Bernoulli **6**(2), 2000, 215–242

Estimating the second largest eigenvalue of a Markov transition matrix

STEVEN T. GARREN[1] and RICHARD L. SMITH[2]

[1]*104 Halsey Hall, Department of Statistics, University of Virginia, Charlottesville VA 22903, USA. E-mail: stg3a@virginia.edu*
[2]*117 New West, Department of Statistics, University of North Carolina, Chapel Hill NC 27599-3260, USA. E-mail: rls@email.unc.edu*

The number of iterations required to estimate accurately the stationary distribution of a Markov chain is determined by a preliminary sample to estimate the convergence rate, which is related to the second largest eigenvalue of the transition operator. The estimator of the second largest eigenvalue, along with those of two nuisance parameters, can be shown to converge to their true values in probability, and a form of the central limit theorem is proved. Explicit expressions for the bias and variance of the asymptotic distribution of this estimator are derived. A theoretical standard is derived against which other estimators of the second largest eigenvalue may be judged. An application is given involving the use of the Gibbs sampler to calculate a posterior distribution.

Keywords: Gibbs sampler; Hilbert–Schmidt operator; Markov chain Monte Carlo; Metropolis–Hastings algorithm

1. Introduction

Markov chain Monte Carlo algorithms have become widely used in statistical inference, especially in Bayesian analysis (Gelfand and Smith 1990; Smith and Roberts 1993) but also for solving non-Bayesian problems involving latent variables (Geyer and Thompson 1992). Particular attention has been paid to the Gibbs sampler (Geman and Geman 1984), though a more general approach is through Hastings's (1970) generalization of the Metropolis *et al*. (1953) procedure. A related idea is data augmentation (Tanner and Wong 1987). In recent research on these methods, a recurrent theme is the use of diagnostic tests to assess the rate of convergence or, more broadly, when to stop sampling. Contrasting approaches are presented by Gelman and Rubin (1992) and Geyer (1992), and a comprehensive review of convergence issues is in the paper by Tierney (1994).

Our aim in this paper is to introduce a new method of assessing the rate of convergence of Markov chain samplers, based on estimation of the second largest eigenvalue of the Markov transition operator. The largest eigenvalue is always one, but the second largest, when it is well defined and strictly less than one in modulus, determines the rate of convergence. A feature of our method is that it is based directly on data generated by the sampler, in contrast with methods based on analytical bounds (see, for example, Lawler and

Sokal 1988; Sinclair and Jerrum 1989; Diaconis and Stroock 1991; Frigessi *et al*. 1993; Rosenthal 1993). However, there are precedents for a data-based approach, notably the papers of Raftery and Lewis (1992) and Roberts (1992). Intermediate between the data-based and analytical approaches is that of Mykland *et al*. (1995), which is based on the regeneration principle.

Our own approach is based on a two-stage sampling procedure, in which data from the first stage are used to estimate the rate of convergence and hence to determine the length of the second stage. This idea was motivated by Raftery and Lewis (1992), who adopted a similar point of view but whose analysis was based on an assumption that the process, reduced to suitable indicator variables, could be viewed as a two-state Markov chain. Although the method seems to produce good results in practice, the justification for treating the two-state reduced process as Markov remains doubtful. The only other approach we are aware of that is based on direct estimation of the rate of convergence is Roberts (1992). However, Roberts did not consider the sampling properties of his proposed diagnostic procedure, whereas a considerable part of the present paper is devoted to obtaining such properties for our procedure.

In Section 2 we define the class of Markov chains we are considering – essentially, ones for which the transition operator is Hilbert–Schmidt – after which we define the estimators and prove sampling properties including consistency and asymptotic normality. Section 3 provides further discussion, including our main result: a theoretical standard is derived against which other estimators of the second largest eigenvalue may be compared in terms of mean square error and the number of iterations. An example based on Bayesian analysis of actual data is given in Section 4.

2. A Markov chain method for sampling algorithms

Consider an irreducible Markov chain $\{X_n,\ n \geqslant 0\}$ on a state space Ω, where X_n is sampled at time n. Let X_0 have a known initial distribution $\Pi^*(\cdot)$, and let $\{X_n,\ n \geqslant 0\}$ have an unknown stationary distribution $\Pi(\cdot)$ as $n \to \infty$. The Radon–Nikodym derivative (assumed to exist) of Π with respect to some measure ν on the measurable space, $(\Omega, \mathcal{F})$ evaluated at y is denoted by $\pi(y)$ or $\Pi(\mathrm{d}y)/\nu(\mathrm{d}y)$, and the Radon–Nikodym derivative of Π^* with respect to ν evaluated at y is denoted by $\pi^*(y)$ or $\Pi^*(\mathrm{d}y)/\nu(\mathrm{d}y)$. Assume that one's objective is to estimate $\Pi(D)$, where D is any fixed non-empty proper subset of Ω. We are choosing a specific D to focus on a particular problem and to simplify the procedure. Define

$$Z_n = I(X_n \in D),$$

where $I(\cdot)$ is the indicator function. Also, let

$$\rho = \Pi(D), \qquad \rho_n = \mathrm{E}(Z_n), \qquad n = 0, 1, \ldots.$$

Assume throughout without loss of generality that ρ is strictly between zero and one. For fixed $M_0 \geqslant 0$ and $N_0 > M_0$, one can estimate ρ by

$$\hat{\rho}_{M_0,N_0} = (N_0 - M_0)^{-1} \sum_{n=M_0+1}^{N_0} Z_n. \tag{1}$$

This estimate of ρ is performed in the second stage of the two-stage sampling procedure. Our aim will be to choose M_0 and N_0 so that the variance and bias of $\hat{\rho}_{M_0,N_0}$ are less than some specified small numbers. This is done by estimating the second largest eigenvalue of the Markov chain, as performed in the first stage of the two-stage sampling procedure.

2.1. Markov chains generated by Hilbert–Schmidt operators

We now make explicit the class of Markov chains we are considering, which is essentially the Hilbert–Schmidt class on $(\Omega, \mathcal{F}, \Pi)$. Let $\mathscr{L}^2$ denote the space of measurable functions $F: \Omega \to \mathbb{R}$ for which

$$\|F\|_{\mathrm{HS}}^2 = \int_\Omega |F(x)|^2 \Pi(\mathrm{d}x) < \infty.$$

Then, $\mathscr{L}^2$ is a Hilbert space with inner product

$$\langle F, G\rangle_{\mathrm{HS}} = \int_\Omega F(x)G(x)\Pi(\mathrm{d}x), \qquad \forall F, G \in \mathscr{L}^2.$$

We shall assume that the Markov chain is reversible and the transition probability measures are absolutely continuous with respect to the dominating measure ν, with densities $A(x, y)$, $x, y \in \Omega$. Assume the function $y \mapsto A(x, y)/\pi(y)$ is in $\mathscr{L}^2$ for each x. This defines an operator $\mathbb{A}$ on $\mathscr{L}^2$, given by

$$\mathbb{A}F(x) = \int_\Omega A(x, y)F(y)\nu(\mathrm{d}y) = \int_\Omega [A(x, y)/\pi(y)]F(y)\Pi(\mathrm{d}y).$$

Then $\mathbb{A}$ is self-adjoint. It is *Hilbert–Schmidt* if

$$\int_\Omega\int_\Omega |A(x, y)/\pi(y)|^2 \Pi(\mathrm{d}y)\Pi(\mathrm{d}x) = \int_\Omega\int_\Omega |A(x, y)|^2 [\pi(x)/\pi(y)]\nu(\mathrm{d}x)\nu(\mathrm{d}y) < \infty. \tag{2}$$

Roberts (1992; 1994) has previously discussed self-adjoint Hilbert–Schmidt operators in the context of Markov chain Monte Carlo methods. The Gibbs sampler is not, in general, a reversible Markov chain, though with some simple modifications (for example, alternating a forward and a backward order of updating) it may be made into one. Then (2) is automatic if Ω is compact, and can often be verified in non-compact cases. On the other hand, the Metropolis–Hastings algorithm is typically not Hilbert–Schmidt, because there is positive probability that the chain remains in the same state, so transition densities do not exist. Smith (1994) has computed a specific example (the independence Metropolis chain) for which discrete expansions of the form (3) below do not exist, and must be replaced by integrals. Our theory does not at present cover such cases. On the other hand, condition (3) below is the one that really matters, and it is conceivable that this could be satisfied for reversible Markov chains without the Hilbert–Schmidt condition.

For a self-adjoint Hilbert–Schmidt operator $\mathbb{A}$ on $\mathscr{L}^2$, there exists an orthonormal basis

$\{e_k(\cdot),\ k = 1, 2, \ldots\}$ of $\mathscr{L}^2$ with eigenvalues $\{\lambda_k,\ k = 1, 2, \ldots\}$ (Dunford and Schwartz, 1963, pp. 1009–1034) such that

$$[\mathbb{A}e_k](\cdot) = \lambda_k e_k(\cdot), \qquad k = 1, 2, \ldots .$$

Since $\mathbb{A}$ is generated by an ergodic Markov chain, then an eigenvalue of $\mathbb{A}$ is $\lambda_1 = 1$, and $e_1(\cdot) \equiv 1$ Π-a.s., and λ_1 is the largest eigenvalue in modulus.

For an ergodic Markov chain $\{X_n,\ n \geqslant 0\}$ satisfying the above condition, the Radon–Nikodym derivative of the probability that X_n is in state $\mathrm{d}y$ during the $(n+m)$th iteration given that X_n is in state x during the nth iteration with respect to $\nu(\mathrm{d}y)$ exists, and for some $r \geqslant 0$ can be written

$$\Pi_{n+m}(\mathrm{d}y|X_n = x)/\nu(\mathrm{d}y) = \sum_{k=1}^{\infty} \alpha_k(y|x)(\lambda_k)^m,\ \Pi\text{-a.e. } x \in \Omega,\ \nu\text{-a.e. } y \in \Omega, \qquad \forall m \geqslant r, \quad (3)$$

where

$$1 = \lambda_1 > |\lambda_2| \geqslant |\lambda_3| \geqslant |\lambda_4| \geqslant \ldots, \quad (4)$$

and λ_2 is real. The $\alpha_k(y|x)$ are real-valued and are defined by

$$\alpha_k(y|x) = e_k(x)e_k(y)\pi(y),\ \Pi\text{-a.e. } x \in \Omega,\ \nu\text{-a.e. } y \in \Omega, \qquad k = 1, 2, \ldots .$$

If (4) holds and

$$\sum_{k=1}^{\infty} |\lambda_k|^r < \infty, \quad (5)$$

then

$$\mathrm{E}Z_n - \rho = \sum_{k=2}^{\infty} (\lambda_k)^n \int_{\Omega}\int_{D} \alpha_k(y|x)\nu(\mathrm{d}y)\Pi^*(\mathrm{d}x), \qquad \forall n \geqslant r, \quad (6)$$

$$= a_2\lambda_2^n + O(|\lambda_\kappa|^n) \qquad \text{as } n \to \infty, \text{ for some } \kappa \geqslant 3 \text{ and some } a_2,$$

assuming that

$$\sup_{\{k=1,2,\ldots\}} \left| \int_{\Omega}\int_{D} \alpha_k(y|x)\nu(\mathrm{d}y)\Pi^*(\mathrm{d}x) \right| < \infty, \quad (7)$$

where

$$\rho = \int_{\Omega}\int_{D} \alpha_1(y|x)\nu(\mathrm{d}y)\Pi^*(\mathrm{d}x). \quad (8)$$

Note that (6) can be rewritten

$$\mathrm{E}Z_n - \rho = a_2\lambda_2^n + \sum_{k=\kappa}^{\infty} a_k(\lambda_k)^n, \quad (9)$$

for some a_k.

If the Markov chain $\{X_n,\ n \geqslant 0\}$ is discrete and finite, then (5) holds for $r = 0$. If the

Markov chain generates a self-adjoint operator of trace class, then (5) holds for $r = 1$. In the more general Hilbert–Schmidt case, (5) holds for $r = 2$, and (6) is satisfied since (7) holds.

Notice that a_2 depends on $\Pi^*(\cdot)$ and D, but λ_2 depends on neither. The parameter ρ depends on D but not on $\Pi^*(\cdot)$. Assuming that the ergodic Markov chain induces a self-adjoint Hilbert–Schmidt operator, (5) implies that the right-hand side of (3) is absolutely convergent a.e. $\nu(\mathrm{d}y)\Pi(\mathrm{d}x)$, for all $n \geqslant r = 2$. The result also holds for $r = 1$ if the operator is trace class (Yosida, 1965, p. 281).

For the self-adjoint Hilbert–Schmidt case, the parameters $\lambda_1, \lambda_2, \ldots$ are the eigenvalues of the operator $\mathbb{A}$. Multiple eigenvalues of λ_2 in modulus are not problematic, provided that $\lambda_2 \neq -\lambda_k$, for all k. It is assumed that

$$|\lambda_2| > |\lambda_\kappa| \quad \text{and} \quad a_2 \neq 0 \tag{10}$$

to avoid non-identifiability problems.

Under (6) one can write

$$\lambda_2^{-M_0-1}(N_0 - M_0)\mathrm{bias}(\hat{\rho}_{M_0,N_0}) \to a_2(1-\lambda_2)^{-1} \qquad \text{as } N_0 - M_0 \to \infty. \tag{11}$$

Furthermore, for any stationary, ergodic, reversible Markov chain, Kipnis and Varadhan (1986) show that

$$(N_0 - M_0)\mathrm{var}(\hat{\rho}_{M_0,N_0}) \to T^* \qquad \text{as } N_0 \to \infty, \tag{12}$$

for fixed M_0 and some finite constant T^*. Geyer (1992) and Besag and Green (1993) state that T^* can be bounded according to

$$T^*[\rho(1-\rho)]^{-1} \leqslant [1 + |\lambda_2|][1 - |\lambda_2|]^{-1}. \tag{13}$$

Also, (6) implies that $|a_\kappa|$ can be bounded for Markov chains which generate Hilbert–Schmidt operators since

$$\begin{aligned}\left|\int_\Omega\int_D \alpha_k(y|x)\nu(\mathrm{d}y)\Pi^*(\mathrm{d}x)\right| &= \left|\int_\Omega\int_D e_k(x)e_k(y)\pi(y)\nu(\mathrm{d}y)\Pi^*(\mathrm{d}x)\right| \\ &\leqslant \left[\int_\Omega |e_k(x)|\pi^*(x)\nu(\mathrm{d}x)\right]\left[\int_D |e_k(y)|\Pi(\mathrm{d}y)\right] \\ &\leqslant \|\pi^*(\cdot)/\pi(\cdot)\|_{\mathrm{HS}}\sqrt{\Pi(D)}\end{aligned}$$

by the Cauchy–Schwarz inequality, where $\|\cdot\|_{\mathrm{HS}}$ denotes the Hilbert–Schmidt norm. By symmetry it follows that

$$|a_k| \leqslant \|\pi^*/\pi\|_{\mathrm{HS}}\sqrt{\min(\rho, 1-\rho)}, \qquad k = 2, 3, \ldots.$$

Hence, for fixed Π^* both

$$|\mathrm{bias}(\hat{\rho}_{M_0,N_0})|[\min(\rho, 1-\rho)]^{-1/2} \text{ and } [\rho(1-\rho)]^{-1}\,\mathrm{var}(\hat{\rho}_{M_0,N_0})$$

can be minimized uniformly for all non-trivial $D \in \Omega$. To estimate (1) accurately, (11) and (12) are made to be sufficiently small by choosing M_0 and $N_0 - M_0$ sufficiently large.

2.2. A least-squares estimator

We have seen in (6) that the key parameter in determining the rate of convergence is λ_2, which under the Hilbert–Schmidt assumptions is also the second largest eigenvalue of the operator $\mathbb{A}$. We also see from (11)–(13) that a_2 and λ_2 are the key parameters required to bound the bias and variance of the estimator $\hat{\rho}_{M_0,N_0}$. It will now be argued that preliminary estimates of ρ, a_2, and λ_2 can be obtained in the first stage of a two-stage procedure. The second stage will be (1), with M_0 and N_0 chosen sufficiently large to obtain a much more accurate estimate of the parameter ρ, which is the ultimate objective.

Assume that (10) holds, and generate L replications $\{X_n^{(l)}, 0 \leqslant n \leqslant N, 1 \leqslant l \leqslant L\}$ of the first N steps of the Markov chain. Similarly, let

$$Z_n^{(l)} = I(X_n^{(l)} \in D), \qquad 0 \leqslant n \leqslant N, 1 \leqslant l \leqslant L.$$

The estimator will depend on the values of $Z_n^{(l)}$ for $M < n \leqslant N$, $1 \leqslant l \leqslant L$. Here L, M and N are integers, and the purpose of the following analysis is to give some guidance as to how they should be chosen. For asymptotic calculations we shall let $M \to \infty$ and write L and N both as functions of M.

The criteria for estimating ρ, a_2, and λ_2 in the first stage are as follows. Define the vector

$$\boldsymbol{\theta}_0 = (\rho, a_2, \lambda_2)^{\mathrm{T}},$$

which are the true values ρ, a_2, and λ_2 to be estimated. Moreover, define the dummy vector

$$\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)^{\mathrm{T}},$$

where θ_1, θ_2 and θ_3 denote generic values of the parameters ρ, a_2, and λ_2 being estimated, respectively. Also, define

$$\rho_n(\boldsymbol{\theta}) = \theta_1 + \theta_2\theta_3^n,$$

and the sum of squares

$$S_M(\boldsymbol{\theta}) = \sum_{n=M+1}^{N} \left[L^{-1} \sum_{l=1}^{L} Z_n^{(l)} - \rho_n(\boldsymbol{\theta}) \right]^2. \tag{14}$$

It will be shown that there exists a value, $\hat{\boldsymbol{\theta}}_M$ say, such that $S_M(\hat{\boldsymbol{\theta}}_M)$ is a relative minimum of $S_M(\cdot)$ and $\hat{\boldsymbol{\theta}}_M \to \boldsymbol{\theta}_0$ in probability as $M \to \infty$. Furthermore, the probability that $\hat{\boldsymbol{\theta}}_M$ absolutely minimizes $S_M(\cdot)$ converges to one, within some open symmetric sphere which converges to $\boldsymbol{\theta}_0$, as $M \to \infty$.

The conditions on N and L as a function of M are now to be determined. First, choose λ_{1*} such that

$$|\lambda_2| < \lambda_{1*} < 1 \quad \text{and} \quad |\lambda_\kappa| < \lambda_{1*}|\lambda_2|. \tag{15}$$

Furthermore, choose N sufficiently large such that

$$N - (1 + \varepsilon_0)M \to \infty \quad \text{as } M \to \infty, \qquad \text{for some } \varepsilon_0 > 0. \tag{16}$$

Also, choose L such that

$$1/\sqrt{L} = O(|\lambda_{1*}\lambda_2|^M/[M^4N]) \qquad \text{as } M \to \infty. \tag{17}$$

Define the matrix

$$\mathbf{J}_M(\boldsymbol{\theta}) = \left\{ 2 \sum_{n=M+1}^{N} \frac{\partial \rho_n(\boldsymbol{\theta})}{\partial \theta_j} \frac{\partial \rho_n(\boldsymbol{\theta})}{\partial \theta_k} \right\} \Bigg|_{1 \leqslant j,k \leqslant 3}, \tag{18}$$

$\|\cdot\|$ to be the L^2 norm, and $\mathbf{t}$ to be a three-dimensional dummy vector. Before showing that a consistent estimator of $\boldsymbol{\theta}_0$ exists as M, N, and L tend to infinity, the following lemma will be proved:

Lemma 2.1. *Under conditions (15) and (16),*

$$\inf_{\{\mathbf{t}:\|\mathbf{t}\|=1\}} \mathbf{t}^{\mathrm{T}} \mathbf{J}_M(\boldsymbol{\theta}_0)\mathbf{t} \geqslant K_0 \lambda_2^{2M} M^{-2}, \qquad \forall M > 0, \text{ some } K_0 > 0.$$

Proof. From (18) one can write

$$\tfrac{1}{2}\mathbf{J}_M(\boldsymbol{\theta}_0)$$

$$= \begin{bmatrix} N-M & \dfrac{\lambda_2^{M+1}}{1-\lambda_2} & \dfrac{a_2[M(1-\lambda_2)+1]\lambda_2^M}{(1-\lambda_2)^2} \\ \dfrac{\lambda_2^{M+1}}{1-\lambda_2} & \dfrac{\lambda_2^{2M+2}}{1-\lambda_2^2} & \dfrac{a_2[M(1-\lambda_2^2)+1]\lambda_2^{2M+1}}{(1-\lambda_2^2)^2} \\ \dfrac{a_2[M(1-\lambda_2)+1]\lambda_2^M}{(1-\lambda_2)^2} & \dfrac{a_2[M(1-\lambda_2^2)+1]\lambda_2^{2M+1}}{(1-\lambda_2^2)^2} & \dfrac{a_2^2(q_1^2M^2+2q_1M+q_2)\lambda_2^{2M}}{(1-\lambda_2^2)^3} \end{bmatrix}$$

$$+ \begin{bmatrix} 0 & O(|\lambda_2|^N) & O(N|\lambda_2|^N) \\ O(|\lambda_2|^N) & O(|\lambda_2|^{2N}) & O(N|\lambda_2|^{2N}) \\ O(N|\lambda_2|^N) & O(N|\lambda_2|^{2N}) & O(N^2|\lambda_2|^{2N}) \end{bmatrix} \qquad \text{as } N-M \to \infty,$$

where

$$q_1 = 1-\lambda_2^2, \qquad q_2 = 1+\lambda_2^2.$$

The eigenvalues of $\frac{1}{2}\mathbf{J}_M(\boldsymbol{\theta}_0)$ are the zeros of the function

$$g_M(\Lambda) = \left| \tfrac{1}{2}\mathbf{J}_M(\boldsymbol{\theta}_0) - \begin{bmatrix} \Lambda & 0 & 0 \\ 0 & \Lambda & 0 \\ 0 & 0 & \Lambda \end{bmatrix} \right|.$$

For any $\varepsilon_1 > 0$, by direct expansion of the determinant, we can show that

$$\left|\frac{\partial g_M(\Lambda)}{\partial \Lambda}\right| \leqslant c_1(N-M)M^2\lambda_2^{2M}, \qquad \forall |\Lambda| < \varepsilon_1 \lambda_2^{2M} M^{-2}, \text{ for some } c_1 < \infty.$$

Hence,

$$\begin{aligned} |g_M(\Lambda) - g_M(0)| &< [c_1(N-M)M^2\lambda_2^{2M}]\varepsilon_1\lambda_2^{2M}M^{-2} \\ &< \varepsilon_1 c_1 (N-M)\lambda_2^{4M}, \qquad \forall |\Lambda| < \varepsilon_1 \lambda_2^{2M} M^{-2}. \end{aligned}$$

Since

$$(N-M)^{-1}\lambda_2^{-4M-4} g_M(0) \to a_2^2(1-\lambda_2^2)^{-4} \qquad \text{as } M \to \infty \text{ and } N - M \to \infty,$$

then

$$g_M(\Lambda) > 0 \text{ whenever } |\Lambda| < \varepsilon_1 \lambda_2^{2M} M^{-2} \qquad \text{as } M \to \infty \text{ and } N - M \to \infty,$$

for sufficiently small ε_1. Hence, the smallest eigenvalue of $\frac{1}{2}\mathbf{J}_M(\boldsymbol{\theta}_0)$ is at least as large as $\varepsilon_1 \lambda_2^{2M} M^{-2}$ as $M \to \infty$ and $N - M \to \infty$. Now, since $\frac{1}{2}\mathbf{J}_M(\boldsymbol{\theta}_0)$ is symmetric, it can be written

$$\tfrac{1}{2}\mathbf{J}_M(\boldsymbol{\theta}_0) = \boldsymbol{\Sigma}^{\mathrm{T}}\mathbf{A}\boldsymbol{\Sigma},$$

where $\boldsymbol{\Sigma}$ is orthonormal, and $\mathbf{A}$ is a diagonal matrix consisting of the eigenvalues of $\frac{1}{2}\mathbf{J}_M(\boldsymbol{\theta}_0)$. Therefore, since $\|\boldsymbol{\Sigma}\mathbf{t}\| = 1$ whenever $\|\mathbf{t}\| = 1$, then

$$\inf_{\{\mathbf{t}:\|\mathbf{t}\|=1\}} \mathbf{t}^{\mathrm{T}}\boldsymbol{\Sigma}^{\mathrm{T}}\mathbf{A}\boldsymbol{\Sigma}\mathbf{t} \geqslant \varepsilon_1 \lambda_2^{2M} M^{-2} \qquad \text{as } M \to \infty \text{ and } N - M \to \infty. \qquad \square$$

Now, using λ_{1*} as defined in (15), define the sphere

$$\boldsymbol{\Theta}_M = \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| = \lambda_{1*}^M\},$$

and define the interior of $\boldsymbol{\Theta}_M$ by

$$\boldsymbol{\Theta}_M^{(\mathrm{int})} = \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| < \lambda_{1*}^M\}. \tag{19}$$

Lemma 2.1 can be strengthened to include a wider range of values of $(\theta_1, \theta_2, \theta_3)$ as shown in the following lemma:

Lemma 2.2. *Under conditions (15) and (16),*

$$\inf_{\{\boldsymbol{\theta}\in\boldsymbol{\Theta}_M \cup \boldsymbol{\Theta}_M^{(\mathrm{int})}\}} \inf_{\{\mathbf{t}:\|\mathbf{t}\|=1\}} \mathbf{t}^{\mathrm{T}}\mathbf{J}_M(\boldsymbol{\theta})\mathbf{t} \geqslant K_1(|\lambda_2| - \lambda_{1*}^M)^{2M} M^{-2}, \qquad \forall M > 0, \ \textit{some } K_1 > 0.$$

Proof. The infimum of $|\theta_3|$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}_M \cup \boldsymbol{\Theta}_M^{(\mathrm{int})}$ is $|\lambda_2| - \lambda_{1*}^M$. The proof then follows from Lemma 2.1 by replacing λ_2 by all $\boldsymbol{\theta} \in \boldsymbol{\Theta}_M \cup \boldsymbol{\Theta}_M^{(\mathrm{int})}$ when determining how large M needs to be. $\square$

Consistency can be shown after using the following lemma, which is stated without proof:

Lemma 2.3. *Under conditions (15)–(17), if* $(\boldsymbol{\theta}_M - \boldsymbol{\theta}_0)^{\mathrm{T}} \nabla S_M(\boldsymbol{\theta}_M) > 0$, $\forall \boldsymbol{\theta}_M \in \boldsymbol{\Theta}_M$, *then there exists* $\hat{\boldsymbol{\theta}}_M \in \boldsymbol{\Theta}_M^{(\mathrm{int})}$ *such that* $\hat{\boldsymbol{\theta}}_M$ *is a relative minimum of* $S_M(\cdot)$.

An asymptotic expression is established in the next lemma to aid in proving Theorem 2.1.

Lemma 2.4. *For any sequence* $\{b_n, n > 0\}$, *one can write*

$$\frac{1}{L}\sum_{l=1}^{L}\sum_{n=M+1}^{N} b_n[Z_n^{(l)} - \rho_n(\boldsymbol{\theta}_0)] = O\left(\sum_{n=M+1}^{N} |b_n|\,|\lambda_\kappa|^n\right) + O_P\left(\sum_{n=M+1}^{N} |b_n|/\sqrt{L}\right) \quad \text{as } M \to \infty.$$

Proof. Observing that one can write

$$\begin{aligned}\frac{1}{L}\sum_{l=1}^{L}\sum_{n=M+1}^{N} b_n[Z_n^{(l)} - \rho_n(\boldsymbol{\theta}_0)] &= \frac{1}{L}\sum_{l=1}^{L}\sum_{n=M+1}^{N} b_n\{[Z_n^{(l)} - \mathrm{E}Z_n^{(l)}] + [\mathrm{E}Z_n^{(l)} - \rho_n(\boldsymbol{\theta}_0)]\} \\ &= \frac{1}{L}\sum_{l=1}^{L}\sum_{n=M+1}^{N} b_n[Z_n^{(l)} - \mathrm{E}Z_n^{(l)}] \\ &\quad + O\left(\sum_{n=M+1}^{N} |b_n|\,|\lambda_\kappa|^n\right) \qquad \text{as } M \to \infty\end{aligned}$$

from (6), and that the Cauchy–Schwarz inequality implies that

$$\mathrm{var}\left\{\frac{1}{L}\sum_{l=1}^{L}\sum_{n=M+1}^{N} b_n[Z_n^{(l)} - \mathrm{E}Z_n^{(l)}]\right\} \leqslant \frac{1}{L}\mathrm{E}\left\{\sum_{n=M+1}^{N} |b_n|\,|Z_n^{(1)} - \mathrm{E}Z_n^{(1)}|\right\}^2,$$

the result follows from Chebyshev's inequality since $|Z_n^{(1)} - \mathrm{E}Z_n^{(1)}| \leqslant 1$ a.s. □

As an immediate consequence of Lemma 2.4, we note that

$$\frac{1}{L}\sum_{l=1}^{L}\sum_{n=M+1}^{N} [Z_n^{(l)} - \rho_n(\boldsymbol{\theta}_0)] = O(|\lambda_\kappa|^M) + O_P([N-M]/\sqrt{L}) \qquad \text{as } M \to \infty. \tag{20}$$

Furthermore, if $|\lambda| < 1$, $c_M = o(1/M)$ as $M \to \infty$, and $\upsilon \geqslant 0$, then Lemma 2.4 also implies that

$$\frac{1}{L}\sum_{l=1}^{L}\sum_{n=M+1}^{N} n^{\upsilon}(\lambda + c_M)^n[Z_n^{(l)} - \rho_n(\boldsymbol{\theta}_0)] = O(M^{\upsilon}|\lambda\lambda_\kappa|^M) + O_P(M^{\upsilon}|\lambda|^M/\sqrt{L}) \qquad \text{as } M \to \infty. \tag{21}$$

Now, consistency of $\hat{\boldsymbol{\theta}}_M$ is proved in the following theorem:

Theorem 2.1. *Under conditions (15)–(17), the probability that there exists a unique relative minimum* $\hat{\boldsymbol{\theta}}_M$ *of* $S_M(\cdot)$ *in the set* $\boldsymbol{\Theta}_M^{(\mathrm{int})}$ *tends to one as* $M \to \infty$. *Uniqueness implies that this relative minimum is also an absolute minimum over the set* $\boldsymbol{\Theta}_M^{(\mathrm{int})}$. *The rate of convergence is governed by*

$$\|\hat{\boldsymbol{\theta}}_M - \boldsymbol{\theta}_0\| = O_P(\lambda_{1*}^M) \qquad \text{as } M \to \infty.$$

Proof. We first prove that $\hat{\boldsymbol{\theta}}_M$ exists. Observe that

$$\frac{\partial^2 \rho_n(\boldsymbol{\theta})}{\partial\theta_j \partial\theta_k} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & n\theta_3^{n-1} \\ 0 & n\theta_3^{n-1} & n(n-1)\theta_2\theta_3^{n-2} \end{bmatrix}, \qquad n = M+1, \ldots, N.$$

Define the set

$$\mathbf{A}_M = \{\boldsymbol{\theta} \in \boldsymbol{\Theta}_M : |\theta_1 - \rho| \leqslant |\lambda_{1*}\lambda_2|^M\},$$

and let $\boldsymbol{\theta}_M \in \boldsymbol{\Theta}_M$. Using a Taylor series expansion about $\boldsymbol{\theta}_0$, there is a $\boldsymbol{\theta}_M^*$ on the line segment connecting $\boldsymbol{\theta}_0$ to $\boldsymbol{\theta}_M$ such that

$$(\boldsymbol{\theta}_M - \boldsymbol{\theta}_0)^{\mathrm{T}} \nabla S_M(\boldsymbol{\theta}_M) = (\boldsymbol{\theta}_M - \boldsymbol{\theta}_0)^{\mathrm{T}} \nabla S_M(\boldsymbol{\theta}_0) + (\boldsymbol{\theta}_M - \boldsymbol{\theta}_0)^{\mathrm{T}} [\nabla^2 S_M(\boldsymbol{\theta}_M^*)](\boldsymbol{\theta}_M - \boldsymbol{\theta}_0). \tag{22}$$

By Lemma 2.3, it suffices to prove that the above expression is strictly positive on $\boldsymbol{\Theta}_M$ with probability tending to one as $M \to \infty$. We do this splitting by into two cases.

Case 1. Assume that $\boldsymbol{\theta}_M \in \mathbf{A}_M$. The right-hand side of (22) is equivalent to $Q_1 + Q_2 + Q_3$, where

$$Q_1 = (\boldsymbol{\theta}_M - \boldsymbol{\theta}_0)^{\mathrm{T}} \nabla S_M(\boldsymbol{\theta}_0),$$

$$Q_2 = -2(\boldsymbol{\theta}_M - \boldsymbol{\theta}_0)^{\mathrm{T}} \sum_{n=M+1}^{N} \left[L^{-1} \sum_{l=1}^{L} Z_n^{(l)} - \rho_n(\boldsymbol{\theta}_M^*) \right] \nabla^2 \rho_n(\boldsymbol{\theta}_M^*)(\boldsymbol{\theta}_M - \boldsymbol{\theta}_0),$$

and

$$Q_3 = (\boldsymbol{\theta}_M - \boldsymbol{\theta}_0)^{\mathrm{T}} \mathbf{J}_M(\boldsymbol{\theta}_M^*)(\boldsymbol{\theta}_M - \boldsymbol{\theta}_0).$$

The objective is to show that Q_3 dominates Q_1 and Q_2 as $M \to \infty$. Lemma 2.2 implies that

$$Q_3 > K_2 |\lambda_{1*}\lambda_2|^{2M} M^{-2}, \qquad \forall M > 0, \text{ some } K_2 > 0.$$

The bound for Q_1 can be split into three terms, corresponding to the three components of $\boldsymbol{\theta}_0$. It follows from (20) that

$$[\boldsymbol{\theta}_M - \boldsymbol{\theta}_0]_1 [\nabla S_M(\boldsymbol{\theta}_0)]_1 = O(|\lambda_{1*}\lambda_2\lambda_\kappa|^M) + O_P(|\lambda_{1*}\lambda_2|^M [N-M]/\sqrt{L}),$$

$$\forall \boldsymbol{\theta}_M \in \mathbf{A}_M, \quad \text{as } M \to \infty.$$

Also, (21) implies that

$$[\boldsymbol{\theta}_M - \boldsymbol{\theta}_0]_2[\nabla S_M(\boldsymbol{\theta}_0)]_2 = O(|\lambda_{1*}\lambda_2\lambda_\kappa|^M) + O_P(|\lambda_{1*}\lambda_2|^M/\sqrt{L}), \quad \forall \boldsymbol{\theta}_M \in \boldsymbol{\Theta}_M, \quad \text{as } M \to \infty,$$

and

$$[\boldsymbol{\theta}_M - \boldsymbol{\theta}_0]_3[\nabla S_M(\boldsymbol{\theta}_0)]_3 = O(M|\lambda_{1*}\lambda_2\lambda_\kappa|^M) + O_P(M|\lambda_{1*}\lambda_2|^M/\sqrt{L}),$$
$$\forall \boldsymbol{\theta}_M \in \boldsymbol{\Theta}_M, \quad \text{as } M \to \infty.$$

From (15) and (17) it follows that

$$Q_1 = o(Q_3) \qquad \text{as } M \to \infty.$$

The term Q_2 is equivalent to $R_1 + R_2$, where

$$R_1 = -2(\boldsymbol{\theta}_M - \boldsymbol{\theta}_0)^{\mathrm{T}} \sum_{n=M+1}^{N} \left[L^{-1} \sum_{l=1}^{L} Z_n^{(l)} - \rho_n(\boldsymbol{\theta}_0) \right] \nabla^2 \rho_n(\boldsymbol{\theta}_M^*)(\boldsymbol{\theta}_M - \boldsymbol{\theta}_0)$$

$$= O(M^2|\lambda_{1*}^2\lambda_2\lambda_\kappa|^M) + O_P(M^2|\lambda_{1*}^2\lambda_2|^M/\sqrt{L}), \qquad \forall \boldsymbol{\theta}_M \in \boldsymbol{\Theta}_M, \ \text{as } M \to \infty, \qquad \text{from (21)},$$

and

$$R_2 = -2(\boldsymbol{\theta}_M - \boldsymbol{\theta}_0)^{\mathrm{T}} \sum_{n=M+1}^{N} [\rho_n(\boldsymbol{\theta}_0) - \rho_n(\boldsymbol{\theta}_M^*)] \nabla^2 \rho_n(\boldsymbol{\theta}_M^*)(\boldsymbol{\theta}_M - \boldsymbol{\theta}_0)$$

$$= O(M^3|\lambda_{1*}^3\lambda_2^2|^M), \qquad \forall \boldsymbol{\theta}_M \in \mathbf{A}_M, \quad \text{as } M \to \infty,$$

since

$$\rho_n(\boldsymbol{\theta}_0) - \rho_n(\boldsymbol{\theta}_M^*) = O(M|\lambda_{1*}\lambda_2|^M), \qquad \forall \boldsymbol{\theta}_M \in \mathbf{A}_M, \ \text{as } M \to \infty, \quad \forall n > M.$$

Again, from (15) and (17) it follows that

$$Q_2 = o(Q_3) \qquad \text{as } M \to \infty.$$

Case 2. Assume that $\boldsymbol{\theta}_M \in \boldsymbol{\Theta}_M \backslash \mathbf{A}_M$. As in case 1, decompose the right-hand side of (22) into $Q_1 + Q_2 + Q_3$. Note that

$$Q_3 = 2 \sum_{n=M+1}^{N} \{[\boldsymbol{\theta}_M - \boldsymbol{\theta}_0]_1 + [\boldsymbol{\theta}_M - \boldsymbol{\theta}_0]_2([\boldsymbol{\theta}_M^*]_3)^n + n[\boldsymbol{\theta}_M - \boldsymbol{\theta}_0]_3[\boldsymbol{\theta}_M^*]_2([\boldsymbol{\theta}_M^*]_3)^{n-1}\}^2.$$

Recalling (16), we ignore the terms in $M < n \leqslant M(1 + \varepsilon_0/2)$. For $n > M(1 + \varepsilon_0/2)$, we have

$$[\boldsymbol{\theta}_M - \boldsymbol{\theta}_0]_2([\boldsymbol{\theta}_M^*]_3)^n + n[\boldsymbol{\theta}_M - \boldsymbol{\theta}_0]_3[\boldsymbol{\theta}_M^*]_2([\boldsymbol{\theta}_M^*]_3)^{n-1} = O(M|\lambda_{1*}\lambda_2^{(1+\varepsilon_0/2)}|^M) \qquad \text{as } M \to \infty. \tag{23}$$

Since

$$|[\boldsymbol{\theta}_M - \boldsymbol{\theta}_0]_1| > |\lambda_{1*}\lambda_2|^M,$$

then $[\boldsymbol{\theta}_M - \boldsymbol{\theta}_0]_1$ dominates the left-hand side of (23) for all M sufficiently large, and for some $K_3 > 0$,

$$Q_3 > K_3 M([\theta_M - \theta_0]_1)^2, \qquad \forall M \text{ sufficiently large.}$$

The term Q_3 dominates R_1 and the last two terms of Q_1, using the same argument as in case 1. The first term of Q_1 is bounded by

$$|[\boldsymbol{\theta}_M - \boldsymbol{\theta}_0]_1|\{O(|\lambda_\kappa|^M) + O_P([N-M]/\sqrt{L})\} \qquad \text{as } M \to \infty,$$

which is $o(Q_3)$ as $M \to \infty$ from (15) and (17). Since

$$\rho_n(\boldsymbol{\theta}_0) - \rho_n(\boldsymbol{\theta}_M^*) = O(|[\boldsymbol{\theta}_M - \boldsymbol{\theta}_0]_1| + M|\lambda_{1*}\lambda_2|^M) \qquad \text{as } M \to \infty, \forall n > M,$$

then

$$R_2 = O(|[\boldsymbol{\theta}_M - \boldsymbol{\theta}_0]_1| M^2 |\lambda_{1*}^2 \lambda_2|^M + M^3 |\lambda_{1*}^3 \lambda_2^2|^M) = o(Q_3) \qquad \text{as } M \to \infty.$$

Hence, Q_3 dominates Q_1 and Q_2 as $M \to \infty$.

Combining cases 1 and 2 results in

$$\mathrm{P}\left(\inf_{\{\boldsymbol{\theta}_M \in \boldsymbol{\Theta}_M\}} (\boldsymbol{\theta}_M - \boldsymbol{\theta}_0)^{\mathrm{T}} \nabla S_M(\boldsymbol{\theta}_M) > 0 \right) \to 1 \qquad \text{as } M \to \infty. \tag{24}$$

Therefore, if one defines

$$\hat{\boldsymbol{\theta}}_M = \begin{cases} \text{any value } \in \boldsymbol{\Theta}_M^{(\mathrm{int})} \text{ which locally minimizes } S_M(\cdot), & \text{if } \exists \text{ such a value,} \\ \text{any estimator of } \boldsymbol{\theta}_0, & \text{otherwise,} \end{cases}$$

then

$$\|\hat{\boldsymbol{\theta}}_M - \boldsymbol{\theta}_0\| = O_P(\lambda_{1*}^M) \qquad \text{as } M \to \infty,$$

which follows from (19), (24), and Lemma 2.3.

Finally to prove uniqueness of $\hat{\boldsymbol{\theta}}_M$ we show that the probability that $\hat{\boldsymbol{\theta}}_M$ absolutely minimizes $S_M(\boldsymbol{\theta}_M)$ for all $\boldsymbol{\theta}_M \in \boldsymbol{\Theta}_M^{(\mathrm{int})}$ tends to one as $M \to \infty$. If $\hat{\boldsymbol{\theta}}_M \in \boldsymbol{\Theta}_M^{(\mathrm{int})}$, but $\hat{\boldsymbol{\theta}}_M$ does not absolutely minimize $S_M(\boldsymbol{\theta}_M)$ for all $\boldsymbol{\theta}_M \in \boldsymbol{\Theta}_M^{(\mathrm{int})}$, then there exist at least two values of $\boldsymbol{\theta}_M \in \boldsymbol{\Theta}_M^{(\mathrm{int})}$ such that $\nabla S_M(\boldsymbol{\theta}_M) = \mathbf{0}$. By continuity there must exist a value $\boldsymbol{\theta}_M^\iota \in \boldsymbol{\Theta}_M^{(\mathrm{int})}$ such that $\nabla^2 S_M(\boldsymbol{\theta}_M^\iota) = \mathbf{0}$. Recalling (22) and the fact that Q_3 dominates Q_2 in both cases 1 and 2, it follows that

$$\mathrm{P}(\nabla^2 S_M(\boldsymbol{\theta}_M^\iota) = \mathbf{0}, \text{ some } \boldsymbol{\theta}_M^\iota \in \boldsymbol{\Theta}_M^{(\mathrm{int})}) \to 0 \qquad \text{as } M \to \infty. \qquad \square$$

2.3. The bias of the asymptotic distribution of the estimator

The goal of this section is to determine the bias of the limiting distribution of $\hat{\boldsymbol{\theta}}_M$ as $M \to \infty$, where $\hat{\boldsymbol{\theta}}_M$ is used to estimate $\boldsymbol{\theta}_0$. Throughout this section assume that (6), (10) and conditions (15)–(17) required by Theorem 2.1 hold. One can determine the asymptotic distribution of $\hat{\boldsymbol{\theta}}_M - \boldsymbol{\theta}_0$ as $M \to \infty$ by noting from a Taylor series expansion about $\boldsymbol{\theta}_0$ that

$$\hat{\boldsymbol{\theta}}_M = \boldsymbol{\theta}_0 - [\nabla^2 S_M(\boldsymbol{\theta}_M^{\dagger})]^{-1} \nabla S_M(\boldsymbol{\theta}_0), \tag{25}$$

where $\boldsymbol{\theta}_M^{\dagger}$ lies on the line segment connecting $\boldsymbol{\theta}_0$ to $\hat{\boldsymbol{\theta}}_M$. Establishing (25) is the motivation for first proving consistency of $\hat{\boldsymbol{\theta}}_M$ in Theorem 2.1. Observe from Chebyshev's inequality and (9) that

$$\nabla S_M(\boldsymbol{\theta}_0) = \mathrm{E}\{\nabla S_M(\boldsymbol{\theta}_0)\} + [O_P(N/\sqrt{L}),\ O_P(|\lambda_2|^M/\sqrt{L}),\ O_P(M|\lambda_2|^M/\sqrt{L})]^{\mathrm{T}} \qquad \text{as } M \to \infty,$$

where

$$\mathrm{E}\{\nabla S_M(\boldsymbol{\theta}_0)\}$$

$$= \begin{bmatrix} -2\displaystyle\sum_{k=\kappa}^{\infty} \frac{a_k(\lambda_k^{M+1} - \lambda_k^{N+1})}{1-\lambda_k} \\ -2\displaystyle\sum_{k=\kappa}^{\infty} \frac{a_k[(\lambda_2\lambda_k)^{M+1} - (\lambda_2\lambda_k)^{N+1}]}{1-\lambda_2\lambda_k} \\ \displaystyle\sum_{k=\kappa}^{\infty} \frac{-2a_2a_k\{[M(1-\lambda_2\lambda_k)+1](\lambda_2\lambda_k)^{M+1} - [N(1-\lambda_2\lambda_k)+1](\lambda_2\lambda_k)^{N+1}\}}{\lambda_2(1-\lambda_2\lambda_k)^2} \end{bmatrix}. \tag{26}$$

Furthermore, defining

$$[\mathbf{J}_M^{\infty}(\boldsymbol{\theta})]_{jk} = \begin{cases} 2(N-M), & \text{if } j = k = 1, \\ 2\displaystyle\sum_{n=M+1}^{\infty} \frac{\partial \rho_n(\boldsymbol{\theta})}{\partial \theta_j} \frac{\partial \rho_n(\boldsymbol{\theta})}{\partial \theta_k}, & \text{otherwise,} \end{cases}$$

it follows from matrix algebra that

$$\nabla^2 S_M(\boldsymbol{\theta}_M^{\dagger}) = \mathbf{J}_M^{\infty}(\boldsymbol{\theta}_0) + \begin{bmatrix} 0 & O_P(M|\lambda_{1*}\lambda_2|^M) & O_P(M^2|\lambda_{1*}\lambda_2|^M) \\ O_P(M|\lambda_{1*}\lambda_2|^M) & O_P(M|\lambda_{1*}\lambda_2^2|^M) & O_P(M^2|\lambda_{1*}\lambda_2^2|^M) \\ O_P(M^2|\lambda_{1*}\lambda_2|^M) & O_P(M^2|\lambda_{1*}\lambda_2^2|^M) & O_P(M^3|\lambda_{1*}\lambda_2^2|^M) \end{bmatrix}. \tag{27}$$

The determinant of $\mathbf{J}_M^{\infty}(\boldsymbol{\theta}_0)$ is strictly greater than zero for all M sufficiently large. Also, the probability that the determinant of $\nabla^2 S_M(\boldsymbol{\theta}_M^{\dagger})$ is strictly greater than zero converges to one as $M \to \infty$. Approximating $[\nabla^2 S_M(\boldsymbol{\theta}_M^{\dagger})]^{-1}$ by $[\mathbf{J}_M^{\infty}(\boldsymbol{\theta}_0)]^{-1}$ plus the order terms, when these inverses exist, the asymptotic distribution of $\hat{\boldsymbol{\theta}}_M$ now can be expressed as follows:

Theorem 2.2. *Under conditions (6) and (15)–(17),*

$$\hat{\boldsymbol{\theta}}_M - \boldsymbol{\theta}_0$$

$$= \begin{bmatrix} \dfrac{(1-\lambda_2)}{(N-M)(1-\lambda_2)-2(1+\lambda_2)} \displaystyle\sum_{k=\kappa}^{\infty} \frac{a_k(\lambda_2-\lambda_k)^2\lambda_k^{M+1}}{(1-\lambda_k)(1-\lambda_2\lambda_k)^2} \\ \dfrac{(N-M)(1-\lambda_2)^2(1+\lambda_2)}{[(N-M)(1-\lambda_2)-2(1+\lambda_2)]\lambda_2^{M+3}} \displaystyle\sum_{k=\kappa}^{\infty} \frac{a_k\lambda_k^{M+1}}{(1-\lambda_2\lambda_k)^2} \{M(1-\lambda_2^2)(\lambda_2-\lambda_k)\lambda_2 \\ + (1+\lambda_2^2)(1-\lambda_2\lambda_k)\} \\ \dfrac{-(1-\lambda_2)^3(1+\lambda_2)^2(N-M)}{a_2[(N-M)(1-\lambda_2)-2(1+\lambda_2)]\lambda_2^{M+1}} \displaystyle\sum_{k=\kappa}^{\infty} \frac{a_k\lambda_k^{M+1}}{(1-\lambda_2\lambda_k)^2} \{M(1-\lambda_2)(1-\lambda_2\lambda_k)+1-\lambda_k\} \end{bmatrix}$$

$$+ \begin{bmatrix} O_P(M^4N^{-1}|\lambda_{1*}\lambda_\kappa|^M + 1/\sqrt{L}) \\ O_P(M^5|\lambda_{1*}\lambda_\kappa/\lambda_2|^M + M^2|\lambda_2|^{-M}/\sqrt{L}) \\ O_P(M^4|\lambda_{1*}\lambda_\kappa/\lambda_2|^M + M|\lambda_2|^{-M}/\sqrt{L}) \end{bmatrix} \quad \text{as } M \to \infty.$$

Proof. The proof follows from (25) and matrix calculations. □

We have shown in Theorem 2.2 that $\hat{\boldsymbol{\theta}}_M - \boldsymbol{\theta}_0$ can be written in terms of the asymptotic bias plus higher-order terms as $M \to \infty$.

2.4. Evaluation of the covariance matrix of the gradient vector

Section 2.3 involved determining the asymptotic bias of $\hat{\boldsymbol{\theta}}_M$ via (25) and approximating $[\nabla^2 S_M(\boldsymbol{\theta}_M^\dagger)]^{-1}$ by $[\mathbf{J}_M^\infty(\boldsymbol{\theta}_0)]^{-1}$ plus higher-order terms. In order to determine the asymptotic behaviour of the covariance matrix of $\hat{\boldsymbol{\theta}}_M$, the asymptotic covariance matrix of the vector $\nabla S_M(\boldsymbol{\theta}_0)$ also needs to be determined as

$$N(1-\varepsilon_2) - 2M \to \infty \qquad \text{as } M \to \infty, \text{ some } 0 < \varepsilon_2 < 1, \tag{28}$$

where (28) is a stronger condition than (16): that is the goal of this section.

Assume that (3) holds for $r = 1$. This condition is always satisfied for chains which induce self-adjoint trace class operators and for discrete, finite Markov chains. Also, assume that (6) and (10) hold. Since $\lambda_1 = 1$ and the stationary distribution of the Markov chain $\{X_n,\ n \geq 0\}$ is independent of the starting state, then (3) implies that

$$\alpha_1(y) = \alpha_1(y|x),\ \Pi\text{-a.e. } x \in \Omega,\ \nu\text{-a.e. } y \in \Omega,$$

so that (8) may be rewritten

$$\rho = \int_D \alpha_1(y)\nu(\mathrm{d}y).$$

Thus, one can write

$$\operatorname{cov}(Z_n^{(1)}, Z_{n+m}^{(1)})$$

$$= P(Z_n^{(1)} = Z_{n+m}^{(1)} = 1) - P(Z_n^{(1)} = 1)P(Z_{n+m}^{(1)} = 1)$$

$$= \sum_{k=1}^{\infty} \lambda_k^n \int_\Omega \int_D \left\{ \alpha_k(y|x) \sum_{l=2}^{\infty} \lambda_l^m \int_D \left[\alpha_l(z|y) - \lambda_l^n \int_\Omega \alpha_l(z|w) \Pi^*(\mathrm{d}w) \right] \nu(\mathrm{d}z) \right\} \nu(\mathrm{d}y) \Pi^*(\mathrm{d}x),$$

$$\forall n \geqslant 0,\ m \geqslant 1. \quad (29)$$

Notice that (29) also holds for $m = 0$ in the Hilbert–Schmidt case since

$$\sum_{l=1}^{\infty} \int_D \alpha_l(z|y) \nu(\mathrm{d}z) = I(y \in D).$$

Therefore,

$$\mathrm{L}\operatorname{cov}\left(\frac{\partial S_M(\boldsymbol{\theta})}{\partial \theta_j}, \frac{\partial S_M(\boldsymbol{\theta})}{\partial \theta_k} \right) = 4 \operatorname{cov}\left(\sum_{n=M+1}^{N} Z_n \frac{\partial \rho_n(\boldsymbol{\theta})}{\partial \theta_j}, \sum_{n=M+1}^{N} Z_n \frac{\partial \rho_n(\boldsymbol{\theta})}{\partial \theta_k} \right),$$

and the covariance matrix of $\nabla S_M(\boldsymbol{\theta}_0)$ can be defined by

$$\mathbf{V}_M(\boldsymbol{\theta}_0) := L \operatorname{var}\{\nabla S_M(\boldsymbol{\theta}_0)\}.$$

Under (28) one can conveniently express

$$\begin{aligned} \tfrac{1}{4}\mathbf{V}_M(\boldsymbol{\theta}_0) = & \begin{bmatrix} T_{11}(N-M) & T_{12}\lambda_2^M & T_{12}a_2 M\lambda_2^{M-1} \\ T_{12}\lambda_2^M & T_{22}\lambda_2^{2M} & T_{22}a_2 M\lambda_2^{2M-1} \\ T_{12}a_2 M\lambda_2^{M-1} & T_{22}a_2 M\lambda_2^{2M-1} & T_{22}a_2^2 M^2\lambda_2^{2M-2} \end{bmatrix} \\ & + \begin{bmatrix} S_{11} & S_{12}\lambda_2^{2M} & S_{13}\lambda_2^M \\ S_{12}\lambda_2^{2m} & S_{22}\lambda_2^{3M} & S_{23}\lambda_2^{2M} \\ S_{13}\lambda_2^M & S_{23}\lambda_2^{2M} & 2S_{23}a_2 M\lambda_2^{2M-1} + S_{33}\lambda_2^{2M} \end{bmatrix} \\ & + \begin{bmatrix} O(|\lambda_2|^M) & O(|\lambda_2|^{3M} + |\lambda_2\lambda_\kappa|^M) & O(M|\lambda_2|^{2M}) \\ O(|\lambda_2|^{3M} + |\lambda_2\lambda_\kappa|^M) & O(|\lambda_2|^{4M} + |\lambda_2^2\lambda_\kappa|^M) & O(M|\lambda_2|^{3M}) \\ O(M|\lambda_2|^{2M}) & O(M|\lambda_2|^{3M}) & O(|\lambda_{1*}\lambda_2^2|^M) \end{bmatrix} \end{aligned} \quad (30)$$

as $M \to \infty$ under conditions (15), (17), and (28), where

$$T_{11} = \sum_{l=2}^{\infty}(1+\lambda_l)(1-\lambda_l)^{-1}\int_D\int_D \alpha_1(y)\alpha_l(z|y)\nu(\mathrm{d}z)\nu(\mathrm{d}y),$$

$$T_{12} = \lambda_2(1-\lambda_2)^{-1}\sum_{l=2}^{\infty}[1-\lambda_2\lambda_l^2][(1-\lambda_l)(1-\lambda_2\lambda_l)]^{-1}\int_D\int_D \alpha_1(y)\alpha_l(z|y)\nu(\mathrm{d}z)\nu(\mathrm{d}y),$$

$$T_{22} = \lambda_2^2[1-\lambda_2^2]^{-1}\sum_{l=2}^{\infty}[1+\lambda_2\lambda_l][1-\lambda_2\lambda_l]^{-1}\int_D\int_D \alpha_1(y)\alpha_l(z|y)\nu(\mathrm{d}z)\nu(\mathrm{d}y),$$

$$S_{11} = -2\sum_{l=2}^{\infty}\lambda_l(1-\lambda_l)^{-2}\int_D\int_D \alpha_1(y)\alpha_l(z|y)\nu(\mathrm{d}z)\nu(\mathrm{d}y),$$

$$S_{12} = \lambda_2^2[1-\lambda_2^2]^{-1}\int_\Omega\int_D\int_D\Bigg(-\alpha_1(y)\alpha_2(z|x)+\sum_{k=2}^{\infty}\alpha_2(y|x)\alpha_k(z|y)+\sum_{l=2}^{\infty}\lambda_l[1+\lambda_2-2\lambda_2\lambda_l]$$
$$\times[(1-\lambda_l)(1-\lambda_2\lambda_l)]^{-1}[-\alpha_1(y)\alpha_2(z|x)+\alpha_2(y|x)\alpha_1(z|y)]\Bigg)\nu(\mathrm{d}z)\nu(\mathrm{d}y)\Pi^*(\mathrm{d}x),$$

$$S_{13} = a_2(1-\lambda_2)^{-1}\int_D\int_D \alpha_1(y)\alpha_l(z|y)\nu(\mathrm{d}z)\nu(\mathrm{d}y)\sum_{l=2}^{\infty}\big\{(1-\lambda_2)^{-1}$$
$$+\lambda_l[[(1-\lambda_2)(1-\lambda_l)]^{-1}+\lambda_2[(1-\lambda_2)(1-\lambda_2\lambda_l)]^{-1}+\lambda_2(1-\lambda_2\lambda_l)^{-2}]\big\},$$

$$S_{22} = \lambda_2^3[1-\lambda_2^3]^{-1}\int_\Omega\int_D\int_D\Bigg(-\alpha_1(y)\alpha_2(z|x)+\sum_{k=2}^{\infty}\alpha_2(y|x)\alpha_k(z|y)$$
$$+2\lambda_2\sum_{l=2}^{\infty}\lambda_l(1-\lambda_2\lambda_l)^{-1}[-\alpha_1(y)\alpha_2(z|x)+\alpha_2(y|x)\alpha_1(z|y)]\Bigg)\nu(\mathrm{d}z)\nu(\mathrm{d}y)\Pi^*(\mathrm{d}x),$$

$$S_{23} = a_2\lambda_2(1-\lambda_2^2)^{-2}\sum_{l=2}^{\infty}\left[1+\frac{\lambda_2\lambda_l(33-\lambda_2^2-2\lambda_2\lambda_l)}{(1-\lambda_2\lambda_l)^2}\right]\int_D\int_D \alpha_1(y)\alpha_l(z|y)\nu(\mathrm{d}z)\nu(\mathrm{d}y)$$

and

$$S_{33} = \alpha_2^2[1-\lambda_2^2]^{-3}\sum_{l=2}^{\infty}(1-\lambda_2\lambda_l)^{-2}\{(1+\lambda_2^2)(1-\lambda_2\lambda_l)^2$$
$$+2\lambda_2\lambda_l(2-\lambda_2^3\lambda_l-\lambda_2\lambda_l)\}\int_D\int_D \alpha_1(y)\alpha_l(z|y)\nu(\mathrm{d}z)\nu(\mathrm{d}y).$$

2.5. A central limit theorem for the estimator

The goal of this section is to prove a central limit theorem for $\hat{\boldsymbol{\theta}}_M$ using Lindeberg's condition. Assume throughout this section that conditions (6), (10), (15), (17) and (28) hold. To condense notation, the matrix $\mathbf{V}_M(\boldsymbol{\theta}_0)$ is expressed by $\mathbf{V}_M$.

It is easy to show that $\mathbf{V}_M$ is positive definite, because for any non-zero $\mathbf{t} = (t_1, t_2, t_3)^{\mathrm{T}}$, we have

$$\mathbf{t}^{\mathrm{T}}\mathbf{V}_M\mathbf{t} = 4\,\mathrm{var}\left\{\sum_{n=M+1}^{N} Z_n^{(1)}(t_1 + t_2\lambda_2^n + t_3 n a_2 \lambda_2^{n-1})\right\} > 0.$$

It will be helpful to define the 3×3 symmetric matrix $\mathbf{V}_M^{-1/2}$ by the following:

$$\mathbf{V}_M^{-1/2}\mathbf{V}_M^{-1/2} = \mathbf{V}_M^{-1}.$$

Since $\mathbf{V}_M$ is symmetric, the matrix $\mathbf{V}_M^{-1/2}$ can be shown to exist. Direct calculation and (30) imply that

$$\begin{aligned}
([\mathbf{V}_M^{-1/2}]_{1,1})^2 + ([\mathbf{V}_M^{-1/2}]_{1,2})^2 + ([\mathbf{V}_M^{-1/2}]_{1,3})^2 &= [\mathbf{V}_M^{-1}]_{1,1} = O(N^{-1}),\\
([\mathbf{V}_M^{-1/2}]_{1,2})^2 + ([\mathbf{V}_M^{-1/2}]_{2,2})^2 + ([\mathbf{V}_M^{-1/2}]_{2,3})^2 &= [\mathbf{V}_M^{-1}]_{2,2} = O(M^2|\lambda_2|^{-2M}),\\
([\mathbf{V}_M^{-1/2}]_{1,3})^2 + ([\mathbf{V}_M^{-1/2}]_{2,3})^2 + ([\mathbf{V}_M^{-1/2}]_{3,3})^2 &= [\mathbf{V}_M^{-1}]_{3,3} = O(|\lambda_2|^{-2M})
\end{aligned} \tag{31}$$

as $M \to \infty$. From these three equations (31) it follows that

$$\mathbf{V}_M^{-1/2} = \begin{bmatrix} O(N^{-1/2}) & O(N^{-1/2}) & O(N^{-1/2}) \\ O(N^{-1/2}) & O(M|\lambda_2|^{-M}) & O(|\lambda_2|^{-M}) \\ O(N^{-1/2}) & O(|\lambda_2|^{-M}) & O(|\lambda_2|^{-M}) \end{bmatrix} \quad \text{as } M \to \infty. \tag{32}$$

One can write

$$\begin{aligned}
&\sqrt{L}\mathbf{V}_M^{-1/2}\mathbf{J}_M^{\infty}(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_M - \boldsymbol{\theta}_0 + [\mathbf{J}_M^{\infty}(\boldsymbol{\theta}_0)]^{-1}\mathrm{E}\{\nabla S_M(\boldsymbol{\theta}_0)\})\\
&\quad = \sqrt{L}\mathbf{V}_M^{-1/2}\mathbf{J}_M^{\infty}(\boldsymbol{\theta}_0)\{[\mathbf{J}_M^{\infty}(\boldsymbol{\theta}_0)]^{-1} - [\nabla^2 S_M(\boldsymbol{\theta}_M^{\dagger})]^{-1}\}\nabla S_M(\boldsymbol{\theta}_0)\\
&\qquad - \sqrt{L}\mathbf{V}_M^{-1/2}[\nabla S_M(\boldsymbol{\theta}_0) - \mathrm{E}\{\nabla S_M(\boldsymbol{\theta}_0)\}] \qquad \text{from (25).}
\end{aligned} \tag{33}$$

We shall show that as $M \to \infty$, the second term on the right-hand side of (33) converges in distribution to a multivariate normal distribution with mean $\mathbf{0}$ and variance $\mathbf{I}_3$, where $\mathbf{0}$ is the zero vector of dimension 3 and $\mathbf{I}_3$ is the 3×3 identity matrix. Also, we shall show that the first term on the right-hand side of (33) tends to $\mathbf{0}$ in probability as $M \to \infty$ whenever the condition

$$L = o(M^{-12}|\lambda_{1*}\lambda_{\kappa}|^{-2M}) \qquad \text{as } M \to \infty \tag{34}$$

holds. This additional condition on L will be shown to result in a central limit theorem on $\hat{\boldsymbol{\theta}}_M$ as $M \to \infty$. This condition is needed because if L were allowed to become arbitrarily large as M increases, then the square of the bias of $\hat{\boldsymbol{\theta}}_M$, which is not a function of L, might

dominate the variance of $\hat{\boldsymbol{\theta}}_M$, which decreases with increasing L. It will be shown in the proof of Theorem 2.3 that values of N and L exist which satisfy (17), (28) and (34) simultaneously.

To prove that the second term on the right-hand side of (33) converges in distribution to $\mathcal{MN}(\mathbf{0}, \mathbf{I}_3)$, it is enough to prove that

$$\sqrt{L}\mathbf{t}^{\mathrm{T}}\mathbf{V}_M^{-1/2}[\nabla S_M(\boldsymbol{\theta}_0) - \mathrm{E}\{\nabla S_M(\boldsymbol{\theta}_0)\}] \xrightarrow{\mathcal{D}} \mathcal{N}(0, \|\mathbf{t}\|^2), \tag{35}$$

for all vectors $\mathbf{t}$ of dimension 3. Since (35) is satisfied trivially for $\mathbf{t} = \mathbf{0}$, assume without loss of generality that $\mathbf{t} \neq \mathbf{0}$. First define

$$\nabla S_M^{(l)}(\boldsymbol{\theta}_0) = -2 \sum_{n=M+1}^{N} [Z_n^{(l)} - \rho_n(\boldsymbol{\theta}_0)](1, \lambda_2^n, na_2\lambda_2^{n-1})^{\mathrm{T}}, \qquad l = 1, \ldots, L.$$

Letting $\varepsilon_3 > 0$, Lindeberg's condition,

$$\lim_{M\to\infty} (L\|\mathbf{t}\|^2)^{-1} \sum_{l=1}^{L} \mathrm{E}\{(\mathbf{t}^{\mathrm{T}}[\nabla S_M^{(l)}(\boldsymbol{\theta}_0) - \mathrm{E}\{\nabla S_M^{(l)}(\boldsymbol{\theta}_0)\}])^2$$

$$I(|\mathbf{t}^{\mathrm{T}}\mathbf{V}_M^{-1/2}[\nabla S_M^{(l)}(\boldsymbol{\theta}_0) - \mathrm{E}\{\nabla S_M^{(l)}(\boldsymbol{\theta}_0)\}]/\sqrt{L}| \geqslant \varepsilon_3\|\mathbf{t}\|)\} \to 0 \qquad \text{as } M \to \infty, \tag{36}$$

for proving (35) will be shown to hold. Since (17), (31) and the bounded nature of $Z_n^{(l)}$ imply that

$$\mathbf{t}^{\mathrm{T}}\mathbf{V}_M^{-1/2}[\nabla S_M^{(l)}(\boldsymbol{\theta}_0) - \mathrm{E}\{\nabla S_M^{(l)}(\boldsymbol{\theta}_0)\}]/\sqrt{L} = O([N - M][N^{-1/2} + M\lambda_2^{-M}])/\sqrt{L}$$

$$= o(1) \quad \text{as } M \to \infty,$$

then (36) holds. This proves (35) for any $\mathbf{t}$, and hence

$$-\sqrt{L}\mathbf{V}_M^{-1/2}[\nabla S_M(\boldsymbol{\theta}_0) - \mathrm{E}\{\nabla S_M(\boldsymbol{\theta}_0)\}] \xrightarrow{\mathcal{D}} \mathcal{MN}(\mathbf{0}, \mathbf{I}_3) \qquad \text{as } M \to \infty. \tag{37}$$

Since (26), (27), (32) and (34) imply that the first term on the right-hand side of (33) converges to $\mathbf{0}$ in probability under conditions (15), (17), (28) and (34), the asymptotic distribution of $\hat{\boldsymbol{\theta}}$ now can be stated.

Theorem 2.3. *Under conditions (6), (15), (17), (28) and (34),*

$$\sqrt{L}\mathbf{V}_M^{-1/2}\mathbf{J}_M^{\infty}(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_M - \boldsymbol{\theta}_0 + [\mathbf{J}_M^{\infty}(\boldsymbol{\theta}_0)]^{-1}\mathrm{E}\{\nabla S_M(\boldsymbol{\theta}_0)\}) \xrightarrow{\mathcal{D}} \mathcal{MN}(\mathbf{0}, \mathbf{I}_3) \qquad \text{as } M \to \infty.$$

Proof. First observe that a large class of values of N and L exists which satisfies conditions (17), (28) and (34) simultaneously; for example, let $N = M^2$ and $L = M^{14}|\lambda_{1*}\lambda_2|^{-2M}$. The result follows by applying to (33) and (37) Slutsky's lemma, which is (for example) Theorem 4.4.6 from Chung (1974, p. 92). □

Using tedious algebra or a computer software package, the covariance matrix of the asymptotic distribution of $\hat{\boldsymbol{\theta}}_M$ can be expressed by

$$L\,\mathrm{var}\{\hat{\boldsymbol{\theta}}_M - ([\mathbf{J}_M^{\infty}(\boldsymbol{\theta}_0)]^{-1} - [\nabla^2 S_M(\boldsymbol{\theta}_M^{\dagger})]^{-1})\nabla S_M(\boldsymbol{\theta}_0)\}$$

$$= [\mathbf{J}_M^{\infty}(\boldsymbol{\theta}_0)]^{-1}\mathbf{V}_M(\theta_0)[\mathbf{J}_M^{\infty}(\boldsymbol{\theta}_0)]^{-1}, \qquad \forall \text{ valid } M, \tag{38}$$

$$= \begin{bmatrix} \dfrac{T_{11}}{N-M} & \dfrac{W_{12}M}{[N-M]\lambda_2^M} & \dfrac{-W_{12}}{a_2[N-M]\lambda_2^{M-1}} \\ \dfrac{W_{12}M}{[N-M]\lambda_2^M} & \dfrac{W_{22}M^2}{\lambda_2^{2M}} & \dfrac{-W_{22}M}{a_2\lambda_2^{2M-1}} \\ \dfrac{-W_{12}}{a_2[N-M]\lambda_2^{M-1}} & \dfrac{-W_{22}M}{a_2\lambda_2^{2M-1}} & \dfrac{W_{22}}{a_2^2\lambda_2^{2M-2}} \end{bmatrix}$$

$$+ \begin{bmatrix} O\left(\dfrac{1}{N^2}\right) & O\left(\dfrac{1}{N|\lambda_2|^M}\right) & O\left(\dfrac{1}{N}\left|\dfrac{\lambda_{1*}}{\lambda_2}\right|^M + \dfrac{1}{N^2|\lambda_2|^M}\right) \\ O\left(\dfrac{1}{N|\lambda_2|^M}\right) & O\left(\dfrac{M}{|\lambda_2|^{2M}}\right) & O\left(\dfrac{1}{|\lambda_2|^{2M}}\right) \\ O\left(\dfrac{1}{N}\left|\dfrac{\lambda_{1*}}{\lambda_2}\right|^M + \dfrac{1}{N^2|\lambda_2|^M}\right) & O\left(\dfrac{1}{|\lambda_2|^{2M}}\right) & O\left(\left|\dfrac{\lambda_{1*}}{\lambda_2^2}\right|^M + \dfrac{1}{N|\lambda_2|^{2M}}\right) \end{bmatrix}$$

as $M \to \infty$, where

$$W_{12} = (1-\lambda_2)(1+\lambda_2)^2[a_2^2\lambda_2^6]^{-1}\{T_{11}a_2^2\lambda_2^4 + T_{12}a_2^2\lambda_2^2(1-\lambda_2)$$
$$+ T_{22}a_2^2(1-\lambda_2)^2(1+\lambda_2) - S_{13}a_2\lambda_2^3(1-\lambda_2)^2(1+\lambda_2)$$
$$- S_{23}a_2\lambda_2(2-\lambda_2-4\lambda_2^2+2\lambda_2^3+2\lambda_2^4-\lambda_2^5) + S_{33}\lambda_2^2(1-\lambda_2^2)^3\}$$

and

$$W_{22} = (1-\lambda_2^2)^4[a_2^2\lambda_2^8]^{-1}[T_{22}a_2^2 - 2S_{23}a_2\lambda_2(1-\lambda_2^2) + S_{33}\lambda_2^2(1-\lambda_2^2)^2].$$

3. Discussion

The issue remaining is how to choose M, N and L in practice. Since we are interested in estimating λ_2 and since

$$L^{-1}\sum_{l=1}^{L} Z_n^{(l)} = \rho + a_2\lambda_2^n + O(|\lambda_\kappa|^n) + O_P(1/\sqrt{L}) \qquad \text{as } L \to \infty, \tag{39}$$

reducing the bias of $\hat{\boldsymbol{\theta}}_M$ requires values of n sufficiently large such that $|\lambda_\kappa/\lambda_2|^n$ is negligible, as Theorem 2.2 quantifies. However, if M and N are chosen too large, then the $a_2\lambda_2^n$ term in (39) becomes negligible relative to the $O_P(1/\sqrt{L})$ term, and for fixed L the variances of $\hat{a}_2$

and $\hat{\lambda}_2$ become large, as (38) quantifies. These variances may be reduced simply by choosing L large. Qualitatively, therefore, the results show that M needs to be moderately large so that $|\lambda_\kappa/\lambda_2|^M$ is negligible, but then L needs to be as large as possible so that $\sqrt{L}|\lambda_2|^M$ is large.

Since computing time may be limited, M, N and L should ideally be chosen to allow a trade-off between bias and variance of $\hat{\lambda}_2$, which are

$$O(M|\lambda_\kappa/\lambda_2|^M) \text{ and } O(L^{-1}|\lambda_2|^{-2M}) \qquad \text{as } M \to \infty, \tag{40}$$

respectively, using Theorems 2.2 and 2.3. Choosing M, N and L to allow an optimal trade-off is possible in theory but not in practice since knowledge of $|\lambda_2|$ and $|\lambda_\kappa|$ is required. Since the bias of $\hat{\lambda}_2$ decays geometrically as $M \to \infty$, M typically need not be chosen too large. The value of N required is $O(M)$ as $M \to \infty$, and the values of L required increases geometrically in M and proportionally to N^2, as implied by (16) and (17), respectively.

In practice one may want to fix values of N and L with L as large as practically possible, and estimate λ_2 using various values of M. The value of M needed to allow an optimal trade-off between bias and variance can be determined theoretically. Using (40) to equate the asymptotic variance to the square of the asymptotic bias, it follows that

$$M = -[\log\sqrt{L} + \log\log L][\log|\lambda_\kappa|]^{-1} + O(1) \qquad \text{as } L \to \infty,$$

and the optimal mean square error is

$$\text{MSE}(\hat{\lambda}_2) = O(L^{\log|\lambda_2|/\log|\lambda_\kappa|-1}[\log L]^{2\log|\lambda_2|/\log|\lambda_\kappa|}) \qquad \text{as } L \to \infty.$$

If one insists that $\sqrt{\text{MSE}}$ be less than some $\varepsilon_4 > 0$, then (40) implies that M and L should be chosen according to

$$M = [-\log\varepsilon_4 + \log(-\log\varepsilon_4)][\log|\lambda_2/\lambda_\kappa|]^{-1} + O(1) \qquad \text{as } \varepsilon_4 \to 0,$$

and

$$L = O(\varepsilon_4^{-2}[-\varepsilon_4^{-1}\log\varepsilon_4]^{2\log|\lambda_2|/\log|\lambda_\kappa/\lambda_2|}) \qquad \text{as } \varepsilon_4 \to 0.$$

The total number of Monte Carlo variables needed is

$$NL = O(ML) = O(\varepsilon_4^{-1}[-\varepsilon_4^{-1}\log\varepsilon_4]^{1+2\log|\lambda_2|/\log|\lambda_\kappa/\lambda_2|}) \qquad \text{as } \varepsilon_4 \to 0. \tag{41}$$

Remark 3.1. Equation (41) provides a standard against which other estimators of λ_2 may be judged. We make no claim that our estimator of λ_2 is the best possible, even in an asymptotic sense, but if some other estimator were proposed, then the sample size needed to achieve a specified mean square error, ε_4^2, could in principle be computed and compared with (41).

We are not aware of any results related to (41) in the current literature on Markov chain Monte Carlo convergence diagnostics.

4. An example

This two-stage sampling procedure is applied to an example, known as the hierarchical Poisson model, which has been studied by Gelfand and Smith (1990) and Tierney (1994) in

the Gibbs sampling context using data previously analysed by Gaver and O'Muircheartaigh (1987). The data, which are listed in Table 1, are denoted by y_i, the number of failures at the ith pump at a nuclear power plant at fixed time t_i, for $i = 1, \ldots, 10$.

The hierarchical Poisson model is now described. Assume that the y_i are independently Poisson distributed with mean $\omega_i t_i$. Also, assume that the ω_i have independent gamma distributions $\Gamma(\alpha, \beta)$, whose densities are $\omega^{\alpha-1}\exp\{-\omega/\beta\}/\beta^{\alpha}\Gamma(\alpha)$. The parameter $1/\beta$ is chosen to have a $\Gamma(\gamma, \delta)$ distribution, where $\gamma = 0.01$ and $\delta = 1$. Moreover, the parameter α is chosen to be 1.802, which is the method of moments estimator as suggested by Gelfand and Smith (1990). The values of α, γ and δ used herein are the same ones as used by both Gelfand and Smith (1990) and Tierney (1994).

To set up the Gibbs sampler, one needs to be able to sample from the conditional distributions $[\omega_i|\omega_j, j \neq i; \mathbf{y}]$, $i = 1, \ldots, 10$, and $[1/\beta|\omega, \mathbf{y}]$. It can be shown that

$$[\omega_i|\beta, \omega_j, j \neq i; \mathbf{y}] = [\omega_i|\beta, y_i]$$
$$\sim \Gamma(\alpha + y_i, (t_i + 1/\beta)^{-1}), \qquad i = 1, \ldots, 10,$$

and

$$[1/\beta|\omega, \mathbf{y}] = [1/\beta|\omega]$$
$$\sim \Gamma\left(\gamma + 10\alpha, \left\{\sum_{i=1}^{10}\omega_i + 1/\delta\right\}^{-1}\right), \qquad i = 1, \ldots, 10.$$

Using (2), it now will be shown that the Hilbert–Schmidt double norm is finite for the hierarchical Poisson model. Letting $f(\cdot)$ denote the appropriate densities, for the hierarchical Poisson model it follows that

Table 1. Number of pump failures at a nuclear power plant

Pump i	y_i	t_i
1	5	94.320
2	1	15.720
3	5	62.880
4	14	125.760
5	3	5.240
6	19	31.440
7	1	1.048
8	1	1.048
9	4	2.096
10	22	10.480

Note: y_i is the number of failures at pump i, and t_i is the time when the failures at pump i are observed.

$$A(\beta^*, \beta)/\pi(\beta) \propto \int_0^\infty \cdots \int_0^\infty \left[\sum_{i=1}^{10} f(\omega_i|\beta^*, y_i)\right] f(\beta|\omega, \mathbf{y})\, \mathrm{d}\omega_1 \ldots \mathrm{d}\omega_{10}$$

$$\times \left[f(\beta) \sum_{j=1}^{10} \int_0^\infty f(y_j|\omega_j^\dagger) f(\omega_j^\dagger|\beta)\, \mathrm{d}\omega_j^\dagger\right]^{-1}$$

$$\propto \int_0^\infty \cdots \int_0^\infty \left[\prod_{j=1}^{10} \omega_j^{\alpha+y_j-1} \exp\{-\omega_j(t_j+\beta^*+\beta)\}\right]$$

$$\times \left[\delta^{-1} + \sum_{k=1}^{10} \omega_k\right]^{\gamma+10\alpha} \mathrm{d}\omega_1 \ldots \mathrm{d}\omega_{10} \prod_{i=1}^{10} [(t_i+\beta^*)(t_i+\beta)]^{\alpha+y_i}.$$

Since

$$\left[\delta^{-1} + \sum_{k=1}^{10} \omega_k\right]^{\gamma+10\alpha} \leqslant \frac{2(\gamma+10\alpha)}{\min_j(t_j)} \exp\left\{(t_i/2)\left[\delta^{-1} + \sum_{k=1}^{10} \omega_k\right] - 1\right\}, \qquad i = 1, \ldots, 10,$$

then

$$A(\beta^*, \beta)/\pi(\beta) \leqslant \text{ constant} \times \sum_{i=1}^{10} [(t_i+\beta^*)(t_i+\beta)(\beta^*+\beta+t_i/2)^{-1}]^{y_i+\alpha}, \tag{42}$$

and the right-hand side of (42) can tend to infinity only if $\beta \to \infty$ and $\beta^* \to \infty$ simultaneously. Therefore, (2) holds since the tails of $\pi(\beta)$ decay exponentially fast as $\beta \to \infty$, and the Hilbert–Schmidt double norm is finite.

Using the data listed in Table 1 the joint posterior distribution of ω and β is approximated using Gibbs sampling. In the first stage of the two-stage sampling procedure, the initial distribution $\Pi^*(\cdot)$ is chosen to be degenerate at $\beta = 0.01$. To satisfy the reversibility condition the variates are sampled in the order $(\omega_1, \ldots, \omega_{10}, \beta)$. The set for the indicator function $Z_n^{(I)}$ is chosen to be $D = [\beta < 0.42]$.

The number of independent replicates is chosen to be $L = 5000$, and the length of any replicate is chosen to be $N = 12$. Allowing M to vary between 0 and 6, the least-squares estimates of and 95% confidence intervals on ρ, a_2, and λ_2 are graphed in Figures 1, 2 and 3, respectively. These confidence intervals have width 2×1.96 times the standard error. The standard errors are calculated by empirical evaluation of the asymptotic covariance matrix (38), using the variability among different runs of the simulation to estimate $\mathbf{V}_M$, the covariance matrix of $\nabla S_M(\boldsymbol{\theta}_0)$ (the so-called information sandwich method). The relatively narrow confidence intervals on λ_2 using $M = 0$, $M = 1$, and possibly even $M = 2$, suggest that $0.3 < \lambda_2 < 0.5$. The confidence intervals using $M \geqslant 3$ are too large to draw reasonable conclusions.

Run lengths longer than 12 are also used, some as large as 50, but not discussed in detail herein. However, estimates of λ_2 are less stable for longer run lengths since the standard

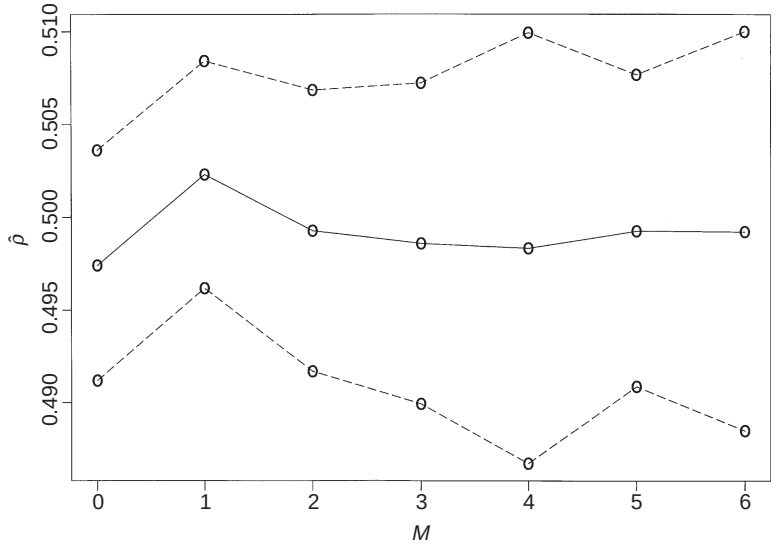

Figure 1. Least-squares estimates of ρ, hierarchical Poisson model

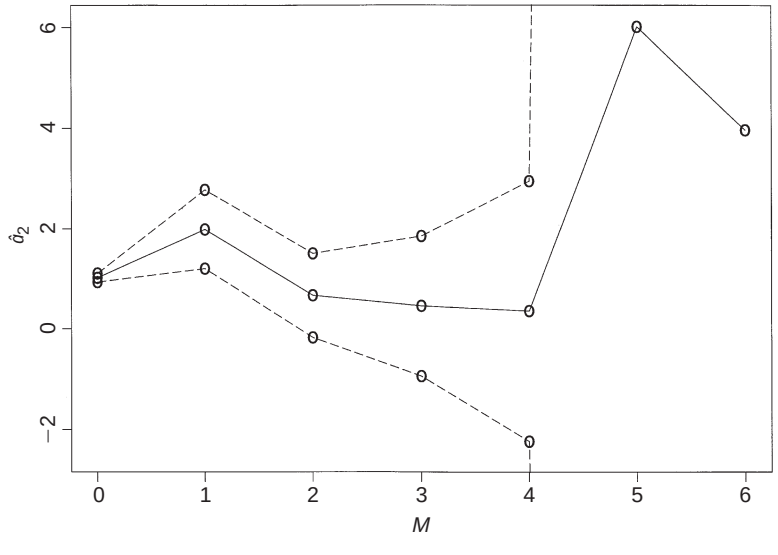

Figure 2. Least-squares estimates of a_2, hierarchical Poisson model

deviation of $\overline{Z}_n$ is large compared to $|a_2\lambda_2^n|$ for large n, as exemplified by (39). Of course if N is too small, then the bias of $\hat{\lambda}_2$ is large. We eventually chose $N = 12$, since all simulations suggest that $|\lambda_2|$ is not close to one. If indeed $|\lambda_2| < 0.5$, then estimating λ_2 accurately becomes difficult but not too pertinent, since covergence is quite rapid.

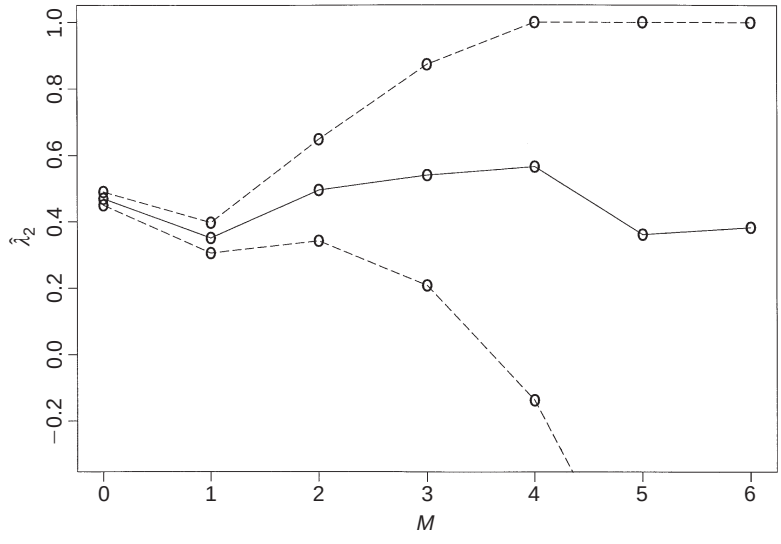

Figure 3. Least-squares estimates of λ_2, hierarchical Poisson model

Other authors also have noted rapid covergence. Using many independent replications, Gelfand and Smith (1990) show that the posterior distribution is practically reached after only $N_0 = 10$ iterations, although they do not specify their initial distribution for (ω, β). Mykland *et al*. (1995) show that the average number of iterations required for the chain to reach a regeneration point is 2.56, suggesting that convergence is rapid. Rosenthal (1995) shows that after N iterations the total variation norm can be bounded by

$$0.976^N + 0.951^N \left[6.2 + \mathrm{E}\left(\sum_{i=1}^{10} \omega_i^{(0)} - 6.5 \right)^2 \right],$$

where the $\omega_i^{(0)}$ are the initial values of the ω_i, and hence an upper bound on λ_2 is 0.976; this appears in the light of our results to be a much too conservative estimate.

In an attempt to estimate the true convergence rate for this example as accurately as possible, the whole procedure has been repeated with L increased to 10^5. Such a large value of L usually would not be practical, but it is here in view of the small value of N and the fast operation of the Gibbs sampler for this problem. We took $N = 12$, various sets D, and degenerate initial distributions on β (Figure 4). The simulations suggest that a reasonable estimate of λ_2 is 0.3, using $M = 1$ for Figure 4a–b, and $M = 0$ for Figure 4c–f. In these figures the solid line represents $\overline{Z}_n$, and the long-dashed lines are pointwise 95% confidence intervals. The short-dashed line represents the least-squares estimate $\hat{\rho} + \hat{\alpha}_2(\hat{\lambda}_2)^n$, using the common estimate $\hat{\lambda}_2 = 0.3$ This line usually falls within the confidence limits, indicating that 0.3 is indeed a reasonable estimate of λ_2.

For the second stage of the two-stage sampling procedure, simulations are generated with

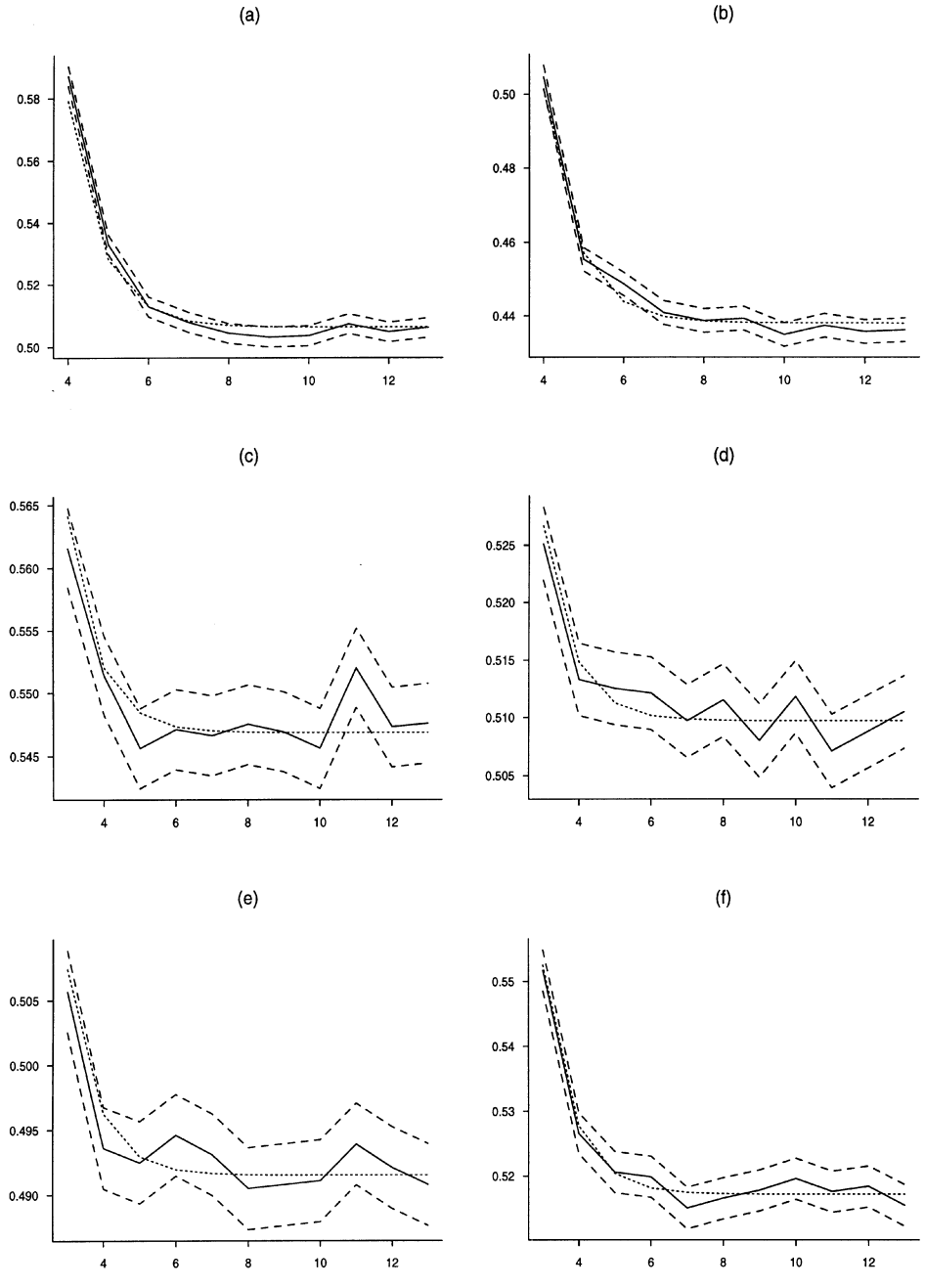

Figure 4. Observed and fitted probabilities, common λ_2 for the following (β, D) pairs: (a) $(0.01, \beta < 0.42)$; (b) $(0.02, \beta < 0.4)$; (c) $(0.02, \omega_1 < 0.07)$; (d) $(0.04, \omega_3 < 0.1)$; (e) $(0.03, \omega_4 < 0.12)$ (f) $(0.04, \omega_2 < 0.14)$.

Table 2. Sample posterior means and standard deviations for hierarchical Poisson model

Parameter	Sample mean	Sample standard deviation
β	0.4408	0.1334
ω_1	0.0703	0.0271
ω_2	0.1539	0.0919
ω_3	0.1040	0.0400
ω_4	0.1234	0.0312
ω_5	0.6293	0.2942
ω_6	0.6141	0.1355
ω_7	0.8285	0.5302
ω_8	0.8320	0.5331
ω_9	1.3026	0.5832
ω_{10}	1.8452	0.3912

Note: These estimates result from using $M_0 = 50$, $N_0 = 10^5$ iterations, and the data listed in Table 1.

$M_0 = 50$ and $N_0 = 10^5$ to minimize the bias and variance of $\hat{\rho}_{M_0,N_0}$. Since $\rho(1-\rho) \leqslant 0.25$, and if $|\lambda_2| \leqslant 0.5$, then (12) and (13) imply that the standard deviation of $\hat{\rho}_{M_0,N_0}$ is less than 0.0027 for all ρ. If $|a_2| \leqslant 3$ and $|\lambda_2| \leqslant 0.5$, then (11) implies that the bias of $\hat{\rho}_{M_0,N_0}$ is negligible for all ρ. The initial distribution $\Pi^*(\cdot)$ for this stage also is chosen to be degenerate at $\beta = 0.01$. The sample means and standard deviations of these N_0 samples of ω and β are listed in Table 2. Using Laplace's method as discussed in Tierney *et al*. (1989), Tierney (1994) reports approximate asymptotic posterior means of ω_1, ω_5 and ω_{10} to be 0.070 28, 0.6279, and 1.8431 and standard deviations to be 0.026 95, 0.2931, and 0.3910, respectively. The 10^5 simulations generated herein using $M_0 = 50$ produce sample means for ω_1, ω_5 and ω_{10} of 0.0703, 0.6293 and 1.8452 and sample deviations of 0.0271, 0.2942 and 0.3912, respectively. The two sets of estimates are in very close agreement.

Acknowledgements

The research of the first author was partially supported by National Institute of Mental Health grant MH53259-01A2. The research of the second author was partially supported by National Science Foundation grant DMS-92-05112. Special thanks are due to Charles R. Baker and Gareth O. Roberts for their many helpful conversations.

References

Besag, J. and Green, P.J. (1993) Spatial statistics and Bayesian computation. *J. Roy. Statist. Soc. Ser. B*, **55**, 25–37.

Chung, K.L. (1974) *A Course in Probability Theory*, 2nd edn. San Diego, CA: Academic Press.

Diaconis, P. and Stroock, D. (1991) Geometric bounds for eigenvalues of Markov chains. *Ann. Appl. Probab.*, **1**, 36–61.

Dunford, N. and Schwartz, J.T. (1963) *Linear Operations, Part II*. New York: Wiley.

Frigessi, A., di Stefano, P., Hwang, C.-R. and Sheu, S.-J., (1993) Convergence rates of the Gibbs sampler, the Metropolis algorithm and other single-site updating dynamics. *J. Roy. Statist. Soc. Ser. B*, **55**, 205–219.

Gaver, D.P. and O'Muircheartaigh, I.G. (1987) Robust empirical Bayes analysis of event rates. *Technometrics*, **29**, 1–15.

Gelfand, A.E. and Smith, A.F.M. (1990) Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.*, **85**, 398–409.

Gelman, A. and Rubin, D.B. (1992) Inference from iterative simulation using multiple sequences. *Statist. Sci.*, **7**, 457–472.

Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intelligence*, **6**, 721–741.

Geyer, C.J. (1992) Practical Markov chain Monte Carlo. *Statist. Sci.*, **7**, 473–483.

Geyer, C.J. and Thompson, E.A. (1992) Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *J. Roy. Statist. Soc. Ser. B.* **54**, 657–699.

Hastings, W.K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

Kipnis, C. and Varadhan, S.R.S. (1986) Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions. *Comm. Math. Phys.*, **104**, 1–19.

Lawler, G.F. and Sokal, A.D. (1988) Bounds on the $\mathscr{L}^2$ spectrum for Markov chains and Markov processes: a generalization of Cheeger's inequality. *Trans. Amer. Math. Soc.*, **309**, 557–580.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953) Equation of state calculations by fast computing machines. *J. Chem. Phys.*, **21**, 1087–1092.

Mykland, P., Tierney, L. and Yu, B. (1995) Regeneration in Markov chain samplers. *J. Amer. Statist. Assoc.*, **90**, 233–241.

Raftery, A.E. and Lewis, S.M. (1992) How many iterations in the Gibbs sampler? In J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith (eds), *Bayesian Statistics 4*, pp. 763–773. New York: Oxford University Press.

Roberts, G.O. (1992) Convergence diagnostics of the Gibbs sampler. In J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith (eds), *Bayesian Statistics 4*, pp. 775–782. New York: Oxford University Press.

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

Rosenthal, J.S. (1993) Rates of convergence for data augmentation of finite sample spaces. *Ann. Appl. Probab.*, **3**, 819–839.

Rosenthal, J.S. (1995) Minorization conditions and convergence rates for Markov chain Monte Carlo. *J. Amer. Statist. Assoc.*, **90**, 558–566.

Sinclair, A. and Jerrum, M. (1989) Approximate counting, uniform generation and rapidly mixing Markov chains. *Inform. and Comput.*, **82**, 93–133.

Smith, A.F.M. and Roberts, G.O. (1993) Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. Roy. Statist. Soc. Ser. B.*, **55**, 3–23.

Smith, R.L. (1994) Exact transition probabilities for the independence Metropolis sampler. Technical Report, University of Cambridge.

Tanner, M.A. and Wong, W.H. (1987) The calculation of posterior distributions by data augmentation.

J. Amer. Statist. Assoc., **82**, 528–540.
Tierney, L. (1994) Markov chains for exploring posterior distributions (with discussion). *Ann. Statist*., **22**, 1701–1762.
Tierney, L., Kass, R.E. and Kadane, J.B. (1989) Fully exponential Laplace approximations to expectations and variances of nonpositive functions. *J. Amer. Statist. Assoc*., **84**, 710–716.
Yosida, K. (1965) *Functional Analysis*. New York: Academic Press.

Received January 1998 and revised October 1998