, in that the confidence regions created at termination have the right coverage probability as $\epsilon \to 0$.

In addition to sequential stopping rules for termination, the univariate central limit theorem in (2) has also been used to terminate simulation using the effective sample size for each component. See Atkinson et al. (2008), Drummond et al. (2006), Giordano et al. (2015), Gong & Flegal (2016), and Kruschke (2014) for a few examples. We develop a multivariate effective sample size, which, to the best of our knowledge, has not been studied in the literature. If $\Lambda$ and $\Sigma$ are of full rank, we define effective sample size as

$$\text{ESS} = n \left\{ \frac{\det(\Lambda)}{\det(\Sigma)} \right\}^{1/p}. \tag{4}$$

For uncorrelated samples, $\Sigma = \Lambda$ and $\text{ESS} = n$. The definition of ESS involves the ratio of generalized variances. This ratio also appears in (3), leading to a key result: terminating according to the relative standard deviation fixed-volume sequential stopping rule is asymptotically equivalent to terminating when the estimated ESS satisfies

$$\hat{\text{ESS}} \geqslant W_{p,\alpha,\epsilon},$$

where $W_{p,\alpha,\epsilon}$ can be calculated a priori and is a function only of the dimension of the estimation problem, the confidence level, and the relative precision desired. Thus, not only do we show that terminating via ESS is a valid procedure, we also provide a theoretically valid, practical lower bound on the number of effective samples required for a desired tolerance level.

Recall that we require a strongly consistent estimator of $\Sigma$. Estimating $\Sigma$ is difficult due to the serial correlation in the Markov chain. Vats et al. (2018) demonstrated strong consistency for a class of multivariate spectral variance estimators, while Dai & Jones (2017) introduced multivariate initial sequence estimators and established their asymptotic validity. However, both estimators are computationally expensive and do not scale well with either $p$ or $n$. Instead, we use the multivariate batch means estimator of $\Sigma$, which is significantly faster to compute and requires weaker moment conditions on $g$ for strong consistency. Our strong consistency result weakens the conditions required in Bednorz & Łatuszyński (2007) and Jones et al. (2006) for the univariate batch means estimator. In particular, we do not require a one-step minorization and only require polynomial ergodicity, as opposed to geometric ergodicity. The condition is fairly

weak since often the existence of a Markov chain central limit theorem itself is demonstrated via polynomial ergodicity or a stronger result; see Jones (2004) for a review. Many Markov chains have been shown to be at least polynomially ergodic. See Doss & Hobert (2010), Hobert & Geyer (1998), Jarner & Hansen (2000), Jarner & Roberts (2002), Johnson et al. (2013), Johnson & Jones (2015), Jones et al. (2014), Jones & Hobert (2004), Khare & Hobert (2013), Marchev & Hobert (2004), Roberts & Polson (1994), Tan et al. (2013), Tan & Hobert (2012), and Vats (2017), among many others.

The multivariate stopping rules significantly improve upon existing univariate methods since termination is dictated by the joint behaviour of the components of the Markov chain and not by the components that mix the slowest, using the inherent multivariate nature of the problem and acknowledging cross-correlations leads to a more realistic understanding of the estimation process, and avoiding corrections for multiple testing yields considerably smaller confidence regions even in moderate-$p$ problems. An illustrative example is given below, where for $i = 1, \ldots, K$, $Y_i$ is a binary response variable and $x_i = (x_{i1}, \ldots, x_{i5})$ are the observed predictors for the $i$th observation. Assume $\tau^2$ is known,

$$Y_i \,|\, x_i, \beta \overset{\text{ind}}{\sim} \text{Ber}\left\{\frac{1}{1+\exp(-x_i\beta)}\right\}, \quad \beta \sim N_5(0, \tau^2 I_5). \tag{5}$$

This simple model produces an intractable posterior, $F$ on $\mathbb{R}^5$. The dataset used is the `logit` dataset in the `R` package `mcmc` (Geyer & Johnson, 2015; R Development Core Team, 2019), where $K = 100$ and $\tau^2 = 1$. The goal is to estimate the posterior mean of $\beta$, $\mathrm{E}_F(\beta) \in \mathbb{R}^5$. Thus $g$ is the identity function. We implement a random walk Metropolis–Hastings algorithm with a multivariate normal proposal distribution $N_5(\,\cdot\,, 0.35^2 I_5)$, where $I_5$ is the $5 \times 5$ identity matrix and the 0.35 scaling approximates the optimal acceptance probability suggested by Roberts et al. (1997). The starting value for the chain is a draw from the prior distribution.

We calculate the Monte Carlo estimate for $\mathrm{E}_F(\beta)$ from a sample of size $10^5$ and use multivariate batch means to estimate $\Sigma$; the estimator is described in § 4. We also implement the univariate methods described by Jones et al. (2006) to estimate $\sigma_i^2$, which ignore the cross-correlation in the $\beta$s. This cross-correlation is often significant as seen in Fig. 1, and can only be captured by multivariate methods. In Fig. 1 we present 90% confidence regions created using multivariate and univariate batch means estimators for $\beta_1$ and $\beta_3$; for the purpose of this figure, we set $p = 2$. This figure illustrates why multivariate methods are likely to outperform univariate methods. The confidence ellipse is the smallest-volume region for a particular level of confidence. Thus, these confidence ellipses should be preferred over other confidence regions.

We obtain coverage probabilities of the confidence regions over 1000 replications for Monte Carlo sample sizes in $\{10^4, 10^5, 10^6\}$. The true posterior mean, $(0.5706, 0.7516, 1.0559, 0.4517, 0.6545)$, was obtained by averaging over $10^9$ iterations. The volume of the confidence region to the $p$th root was also observed. Table 1 summarizes the results. Although the uncorrected univariate methods produce the smallest confidence regions, their coverage probabilities are far from desirable. For a large Monte Carlo sample size, multivariate batch means produce 90% coverage probabilities with systematically lower volume than Bonferroni-corrected univariate batch means.

Thus even easy Markov chain Monte Carlo problems produce samples with complex dependence structures. Not only do we gain more information about the Monte Carlo error by using multivariate methods, we also avoid conservative Bonferroni corrections.

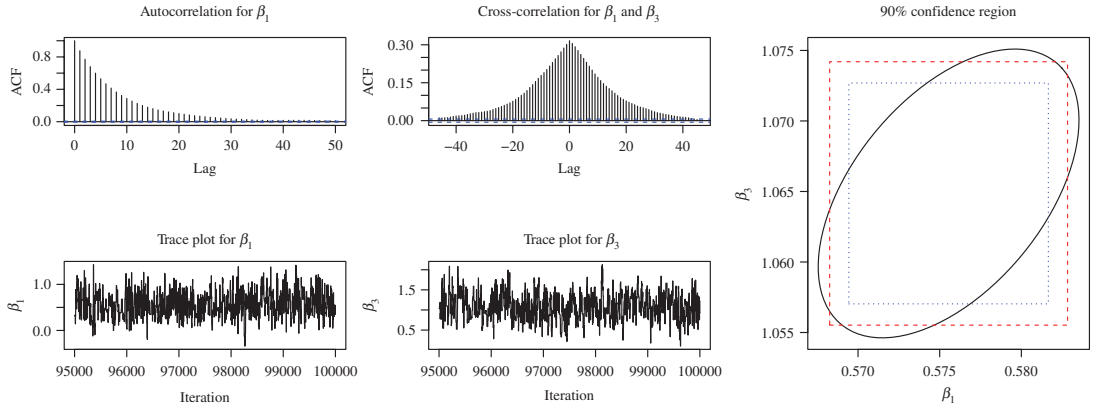


Fig. 1. Autocorrelation plot for $\beta_1$, cross-correlation plot of $\beta_1$ and $\beta_3$, trace plots for $\beta_1$ and $\beta_3$, and 90% confidence regions for $\beta_1$ and $\beta_3$. The ellipse is constructed using multivariate batch means, the dotted box using uncorrected univariate batch means, and the dashed box using Bonferroni-corrected univariate batch means.

Table 1. *Volume to the pth* ($p = 5$) *root and coverage probabilities for* 90% *confidence regions with 1000 replications and standard errors in parentheses*

| | $n$ | Multivariate | Bonferroni-corrected univariate | Uncorrected univariate |
|---|---|---|---|---|
| Volume to the $p$th root | $10^4$ | 0.056 (7.22e-05) | 0.066 (9.23e-05) | 0.046 (6.48e-05) |
| | $10^5$ | 0.018 (1.09e-05) | 0.021 (1.42e-05) | 0.015 (1.00e-05) |
| | $10^6$ | 0.006 (1.70e-06) | 0.007 (2.30e-06) | 0.005 (1.60e-06) |
| Coverage probabilities | $10^4$ | 0.876 (0.0104) | 0.889 (0.0099) | 0.596 (0.0155) |
| | $10^5$ | 0.880 (0.0103) | 0.910 (0.0090) | 0.578 (0.0156) |
| | $10^6$ | 0.894 (0.0097) | 0.903 (0.0094) | 0.627 (0.0153) |

## 2. Termination rules

Let $\sigma^2_{n,i}$ be a strongly consistent estimator of $\sigma_i^2$ and $t_*$ be an appropriate $t$-distribution quantile. For a desired tolerance, $\epsilon_i$, Jones et al. (2006) considered the fixed-width sequential stopping rule which terminates simulation the first time after $n^* \geqslant 0$ iterations, for all $i$ components, that

$$t_* \frac{\sigma_{n,i}}{n^{1/2}} + n^{-1} \leqslant \epsilon_i.$$

The role of $n^*$ is to ensure a minimum simulation effort, as defined by the user, so as to avoid poor initial estimates of $\sigma_i^2$. This rule laid the foundation for termination based on the quality of estimation of $\theta$. As a consequence, estimation is reliable in the sense that if the procedure is repeated again, the estimates will not be vastly different (Flegal et al., 2008). However, implementing the fixed-width sequential stopping rule can be challenging since careful analysis is required for choosing $\epsilon_i$ for each $\theta_{n,i}$, which can be tedious or even impossible for large $p$; to ensure the right coverage probability, $t_*$ is chosen to account for multiple confidence intervals. Thus when $p$ is even moderately large, the termination rule can be aggressively conservative; simulation stops when each component satisfies the termination criterion; therefore, all cross-correlations are ignored and termination is governed by the slowest-mixing components; and it ignores correlation in the target distribution.

We construct a class of multivariate sequential termination rules that lead to asymptotically valid confidence regions. Let $T^2_{1-\alpha,p,q}$ denote the $1-\alpha$ quantile of a Hotelling's $T$-squared

distribution with dimensionality parameter $p$ and degrees of freedom $q$. Throughout this section and the next, we assume that $\Sigma_n$ is a strongly consistent estimator of $\Sigma$ and $q$ is determined by the choice of $\Sigma_n$. A $100(1-\alpha)\%$ confidence region for $\theta$ is the set

$$C_\alpha(n) = \left\{\theta \in \mathbb{R}^p : n(\theta_n - \theta)^{\mathrm{T}} \Sigma_n^{-1} (\theta_n - \theta) < T^2_{1-\alpha,p,q}\right\}.$$

Then $C_\alpha(n)$ forms a $p$-dimensional ellipsoid oriented along the directions of the eigenvectors of $\Sigma_n$. The volume of $C_\alpha(n)$ is

$$\mathrm{Vol}\{C_\alpha(n)\} = \frac{2\pi^{p/2}}{p\Gamma(p/2)} \left(\frac{T^2_{1-\alpha,p,q}}{n}\right)^{p/2} |\Sigma_n|^{1/2}. \tag{6}$$

Since $p$ is fixed and $\Sigma_n \to \Sigma$ with probability 1, $\mathrm{Vol}\{C_\alpha(n)\} \to 0$, with probability 1, as $n \to \infty$. If $\epsilon > 0$ and $s(n)$ is a positive real-valued function defined on the positive integers, then a fixed-volume sequential stopping rule terminates the simulation at the random time

$$T(\epsilon) = \inf\left\{n \geqslant 0 : \mathrm{Vol}\{C_\alpha(n)\}^{1/p} + s(n) \leqslant \epsilon\right\}. \tag{7}$$

Glynn & Whitt (1992) provide conditions so that terminating at $T(\epsilon)$ yields confidence regions that are asymptotically valid in that, as $\epsilon \to 0$, $\mathrm{pr}[\theta \in C_\alpha\{T(\epsilon)\}] \to 1-\alpha$. In particular, they let $s(n) = \epsilon I(n < n^*) + n^{-1}$, which ensures simulation does not terminate before $n^*$ iterations. The rule in (7) can be difficult to implement in practice since the choice of $\epsilon$ depends on the units of $\theta$, which has to be carefully chosen for every application. We present a more natural alternative to (7), which we then connect to the concept of effective sample size.

Let $\|\cdot\|$ denote the Euclidean norm. Let $K(F,g) > 0$ be a functional and suppose $K_n > 0$ is an estimator of $K(F,g)$; for example, $K(F,g) = \|\theta\|$ and $K_n = \|\theta_n\|$. We call $K(F,g)$ the relative metric. Assume $s(n) = o(n^{-1/2})$ and define

$$T^*(\epsilon) = \inf\left\{n \geqslant 0 : \mathrm{Vol}(C_\alpha(n))^{1/p} + s(n) \leqslant \epsilon K_n\right\}.$$

THEOREM 1. *Let $g : \mathcal{X} \to \mathbb{R}^p$ be such that $\mathrm{E}_F\left(\|g\|^{2+\delta}\right) < \infty$ for some $\delta > 0$ and let $\{X_t\}$ be an $F$-invariant polynomially ergodic Markov chain of order $m > (1+\epsilon_1)(1+2/\delta)$ for some $\epsilon_1 > 0$. If $K_n \to K(F,g)$ with probability 1 and $\Sigma_n \to \Sigma$ with probability 1 as $n \to \infty$, then as $\epsilon \to 0$, $T^*(\epsilon) \to \infty$ and $\Pr[\theta \in C_\alpha\{T^*(\epsilon)\}] \to 1-\alpha$.*

Proofs of all the theorems are provided in the Supplementary Material.

*Remark* 1. Theorem 1 applies when $K(F,g) = K_n = 1$. This choice of relative metric leads to the absolute-precision fixed-volume sequential stopping rule, a multivariate generalization of the procedure considered by Jones et al. (2006).

Suppose $K(F,g) = \det(\Lambda)^{1/2p} = \det(\mathrm{var}_F g)^{1/2p}$, and if $\Lambda_n$ is the usual sample covariance matrix for $\{Y_t\}$, set $K_n = \det(\Lambda_n)^{1/2p}$ as long as $\Lambda_n$ is positive definite. Then $K_n \to K(F,g)$, with probability 1, as $n \to \infty$, and $T^*(\epsilon)$ is the first time the variability in estimation, measured via the volume of the confidence region, is an $\epsilon$th fraction of the variability in the target distribution. Set $s(n) = \epsilon \det(\Lambda)^{1/2p} I(n < n^*) + n^{-1}$. The relative standard deviation fixed-volume sequential stopping rule is formalized as terminating at the random time

$$T_{\mathrm{SD}}(\epsilon) = \inf\left\{n \geqslant 0 : \mathrm{Vol}(C_\alpha(n))^{1/p} + s(n) \leqslant \epsilon \det(\Lambda_n)^{1/2p}\right\}. \tag{8}$$

*Remark* 2. For $p = 1$, $T_{SD}(\epsilon)$ reduces to the relative standard deviation fixed-width sequential stopping rule of Flegal & Gong (2015).

## 3. Effective sample size

Let $\rho(Y_1^{(i)}, Y_{1+k}^{(i)})$ be the lag-$k$ correlation for the $i$th component of $Y_1$ and $\lambda_i^2$ be the $i$th diagonal element of $\Lambda$. Gong & Flegal (2016), Kass et al. (1998), Liu (2008), and Robert & Casella (2013) define ESS for the $i$th component as

$$\text{ESS}_i = \frac{n}{1 + 2\sum_{k=1}^{\infty} \rho(Y_1^{(i)}, Y_{1+k}^{(i)})} = n\frac{\lambda_i^2}{\sigma_i^2}. \tag{9}$$

A strongly consistent estimator of $\text{ESS}_i$ is obtained through strongly consistent estimators of $\lambda_i^2$ and $\sigma_i^2$ via the sample variance, $\lambda_{n,i}^2$, and univariate batch means estimators, $\sigma_{n,i}^2$, respectively. Then $\text{ESS}_i$ is estimated for each component separately and a conservative estimate of the overall effective sample size is the minimum of all $\text{ESS}_i$. Consequently, the effective sample size estimate is dictated by the components that mix the slowest.

In (4), instead of using just the diagonals of $\Lambda$ and $\Sigma$ to define effective sample size, we use the matrices themselves. As a univariate measure of spread for a multivariate distribution, Wilks (1932) described the determinant of the covariance matrix of a random vector as its generalized variance. Wilks (1932) and SenGupta (1987) recommend the use of the $p$th root of the generalized variance. When $p = 1$, ESS reduces to the univariate effective sample size in (9). Let $\Lambda_n$ be the sample covariance matrix of $\{Y_t\}$ and $\Sigma_n$ be a strongly consistent estimator of $\Sigma$. Then a strongly consistent estimator of ESS is

$$\hat{\text{ESS}} = n\left\{\frac{\det(\Lambda_n)}{\det(\Sigma_n)}\right\}^{1/p}.$$

We now consider a lower bound for the effective sample size. Rearranging the defining inequality in (8) yields that when $n \geqslant n^*$,

$$\hat{\text{ESS}} \geqslant \left[\left\{\frac{2\pi^{p/2}}{p\Gamma(p/2)}\right\}^{1/p}\left(T_{1-\alpha,p,q}^2\right)^{1/2} + \frac{\det(\Sigma_n)^{-1/2p}}{n^{1/2}}\right]^2 \frac{1}{\epsilon^2} \approx \frac{2^{2/p}\pi}{(p\Gamma(p/2))^{2/p}}\frac{T_{1-\alpha,p,q}^2}{\epsilon^2}.$$

Thus, terminating at $T_{SD}(\epsilon)$ is equivalent to terminating the first time $\hat{\text{ESS}}$ is larger than a lower bound. This lower bound is a function of $n$ through $q$ and thus is difficult to calculate a priori. However, as $n \to \infty$, $T_{p,q}^2$ converges to a $\chi_p^2$, leading to the following approximation:

$$\hat{\text{ESS}} \geqslant \frac{2^{2/p}\pi}{\{p\Gamma(p/2)\}^{2/p}}\frac{\chi_{1-\alpha,p}^2}{\epsilon^2}. \tag{10}$$

One can a priori determine the number of effective samples required for their choice of $\epsilon$ and $\alpha$. As $p \to \infty$, the lower bound in (10) converges to $2\pi e/\epsilon^2$. Thus for large $p$, the lower bound is mainly determined by the choice of $\epsilon$. On the other hand, for a fixed $\alpha$, having obtained $W$ effective samples, we can use the lower bound to determine the relative precision, $\epsilon$, in the estimation. In this way, (10) can be used to make informed decisions regarding termination.

*Example* 1. Suppose $p = 5$, as in the logistic regression setting of § 1, and for a 95% confidence region we want a relative precision of $\epsilon = 0.05$; that is, the Monte Carlo error is 5% of the variability in the target distribution . This requires $\hat{\text{ESS}} \geqslant 8605$. On the other hand, if we simulate until $\hat{\text{ESS}} = 10\,000$, we obtain a relative precision of $\epsilon = 0.0464$.

*Remark* 3. Let $n_{\text{pos}}$ be the smallest integer such that $\Sigma_{n_{\text{pos}}}$ and $\Lambda_{n_{\text{pos}}}$ are positive definite; in the next section we will discuss how to choose $n_{\text{pos}}$ for the multivariate batch means estimator. In light of the lower bound in (10), a natural choice of $n^*$ is

$$n^* \geqslant \max\left[n_{\text{pos}}, \frac{2^{2/p}\pi}{\{p\Gamma(p/2)\}^{2/p}} \frac{\chi^2_{1-\alpha,p}}{\epsilon^2}\right].$$

## 4. STRONG CONSISTENCY OF THE MULTIVARIATE BATCH MEANS ESTIMATOR

We present the multivariate batch means estimator and provide conditions for strong consistency. Let $n = a_n b_n$, where $a_n$ is the number of batches and $b_n$ is the batch size. For $k = 0, \ldots, a_n - 1$, define $\bar{Y}_k = b_n^{-1} \sum_{t=1}^{b_n} Y_{kb_n+t}$. Then $\bar{Y}_k$ is the mean vector for batch $k$ and the multivariate batch means estimator of $\Sigma$ is

$$\Sigma_n = \frac{b_n}{a_n - 1} \sum_{k=0}^{a_n-1} \left(\bar{Y}_k - \theta_n\right)\left(\bar{Y}_k - \theta_n\right)^{\mathrm{T}}.$$

For the multivariate batch means estimator, $q$ in (6) is $a_n - p$. In addition, $\Sigma_n$ is singular if $a_n < p$; thus $n_{\text{pos}}$ is the smallest $n$ such that $a_n > p$.

The univariate batch means estimator has been well studied (Damerdji, 1994; Flegal & Jones, 2010; Glynn & Iglehart, 1990; Glynn & Whitt, 1991; Jones et al., 2006). Glynn & Whitt (1991) showed that the batch means estimator cannot be consistent for fixed batch size $b_n$. Damerdji (1994, 1995), Jones et al. (2006), and Flegal & Jones (2010) established strong consistency and mean square consistency for when both the batch size and the number of batches increase with $n$. The multivariate extension was first introduced by Chen & Seila (1987). For steady-state simulation, Charnes (1995) and Muñoz & Glynn (2001) studied confidence regions for $\theta$ based on the multivariate batch means estimator; however, its asymptotic properties remain unexplored. We present conditions for strong consistency of $\Sigma_n$ in estimating $\Sigma$ for Markov chain Monte Carlo, but our results hold for more general processes. Our main assumption on the process is that of a strong invariance principle.

*Condition* 1. Let $\{B(t), t \geqslant 0\}$ be a $p$-dimensional standard Brownian motion. There exists a $p \times p$ lower triangular matrix $L$, a nonnegative increasing function $\gamma$ on the positive integers, a finite random variable $D$, and a sufficiently rich probability space such that, with probability 1, as $n \to \infty$,

$$\|n(\theta_n - \theta) - LB(n)\| < D\gamma(n). \tag{11}$$

*Condition* 2. The batch size $b_n$ satisfies the following conditions:

(a) the batch size $b_n$ is an integer sequence such that $b_n \to \infty$, and $n/b_n \to \infty$ as $n \to \infty$, where $b_n$ and $n/b_n$ are increasing;
(b) there exists a constant $c \geqslant 1$ such that $\sum_n \left(b_n n^{-1}\right)^c < \infty$.

THEOREM 2. *Let $g$ be such that $E_F(\|g\|^{2+\delta}) < \infty$ for some $\delta > 0$. Let $\{X_t\}$ be an $F$-invariant polynomially ergodic Markov chain of order $m > (1+\epsilon_1)(1+2/\delta)$ for some $\epsilon_1 > 0$. Then (11) holds with $\gamma(n) = n^{1/2-\lambda}$ for some $\lambda > 0$. If Condition 2 holds and $b_n^{-1/2}(\log n)^{1/2} n^{1/2-\lambda} \to 0$ as $n \to \infty$, then $\Sigma_n \to \Sigma$, with probability 1, as $n \to \infty$.*

*Remark* 4. Theorem 2 holds more generally outside the context of Markov chains for processes that satisfy Condition 1. This includes independent processes (Berkes & Philipp, 1979; Einmahl, 1989; Zaitsev, 1998), martingale sequences (Eberlein, 1986), renewal processes (Horvath, 1984), and $\phi$-mixing and strongly mixing processes (Kuelbs & Philipp, 1980; Dehling & Philipp, 1982). The general statement of the theorem is provided in the Supplementary Material.

*Remark* 5. Using Theorem 4 from Kuelbs & Philipp (1980), Vats et al. (2018) established Condition 1 with $\gamma(n) = n^{1/2-\lambda}$, for some $\lambda > 0$ for polynomially ergodic Markov chains. We use their result directly. Kuelbs & Philipp (1980) show that $\lambda$ only depends on $p$, $\epsilon$, and $\delta$; however, the exact relationship remains an open problem. For slow-mixing processes $\lambda$ is closer to 0, while for fast-mixing processes $\lambda$ is closer to $1/2$ (Damerdji, 1991, 1994).

*Remark* 6. It is natural to consider $b_n = \lfloor n^\nu \rfloor$ for $0 < \nu < 1$. Here $\nu > 1 - 2\lambda$ to ensure $b_n^{-1/2}(\log n)^{1/2} n^{1/2-\lambda} \to 0$ as $n \to \infty$. Thus smaller batch sizes suffice for fast-mixing processes, and slow-mixing processes require larger batch sizes. This reinforces our intuition that higher correlation calls for larger batch sizes.

*Remark* 7. For $p = 1$, Jones et al. (2006) proved strong consistency of the batch means estimator under the stronger assumption of geometric ergodicity and a one-step minorization. Thus, in Theorem 2, while extending the result of strong consistency to $p \geqslant 1$ we also weaken the conditions for the univariate case.

## 5. EXAMPLES

### 5.1. *Simulation set-up*

In the examples below, we present a Markov chain with invariant distribution $F$; we specify $g$, and are interested in estimating $\mathrm{E}_F(g)$. We consider the finite-sample performance of the relative standard deviation fixed-volume sequential stopping rules based on 1000 independent replications and compare them to the relative standard deviation fixed-width sequential stopping rules; see Flegal & Gong (2015) and the Supplementary Material. In each case we construct 90% confidence regions for various choices of $\epsilon$ and specify our choice of $n^*$ and $b_n$. The sequential stopping rules are checked at 10% increments of the current Monte Carlo sample size.

### 5.2. *Bayesian logistic regression*

We continue the Bayesian logistic regression example of § 1. Recall that a random walk Metropolis–Hastings algorithm was implemented to sample from the intractable posterior. As a consequence of Theorem 3 below and the fact that $F$ has a moment generating function, the conditions of Theorems 1 and 2 hold.

THEOREM 3. *The random walk-based Metropolis–Hastings algorithm with invariant distribution given by the posterior from (5) is geometrically ergodic.*

Table 2. *Termination iteration, estimated effective sample size at termination, and coverage probability at termination at* 90% *nominal level. Standard errors are in parentheses*

| | | Multivariate | Bonferroni-corrected univariate | Uncorrected univariate |
|---|---|---|---|---|
| Termination iteration | $\epsilon = 0.05$ | 133 005 (196) | 201 497 (391) | 100 445 (213) |
| | $\epsilon = 0.02$ | 844 082 (1158) | 1 262 194 (1880) | 629 898 (1036) |
| | $\epsilon = 0.01$ | 3 309 526 (1837) | 5 046 449 (7626) | 2 510 673 (3150) |
| Effective sample size | $\epsilon = 0.05$ | 7712 (9) | 9270 (13) | 4643 (7) |
| | $\epsilon = 0.02$ | 47 862 (51) | 57 341 (65) | 28 768 (36) |
| | $\epsilon = 0.01$ | 186 103 (110) | 228 448 (271) | 113 831 (116) |
| Coverage probabilities | $\epsilon = 0.05$ | 0.889 (0.0099) | 0.909 (0.0091) | 0.569 (0.0157) |
| | $\epsilon = 0.02$ | 0.896 (0.0097) | 0.912 (0.0090) | 0.606 (0.0155) |
| | $\epsilon = 0.01$ | 0.892 (0.0098) | 0.895 (0.0097) | 0.606 (0.0155) |

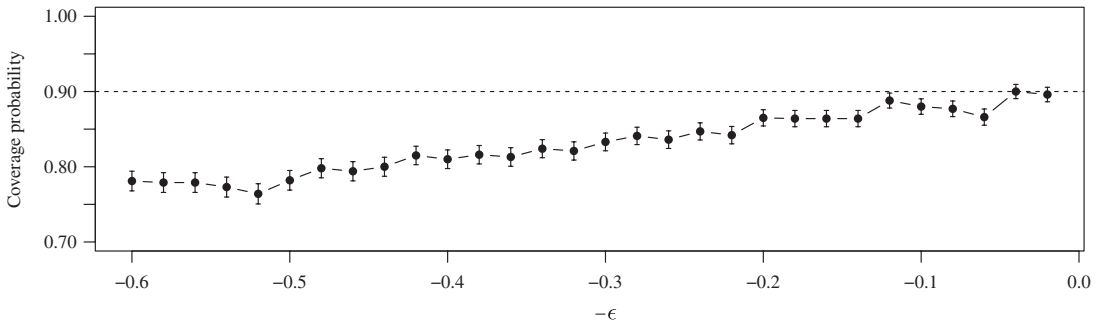


Fig. 2. Plot of coverage probability with confidence bands as $\epsilon$ decreases at 90% nominal rate.

Motivated by the autocorrelation plot in Fig. 1, $b_n$ was set to $\lfloor n^{1/2} \rfloor$ and $n^* = 1000$. For calculating coverage probabilities, we declare the truth as the posterior mean from an independent simulation of length $10^9$. The results are presented in Table 2. As before, the univariate uncorrected method has poor coverage probabilities. For $\epsilon = 0.02$ and 0.01, the coverage probabilities for both the multivariate and Bonferroni-corrected univariate regions are at 90%. However, multivariate termination is earlier.

Recall from Theorem 1 that as $\epsilon$ decreases to zero, the coverage probability of confidence regions created at termination converges to the nominal level. This is demonstrated in Fig. 2, where we present the coverage probability over 1000 replications as $-\epsilon$ increases.

### 5.3. *Vector autoregressive process*

Consider the vector autoregressive process of order 1. For $t = 1, 2, \ldots,$

$$Y_t = \Phi Y_{t-1} + \epsilon_t,$$

where $Y_t \in \mathbb{R}^p$, $\Phi$ is a $p \times p$ matrix, $\epsilon_t \overset{\text{iid}}{\sim} N_p(0, \Omega)$, and $\Omega$ is a $p \times p$ positive-definite matrix. We set $Y_0$ to be the zero vector. The matrix $\Phi$ determines the nature of the correlation in the process. This Markov chain has invariant distribution $F = N_p(0, V)$, where $\text{vec}(V) = (I_{p^2} - \Phi \otimes \Phi)^{-1}\text{vec}(\Omega)$

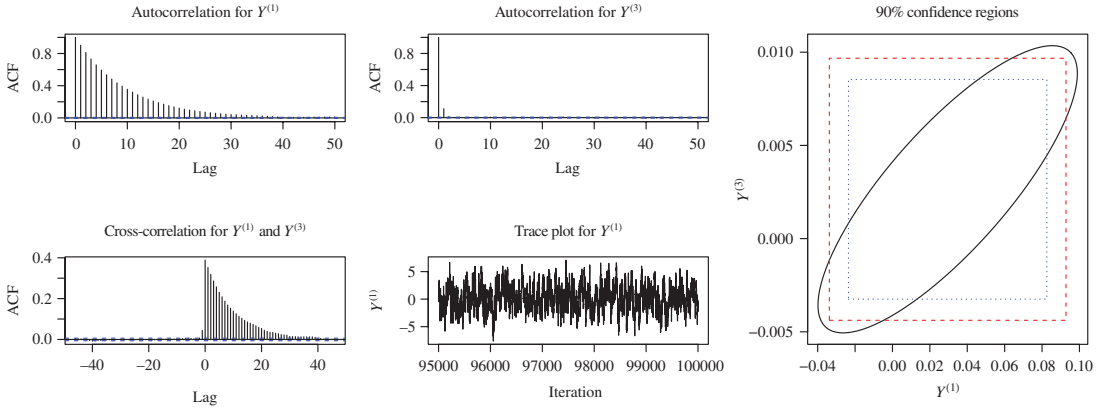


Fig. 3. Autocorrelation and cross-correlation plots for $Y^{(1)}$ and $Y^{(3)}$, trace plot for $Y^{(1)}$, and 90% confidence regions for $Y^{(1)}$ and $Y^{(3)}$. The solid ellipse is constructed using multivariate batch means, the dotted box using univariate batch means uncorrected, and the dashed box using Bonferroni-corrected univariate batch means.

with $\otimes$ denoting the Kronecker product, and is geometrically ergodic when the spectral radius of $\Phi$ is less than 1 (Tjøstheim, 1990).

Consider the goal of estimating the mean of $F$, $\mathrm{E}_F(Y) = 0$ with $\bar{Y}_n$. Then

$$\Sigma = (I_p - \Phi)^{-1} V + V (I_p - \Phi)^{-1} - V.$$

Let $p = 5$, $\Phi = \mathrm{diag}(0.9, 0.5, 0.1, 0.1, 0.1)$, and $\Omega$ be the AR(1) covariance matrix with autocorrelation 0.9. Since the first eigenvalue of $\Phi$ is large, the first component mixes slowest. We sample the process for $10^5$ iterations and in Fig. 3 present the autocorrelation plot for $Y^{(1)}$ and $Y^{(3)}$ and the cross-correlation plot between $Y^{(1)}$ and $Y^{(3)}$ in addition to the trace plot for $Y^{(1)}$. Notice that $Y^{(1)}$ exhibits higher autocorrelation than $Y^{(3)}$ and there is significant cross-correlation between $Y^{(1)}$ and $Y^{(3)}$. Figure 3 also displays joint confidence regions for $Y^{(1)}$ and $Y^{(3)}$. Recall that the true mean is $(0, 0)$ and is present in all three regions, but the ellipse produced by multivariate batch means has significantly smaller volume than the univariate batch means boxes. The orientation of the ellipse is determined by the cross-correlations witnessed in Fig. 3.

We set $n^* = 1000$, $b_n = \lfloor n^{1/2} \rfloor$ and $\epsilon$ in $\{0.05, 0.02, 0.01\}$, and at termination calculate the coverage probabilities and effective sample size. Results are presented in Table 3. As $\epsilon$ decreases, termination time increases and coverage probabilities tend to the 90% nominal level for each method. Also, the uncorrected methods produce confidence regions with undesirable coverage probabilities and thus are not of interest. Consider $\epsilon = 0.02$ in Table 3. Termination for multivariate batch means is at $8.8 \times 10^4$ iterations compared to $1.1 \times 10^6$ for Bonferroni-corrected univariate batch means, with effective sample sizes of $4.8 \times 10^4$ and $5.7 \times 10^4$, respectively. This is because the leading component $Y^{(1)}$ mixes much slower than the other components and defines the behaviour of the univariate effective sample size.

The true univariate and multivariate effective sample sizes can be obtained in closed form in this case since both $\Sigma$ and $V$ are known. For a Monte Carlo sample size of $10^5$, the univariate effective sample sizes for the five components are 5263, 33 333, 81 818, 81 818, and 81 818, while the true ESS is 55 188. In the absence of cross-correlations, ESS is the geometric mean of the univariate $\mathrm{ESS}_i$, which is 39 494 here. Thus, by accounting for the cross-correlation structure of $V$ and $\Sigma$, we see an increase in effective sample size.

Table 3. *Termination iteration, estimated* ESS *at termination, and coverage probability at termination at* 90% *nominal level. Standard errors are in parentheses*

| | | Multivariate | Bonferroni univariate | Uncorrected univariate |
|---|---|---|---|---|
| Termination iteration | $\epsilon = 0.05$ | 14 574 (27) | 169 890 (393) | 83 910 (222) |
| | $\epsilon = 0.02$ | 87 682 (118) | 1 071 449 (1733) | 533 377 (1015) |
| | $\epsilon = 0.01$ | 343 775 (469) | 4 317 599 (5358) | 2 149 042 (3412) |
| Effective sample size | $\epsilon = 0.05$ | 8170 (11) | 9298 (13) | 4658 (7) |
| | $\epsilon = 0.02$ | 48 659 (50) | 57 392 (68) | 28 756 (37) |
| | $\epsilon = 0.01$ | 190 198 (208) | 228 772 (223) | 114 553 (137) |
| Coverage probabilities | $\epsilon = 0.05$ | 0.911 (0.0090) | 0.940 (0.0075) | 0.770 (0.0133) |
| | $\epsilon = 0.02$ | 0.894 (0.0097) | 0.950 (0.0069) | 0.769 (0.0133) |
| | $\epsilon = 0.01$ | 0.909 (0.0091) | 0.945 (0.0072) | 0.779 (0.0131) |

### 5.4. *Bayesian lasso*

Let $y$ be a $K \times 1$ response vector and $X$ be a $K \times r$ matrix of predictors. We consider the following Bayesian lasso formulation of Park & Casella (2008):

$$\begin{aligned} y \mid \beta, \sigma^2, \tau^2 &\sim N_K(X\beta, \sigma^2 I_n), \\ \beta \mid \sigma^2, \tau^2 &\sim N_r(0, \sigma^2 D_\tau), \quad D_\tau = \operatorname{diag}(\tau_1^2, \ldots, \tau_r^2), \\ \sigma^2 &\sim \text{IG}(\alpha, \xi), \\ \tau_j^2 &\overset{\text{iid}}{\sim} \text{Exp}\left(\frac{\tilde{\lambda}^2}{2}\right) \quad (j = 1, \ldots, r), \end{aligned}$$

where $\tilde{\lambda}, \alpha$, and $\xi$ are fixed and the $\text{IG}(a, b)$ distribution has density proportional to $x^{-a-1}\exp(-b/x)$. We use the Gibbs sampler described in Khare & Hobert (2013) to draw approximate samples from the posterior. Khare & Hobert (2013) showed that for $K \geqslant 3$, the Gibbs sampler is geometrically ergodic for arbitrary $r$, $X$, and $\tilde{\lambda}$.

We fit this model using the cookie dough dataset of Osborne et al. (1984). The data were collected to test the feasibility of near-infrared spectroscopy for measuring the composition of biscuit dough pieces. There are 72 observations. The response variable is the amount of dry flour content measured and the predictor variables are 25 measurements of spectral data spaced equally between 1100 to 2498 nanometres. Since we are interested in estimating the posterior mean for $(\beta, \tau^2, \sigma^2)$, $p = 51$. The data are available in the `R` package `ppls` and the Gibbs sampler is implemented by the function `blasso` in the `R` package `monomvn`. The starting values are set to be the least square estimates. The truth was declared by averaging posterior means from 1000 independent runs each of length $10^6$. We set $n^* = 2 \times 10^4$ and $b_n = \lfloor n^{1/3} \rfloor$.

In Table 4 we present termination results from 1000 replications. With $p = 51$, the uncorrected univariate regions produce confidence regions with low coverage probabilities. The Bonferroni-corrected univariate and multivariate methods provide competitive coverage probabilities. However, multivariate termination is significantly earlier than univariate termination over all values of $\epsilon$. For $\epsilon = 0.05$ and 0.02 we observe zero standard error for termination using multivariate batch means since termination is achieved at the same 10% increment over all 1000 replications. Thus the variability in those estimates is less than 10% of the size of the estimate.

Table 4. *Termination iteration, estimated* ESS *at termination, and coverage probability at termination at* 90% *nominal level. Standard errors are in parentheses*

| | | Multivariate | Bonferroni univariate | Uncorrected univariate |
|---|---|---|---|---|
| Termination iteration | $\epsilon = 0.05$ | 20 000 (0) | 69 264 (76) | 20 026 (7) |
| | $\epsilon = 0.02$ | 69 045 (0) | 445 754 (664) | 122 932 (103) |
| | $\epsilon = 0.01$ | 271 088 (393) | 1 765 008 (431) | 508 445 (332) |
| Effective sample size | $\epsilon = 0.05$ | 15631 (4) | 16143 (15) | 4778 (6) |
| | $\epsilon = 0.02$ | 52 739 (8) | 101 205 (122) | 28 358 (24) |
| | $\epsilon = 0.01$ | 204 801 (283) | 395 480 (163) | 115 108 (74) |
| Coverage probabilities | $\epsilon = 0.05$ | 0.898 (0.0096) | 0.896 (0.0097) | 0.010 (0.0031) |
| | $\epsilon = 0.02$ | 0.892 (0.0098) | 0.905 (0.0093) | 0.009 (0.0030) |
| | $\epsilon = 0.01$ | 0.898 (0.0096) | 0.929 (0.0081) | 0.009 (0.0030) |

### 5.5. *Bayesian dynamic spatial-temporal model*

Gelfand et al. (2005) propose a Bayesian hierarchical model for modelling univariate and multivariate dynamic spatial data, viewing time as discrete and space as continuous. The methods in their paper have been implemented in the R package `spBayes`. We present a simpler version of the dynamic model as described by Finley et al. (2015).

Let $s = 1, \ldots, N_s$ be location sites, $t = 1, \ldots, N_t$ be time-points, and the observed measurement at location $s$ and time $t$ be denoted by $y_t(s)$. In addition, let $x_t(s)$ be the $r \times 1$ vector of predictors, observed at location $s$ and time $t$, and let $\beta_t$ be the $r \times 1$ vector of coefficients. Let $\mathrm{GP}\{0, \sigma_t^2 \rho(\cdot; \phi_t)\}$ denote a spatial Gaussian process with covariance function $\sigma_t^2 \rho(\cdot; \phi_t)$. Here $\sigma_t^2$ is the spatial variance component and $\rho(\cdot, \phi_t)$ is a correlation function with exponential decay. For $t = 1, \ldots, N_t$,

$$y_t(s) = x_t(s)^{\mathrm{T}} \beta_t + u_t(s) + \epsilon_t(s), \quad \epsilon_t(s) \overset{\text{ind}}{\sim} N(0, \tau_t^2); \tag{12}$$

$$\beta_t = \beta_{t-1} + \eta_t, \quad \eta_t \overset{\text{iid}}{\sim} N(0, \Sigma_\eta);$$

$$u_t(s) = u_{t-1}(s) + w_t(s), \quad w_t(s) \overset{\text{ind}}{\sim} \mathrm{GP}\{0, \sigma_t^2 \rho(\cdot; \phi_t)\}. \tag{13}$$

Equation (12) is referred to as the measurement equation and $\epsilon_t(s)$ denotes the measurement error, assumed to be independent of location and time. Equation (13) contains the transition equations which emulate the Markovian nature of dependence in time. To complete the Bayesian hierarchy, the following priors are assumed:

$$\begin{aligned} \beta_0 &\sim N(m_0, C_0), \quad u_0(s) \equiv 0; \\ \tau_t^2 &\sim \mathrm{IG}(a_\tau, b_\tau), \quad \sigma_t^2 \sim \mathrm{IG}(a_s, b_s); \\ \Sigma_\eta &\sim \mathrm{IW}(a_\eta, B_\eta), \quad \phi_t \sim \mathrm{Un}(a_\phi, b_\phi), \end{aligned}$$

where the IW distribution has density proportional to $\det(\Sigma_\eta)^{-(a_\eta+q+1)/2} \exp\{-\mathrm{tr}(B_\eta \Sigma_\eta^{-1})/2\}$ and $\mathrm{IG}(a, b)$ has density proportional to $x^{-a-1} \exp(-b/x)$. We fit the model to the `NETemp` dataset in the `spBayes` package. This dataset contains monthly temperature measurements from 356 weather stations on the east coast of the USA collected from January 2000 to December 2010. The elevation of the weather stations is also available as a covariate. We choose a subset of the

Table 5. *Termination iteration, estimated* ESS *at termination, and coverage probability at termination at* 90% *nominal level. Standard errors are in parentheses*

| | | Multivariate | Bonferroni univariate | Uncorrected univariate |
|---|---|---|---|---|
| Termination iteration | $\epsilon = 0.10$ | 50 000 (0) | 1 200 849 (2 8315) | 311 856 (491) |
| | $\epsilon = 0.05$ | 50 030 (12) | - | 1 716 689 (2178) |
| | $\epsilon = 0.02$ | 132 748 (174) | - | - |
| Effective sample size | $\epsilon = 0.10$ | 55 170 (20) | 3184 (75) | 1130 (1) |
| | $\epsilon = 0.05$ | 55 190 (20) | - | 4525 (4) |
| | $\epsilon = 0.02$ | 105 166 (97) | - | - |
| Coverage probabilities | $\epsilon = 0.10$ | 0.882 (0.0102) | 0.625 (0.0153) | 0.007 (0.0026) |
| | $\epsilon = 0.05$ | 0.881 (0.0102) | - | 0.016 (0.0040) |
| | $\epsilon = 0.02$ | 0.864 (0.0108) | - | - |

data with 10 weather stations for the year 2000, and fit the model with an intercept. The resulting posterior has $p = 185$ components.

A componentwise Metropolis–Hastings sampler (Johnson et al., 2013; Jones et al., 2014) is described in Gelfand et al. (2005) and implemented in the `spDynLM` function in the `spBayes` package. Default hyperparameter settings and starting values were used. The rate of convergence for this sampler has not been studied; thus we do not know if the conditions of our theoretical results are satisfied. Our goal is to estimate the posterior mean of $\{\beta_t, u_t(s), \sigma_t^2, \Sigma_\eta, \tau_t^2, \phi_t\}$. The truth was declared by averaging over 1000 independent runs of length $2 \times 10^6$ samples. We set $b_n = \lfloor n^{1/2} \rfloor$ and $n^* = 5 \times 10^4$ so that $a_n > p$ to ensure positive definitiveness of $\Sigma_n$.

Due to the Markovian transition equations in (13), $\beta_t$ and $u_t$ exhibit a significant covariance structure in the posterior distribution, leading to thin confidence ellipsoids. Since $p = 185$ here, for smaller values of $\epsilon$ it was not possible to store the Markov chain Monte Carlo output in memory on a 8 gigabyte machine using Bonferroni-corrected univariate batch means. As a result, in Table 5, the univariate methods could not be implemented for smaller $\epsilon$ values. For $\epsilon = 0.10$, multivariate termination was at $n^* = 5 \times 10^4$ for every replication, resulting in an observed coverage probability of 88%. Both univariate methods have far lower coverage probabilities.

## 6. DISCUSSION

Multivariate analysis of Markov chain Monte Carlo output data has received little attention. Seila (1982) and Chen & Seila (1987) built a framework for multivariate analysis for a Markov chain using regenerative simulation. Since establishing regenerative properties for a Markov chain requires a separate analysis for every problem and will not work well in componentwise Metropolis–Hastings samplers, the application of their work is limited. Paul et al. (2012) used a specific multivariate Markov chain central limit theorem for their Markov chain Monte Carlo convergence diagnostic. More recently, Vats et al. (2018) showed strong consistency of the multivariate spectral variance estimators of $\Sigma$, which could potentially substitute for the multivariate batch means estimator in applications to termination rules. However, outside of toy problems where $p$ is small, the multivariate spectral variance estimator is computationally demanding compared to the multivariate batch means estimator. The multivariate batch means and spectral variance estimators along with multivariate ESS have been implemented in the `R` package `mcmcse` (Flegal et al., 2017).

There are two aspects of the proposed methodology that will benefit from further research. First, the rate of convergence of the Markov chain affects the choice of $b_n$ through the $\lambda$ in the strong invariance principle. Aside from Damerdji (1995) and Flegal & Jones (2010), little work has been done in optimal batch size selection for batch means estimators. This area warrants further research in both asymptotic and finite-sample optimal batch size selection. In the Supplementary Material we study the effect of different batch sizes on simulation termination using the relative standard deviation fixed-volume sequential stopping rule. We notice that a decrease in the tolerance level $\epsilon$ decreases the sensitivity of termination to the choice of $b_n$. We have found that using a large batch size such as $b_n = \lfloor n^{1/2} \rfloor$ often works well. Second, when using the multivariate batch means estimator, a truly large $p$ requires a large Monte Carlo sample size to ensure a positive-definite estimate of $\Sigma$. It would be interesting to investigate the use of dimension reduction techniques or high-dimensional asymptotics to address this problem.

## Acknowledgement

Flegal's work was partially supported by the National Science Foundation. Jones was partially supported by the National Science Foundation and the National Institutes for Health.

## Supplementary material

Supplementary material available at *Biometrika* online includes proofs of the theorems, a description of the univariate termination rules, an analysis of the sensitivity of the batch means estimator to the batch size, and a comparison of batch means and spectral variance estimators.

## References

Atchadé, Y. F. (2011). Kernel estimators of asymptotic variance for adaptive Markov chain Monte Carlo. *Ann. Statist.* **39**, 990–1011.

Atchadé, Y. F. (2016). Markov chain Monte Carlo confidence intervals. *Bernoulli* **22**, 1808–38.

Atkinson, Q. D., Gray, R. D. & Drummond, A. J. (2008). mtDNA variation predicts population size in humans and reveals a major southern Asian chapter in human prehistory. *Molec. Biol. Evol.* **25**, 468–74.

Bednorz, W. & Łatuszyński, K. (2007). A few remarks on "Fixed-width output analysis for Markov chain Monte Carlo" by Jones et al. *J. Am. Statist. Assoc.* **102**, 1485–6.

Berkes, I. & Philipp, W. (1979). Approximation theorems for independent and weakly dependent random vectors. *Ann. Prob.* **7**, 29–54.

Charnes, J. M. (1995). Analyzing multivariate output. In *Proc. 27th Conf. on Winter Simulation*. IEEE Computer Society.

Chen, D.-F. R. & Seila, A. F. (1987). Multivariate inference in stationary simulation using batch means. In *Proc. 19th Conf. on Winter simulation*. Association for Computing Machinery.

Dai, N. & Jones, G. L. (2017). Multivariate initial sequence estimators in Markov chain Monte Carlo. *J. Mult. Anal.* **159**, 184–99.

Damerdji, H. (1991). Strong consistency and other properties of the spectral variance estimator. *Manag. Sci.* **37**, 1424–40.

Damerdji, H. (1994). Strong consistency of the variance estimator in steady-state simulation output analysis. *Math. Oper. Res.* **19**, 494–512.

Damerdji, H. (1995). Mean-square consistency of the variance estimator in steady-state simulation output analysis. *Oper. Res.* **43**, 282–91.

Dehling, H. & Philipp, W. (1982). Almost sure invariance principles for weakly dependent vector-valued random variables. *Ann. Prob.* **10**, 689–701.

Doss, H. & Hobert, J. P. (2010). Estimation of Bayes factors in a class of hierarchical random effects models using a geometrically ergodic MCMC algorithm. *J. Comp. Graph. Statist.* **19**, 295–312.

Drummond, A. J., Ho, S. Y., Phillips, M. J. & Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. *PLoS Biology* **4**, 699.

Eberlein, E. (1986). On strong invariance principles under dependence assumptions. *Ann. Prob.* **14**, 260–70.

EINMAHL, U. (1989). Extensions of results of Komlós, Major, and Tusnády to the multivariate case. *J. Mult. Anal.* **28**, 20–68.

FINLEY, A., BANERJEE, S. & GELFAND, A. (2015). spBayes for large univariate and multivariate point-referenced spatio-temporal data models. *J. Statist. Software* **63**, 1–28.

FLEGAL, J. M. & GONG, L. (2015). Relative fixed-width stopping rules for Markov chain Monte Carlo simulations. *Statist. Sinica* **25**, 655–76.

FLEGAL, J. M., HARAN, M. & JONES, G. L. (2008). Markov chain Monte Carlo: Can we trust the third significant figure? *Statist. Sci.* **23**, 250–60.

FLEGAL, J. M., HUGHES, J., VATS, D. & DAI, N. (2017). *mcmcse: Monte Carlo Standard Errors for MCMC*. Riverside, CA and Minneapolis, MN. R package version 1.3-2.

FLEGAL, J. M. & JONES, G. L. (2010). Batch means and spectral variance estimators in Markov chain Monte Carlo. *Ann. Statist.* **38**, 1034–70.

GELFAND, A. E., BANERJEE, S. & GAMERMAN, D. (2005). Spatial process modelling for univariate and multivariate dynamic spatial data. *Environmetrics* **16**, 465–79.

GELMAN, A. & RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statist. Sci.* **7**, 457–72.

GEYER, C. J. (2011). Introduction to Markov chain Monte Carlo. In *Handbook of Markov Chain Monte Carlo*, S. Brooks, A. Gelman, X.-L. Meng & G. L. Jones, eds. Boca Raton: Chapman & Hall, pp. 3–48.

GEYER, C. J. & JOHNSON, L. T. (2015). *mcmc: Markov Chain Monte Carlo*. Minneapolis, MN. R package version 0.9-4.

GIORDANO, R., BRODERICK, T. & JORDAN, M. (2015). Linear response methods for accurate covariance estimates from mean field variational Bayes. In *Proc. 28th Int. Conf. Neural Information Processing Systems*, vol. 1. Cambridge, Massachusetts: MIT Press, pp. 1441–9.

GLYNN, P. W. & IGLEHART, D. L. (1990). Simulation output analysis using standardized time series. *Math. Oper. Res.* **15**, 1–16.

GLYNN, P. W. & WHITT, W. (1991). Estimating the asymptotic variance with batch means. *Oper. Res. Lett.* **10**, 431–5.

GLYNN, P. W. & WHITT, W. (1992). The asymptotic validity of sequential stopping rules for stochastic simulations. *Ann. Appl. Prob.* **2**, 180–98.

GONG, L. & FLEGAL, J. M. (2016). A practical sequential stopping rule for high-dimensional Markov chain Monte Carlo. *J .Comp. Graph. Statist.* **25**, 684–700.

HOBERT, J. P. & GEYER, C. J. (1998). Geometric ergodicity of Gibbs and block Gibbs samplers for a hierarchical random effects model. *J. Mult. Anal.* **67**, 414–30.

HORVATH, L. (1984). Strong approximation of extended renewal processes. *Ann. Prob.* **12**, 1149–66.

JARNER, S. F. & HANSEN, E. (2000). Geometric ergodicity of Metropolis algorithms. *Stoch. Proces. Appl.* **85**, 341–61.

JARNER, S. F. & ROBERTS, G. O. (2002). Polynomial convergence rates of Markov chains. *Ann. Appl. Prob.* **12**, 224–47.

JOHNSON, A. A. & JONES, G. L. (2015). Geometric ergodicity of random scan Gibbs samplers for hierarchical one-way random effects models. *J. Mult. Anal.* **140**, 325–42.

JOHNSON, A. A., JONES, G. L. & NEATH, R. C. (2013). Component-wise Markov chain Monte Carlo. *Statist. Sci.* **28**, 360–75.

JONES, G. L. (2004). On the Markov chain central limit theorem. *Prob. Surveys* **1**, 299–320.

JONES, G. L., HARAN, M., CAFFO, B. S. & NEATH, R. (2006). Fixed-width output analysis for Markov chain Monte Carlo. *J. Am. Statist. Assoc.* **101**, 1537–47.

JONES, G. L. & HOBERT, J. P. (2004). Sufficient burn-in for Gibbs samplers for a hierarchical random effects model. *Ann. Statist.* **32**, 784–817.

JONES, G. L., ROBERTS, G. O. & ROSENTHAL, J. S. (2014). Convergence of conditional Metropolis-Hastings samplers. *Adv. Appl. Prob.* **46**, 422–45.

KASS, R. E., CARLIN, B. P., GELMAN, A. & NEAL, R. M. (1998). Markov chain Monte Carlo in practice: A roundtable discussion. *Am. Statistician* **52**, 93–100.

KHARE, K. & HOBERT, J. P. (2013). Geometric ergodicity of the Bayesian lasso. *Electron. J. Statist.* **7**, 2150–63.

KRUSCHKE, J. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. London: Academic Press.

KUELBS, J. & PHILIPP, W. (1980). Almost sure invariance principles for partial sums of mixing B-valued random variables. *Ann. Prob.* **8**, 1003–36.

LIU, J. S. (2008). *Monte Carlo Strategies in Scientific Computing*. New York: Springer.

MARCHEV, D. & HOBERT, J. P. (2004). Geometric ergodicity of van Dyk and Meng's algorithm for the multivariate Student's *t* model. *J. Am. Statist. Assoc.* **99**, 228–38.

MUÑOZ, D. F. & GLYNN, P. W. (2001). Multivariate standardized time series for steady-state simulation output analysis. *Oper. Res.* **49**, 413–22.

OSBORNE, B. G., FEARN, T., MILLER, A. R. & DOUGLAS, S. (1984). Application of near infrared reflectance spectroscopy to the compositional analysis of biscuits and biscuit doughs. *J . Sci. Food Agric.* **35**, 99–105.

PARK, T. & CASELLA, G. (2008). The Bayesian lasso. *J. Am. Statist. Assoc.* **103**, 681–6.

PAUL, R., MACEACHERN, S. N. & BERLINER, L. M. (2012). Assessing convergence and mixing of MCMC implementations via stratification. *J. Comp. Graph. Statist.* **21**, 693–712.

R Development Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, http://www.R-project.org.

Robert, C. P. & Casella, G. (2013). *Monte Carlo Statistical Methods*. New York: Springer.

Roberts, G. O., Gelman, A. & Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Prob.* **7**, 110–20.

Roberts, G. O. & Polson, N. G. (1994). On the geometric convergence of the Gibbs sampler. *J. R. Statist. Soc.* B **56**, 377–84.

Seila, A. F. (1982). Multivariate estimation in regenerative simulation. *Oper. Res. Lett.* **1**, 153–6.

SenGupta, A. (1987). Tests for standardized generalized variances of multivariate normal populations of possibly different dimensions. *J. Mult. Anal.* **23**, 209–19.

Tan, A. & Hobert, J. P. (2012). Block Gibbs sampling for Bayesian random effects models with improper priors: Convergence and regeneration. *J. Comp. Graph. Statist.* **18**, 861–78.

Tan, A., Jones, G. L. & Hobert, J. P. (2013). On the geometric ergodicity of two-variable Gibbs samplers. In *Advances in Modern Statistical Theory and Applications: A Festschrift in Honor of Morris L. Eaton*, G. L. Jones & X. Shen, eds. Beachwood, Ohio: Institute of Mathematical Statistics, pp. 25–42.

Tjøstheim, D. (1990). Non-linear time series and Markov chains. *Adv. Appl. Prob.* **22**, 587–611.

Vats, D. (2017). Geometric ergodicity of Gibbs samplers in Bayesian penalized regression models. *Electron. J. Statist.* **11**, 4033–64.

Vats, D., Flegal, J. M. & Jones, G. L. (2018). Strong consistency of multivariate spectral variance estimators in Markov chain Monte Carlo. *Bernoulli* **24**, 1860–909.

Wilks, S. S. (1932). Certain generalizations in the analysis of variance. *Biometrika* **24**, 471–94.

Zaitsev, A. Y. (1998). Multidimensional version of the results of Komlós and Tusnády for vectors with finite exponential moments. *ESAIM: Prob. Statist.* **2**, 41–108.

[*Received on* 3 *February* 2018. *Editorial decision on* 26 *September* 2018]