Empirical estimators based on MCMC data

Priscilla E. Greenwood   Wolfgang Wefelmeyer

1 Introduction

Suppose we want to calculate the expectation of a function f under a distribution π on some space E. If E is of high dimension, or if π is defined indirectly, it may be difficult to calculate the expectation $\pi f = E_\pi f = \int \pi(dx) f(x)$ analytically or even by numerical integration. (The notation πf will be used throughout the paper.) The classical Monte Carlo method generates i.i.d. realizations $X^0, \dots, X^n$ from π, and approximates πf by the *empirical estimator*

$$E_n f = \frac{1}{n} \sum_{i=1}^{n} f(X^i).$$

If f is π-integrable, the estimator is strongly consistent. If f is π-square-integrable, the estimator is asymptotically normal with variance $\pi(f - \pi f)^2$. Often, however, this Monte Carlo method is difficult to implement. One reason is that high dimensional distributions are hard to simulate. Additional difficulties arise when π is defined indirectly, as in many Bayesian modeling situations, or known only up to a normalizing constant, as is usually the case for random fields.

Markov chain Monte Carlo methods (MCMC) generate realizations $X^0, \dots, X^n$ of a Markov chain with π as invariant law. (Here and in the following, by Markov chain we mean a discrete-time Markov process with arbitrary state space, not a continuous-time Markov process with discrete state space.) Again, the empirical estimator $E_n f$ can be used to approximate πf, and we have an explicit expression for the asymptotic variance of the estimator from the ergodic theory for Markov chains; see Section 2.

Choice of an MCMC method amounts to choice of a transition distribution Q from the large family of those with invariant law π. One important criterion is the speed with which the law of the Markov chain converges to π. This problem is well-studied, together with the associated question of how long the sampler must run until the observations are satisfactorily close to stationarity. Recent references are Schervish and Carlin (1992), Chan (1993), Frigessi, Hwang, Sheu and Di Stefano (1993), Tierney (1994), Meyn and Tweedie (1994), Ingrassia (1994), Roberts and Polson (1994), Athreya, Doss and Sethuraman (1996), Rosenthal (1995), Mengersen and Tweedie (1996), Roberts and Tweedie (1996), Johnson (1996), Roberts and Sahu (1997), Kira and Ji (1997), Robert (1998), Diaconis and Saloff-Coste (1998), Jerrum (1998), Roberts and Rosenthal (1998), Roberts and Tweedie (1999, 2000) and Jarner and Roberts (2000). The initial observations from this "burn-in" period are usually discarded.

At this point the transition distribution, Q, used in the sampler may be changed to one which is optimized according to a different criterion. Now the simulated data will be used to estimate πf using either the empirical estimator or possibly an improved estimator which exploits some property of the sampler. It is reasonable to judge the sampler by the asymptotic variance of the empirical estimator. This criterion is utilized by Peskun (1973), Frigessi, Hwang and Younes (1992), Green and Han (1992), Liu, Wong and Kong (1994, 1995), Clifford and Nicholls (1995), Liu (1996), Fishman (1996), and Mira and Tierney (1999). This survey will be about efforts to choose a sampler and an estimator of πf where one starts from an already (approximately) stationary initial distribution. For a short overview see Wefelmeyer (1998).

MCMC methods originated with the study of interacting particle systems (Metropolis, Metropolis, Rosenbluth, Teller and Teller, 1953). More recently, MCMC methods have been applied extensively to image analysis, starting with the Gibbs sampler of Grenander (1983) and Geman and Geman (1984), and to Bayesian statistics (Smith and Roberts, 1993), spatial statistics (Besag and Green, 1993, and Graham, 1994), expert systems (Pearl, 1987, Spiegelhalter, Dawid, Lauritzen and Cowell, 1993), incomplete data problems (Tanner and Wong, 1987), and hierarchical models (Gelfand, Hills, Racine-Poon, A. and Smith, 1990).

The algorithm of Metropolis, Metropolis, Rosenbluth, Teller and Teller (1953) and its generalization by Hastings (1970) construct MCMC samplers as follows. Let $K(x, dy)$ be a *candidate* transition distribution on E. Write $\varepsilon_x(dy)$ for the one-point probability measure with mass at x. Find a function $\alpha(x, y)$ with values in $[0, 1]$ such that

$$Q(x, dy) = K(x, dy)\alpha(x, y) + \varepsilon_x(dy) \int Q(x, dz)(1 - \alpha(x, z))$$

is in *detailed balance* with π,

$$\pi(dx)Q(x, dy) = \pi(dy)Q(y, dx). \tag{1.1}$$

This implies that π is the invariant law of Q, and the chain is *reversible* under the stationary distribution. Assume, for simplicity, that $K(x, dy)$ has density $k(x, y)$, except perhaps for an atom at $x = y$. We refer to Tierney (1998) for a more general discussion. The *Metropolis algorithm* takes $k(x, y)$ symmetric, and

$$\alpha(x, y) = \begin{cases} \min\{1, \frac{\pi(y)}{\pi(x)}\}, & \pi(x)k(x, y) > 0, \\ 1, & \pi(x)k(x, y) = 0. \end{cases}$$

The *Hastings algorithm* does not assume k to be symmetric, and takes

$$\alpha(x, y) = \begin{cases} \min\{1, \frac{\pi(y)k(y,x)}{\pi(x)k(x,y)}\}, & \pi(x)k(x, y) > 0, \\ 1, & \pi(x)k(x, y) = 0. \end{cases}$$

The *independence Hastings algorithm* uses independent candidate realizations, $k(x,y) = k(y)$,

$$\alpha(x,y) = \begin{cases} \min\{1, \frac{\pi(y)k(x)}{\pi(x)k(y)}\}, & \pi(x)k(y) > 0, \\ 1, & \pi(x)k(y) = 0. \end{cases}$$

The algorithms accept the proposal from $K(x,dy)$ with probability $\alpha(x,y)$. If the proposal is rejected, the same position is retained by the chain and the next transition is considered. Tierney and Mira (1999) show that performance is improved if, upon rejection, instead of moving on to the next transition, another attempt to move is made by proposing a new candidate, generated from a different distribution, which is allowed to depend on the previously rejected value. This idea of delaying the rejection and adapting the proposal distribution is generalized to a more flexible class of methods in Green and Mira (1999). These methods apply in particular to settings in which the dimension varies. Optimal scaling of $K(x,dy)$ for high-dimensional state space E is discussed in Gelman, Roberts and Gilks (1996) and Roberts, Gelman and Gilks (1997).

Auxiliary variable algorithms (also called *substitution sampler* or *data augmentation*) consider π as the marginal of an appropriate distribution. For notational convenience, we write π_1 and E_1 for the distribution and state space of interest. Introduce a new state space E_2 and a distribution $\pi(dx_1, dx_2)$ on $E = E_1 \times E_2$, with first marginal $\pi_1(dx_1)$. The distribution π can be factored into marginal and conditional distributions in two different ways:

$$\pi(dx) = \pi_1(dx_1)p_2(x_1, dx_2) = p_1(x_2, dx_1)\pi_2(dx_2).$$

The *auxiliary variable algorithm* is the Markov chain with transition distribution

$$Q(x_1, dy_1) = \int p_2(x_1, dx_2)p_1(x_2, dy_1). \tag{1.2}$$

The algorithm of Swendsen and Wang (1987), see also Edwards and Sokal (1988), is a data augmentation algorithm. The monograph of Tanner (1996) has a chapter on data augmentation. We refer also to Besag and Green (1993), Higdon (1998) and Mira and Tierney (1999).

Another example of an auxiliary variable method is the slice sampler. See Neal (2000), Damien, Wakefield and Walker (1999), Roberts and Rosenthal (1999) and Fishman (1999). The underlying idea is that, as in ordinary rejection sampling, one can simulate from a distribution by simulating uniformly from under its density. For the *simple slice sampler*, we write again π_1 and E_1 for the distribution and state space of interest. We assume that $\pi_1(dx_1)$ has density proportional to $f(x_1)$, and choose a factorization

$$f(x_1) = f_1(x_1)f_2(x_1),$$

where $\sup_{x_1} f_2(x_1) = 1$. We take $E_2 = (0, \infty)$ and $p_2(x_1, dx_2)$ the uniform distribution on $(0, f_2(x_1))$, and introduce the joint distribution

$$\pi(dx) = \pi_1(dx_1)p_2(x_1, dx_2).$$

The conditional distribution $p_1(x_2, dx_1)$ of x_1 given x_2 has density proportional to

$$f_1(x_1)1_{(f_2(x_1)>x_2)}(x_1).$$

The transition distribution of the simple slice sampler is now defined by (1.2).

The *Gibbs sampler* requires that the state space E is a product space, say $E = E_1 \times \ldots \times E_k$. For each $j = 1, \ldots, k$, we can express $x \in E$ by separating out the j-th component: $x = (x_j, x_{-j})$, where x_{-j} is obtained from x by omitting the j-th component x_j. Factor π in k different ways,

$$\pi(dx) = \pi_{-j}(dx_{-j})p_j(x_{-j}, dx_j), \quad j = 1, \ldots, k, \tag{1.3}$$

with $p_j(x_{-j}, dx_j)$ the one-dimensional conditional distribution under π of x_j given x_{-j}, and $\pi_{-j}(dx_{-j})$ the $(k-1)$-dimensional marginal distribution of x_{-j}. Gibbs samplers successively use the transition distributions

$$Q_j(x, dy) = p_j(x_{-j}, dy_j)\varepsilon_{x_{-j}}(dy_{-j})$$

which change only the j-th component of x.

The Gibbs sampler with *deterministic* (and *cyclic*) sweep applies Q_j cyclically according to the numbering $j = 1, \ldots, k$ of the components. The transition distribution at time $i = (q-1)k + j$ is Q_j. The chain is neither homogeneous nor reversible. For $j = 1, \ldots, k-1$, the realization $X^{(q-1)k+j}$ is determined by $X^{(q-1)k}$ and X^{qk} as $(X^{qk}_{\leq j}, X^{(q-1)k}_{>j})$, where $x_{\leq j} = (x_1, \ldots, x_j)$, $x_{>j} = (x_{j+1}, \ldots, x_k)$. Hence nothing is lost if we observe only the chain $X^{(q-1)k}$, $q = 1, 2, \ldots$. By "Gibbs sampler" one often means this subchain of full sweeps, with transition distribution

$$Q_{(d)}(x, dy) = Q_1 \cdots Q_k(x, dy) = \prod_{j=1}^{k} p_j(y_{<j}, x_{>j}, dy_j), \tag{1.4}$$

The subscript (d) stands for *deterministic.* The auxiliary variable method may be viewed as the marginal of a two-step full sweep Gibbs sampler, $k = 2$.

For the Gibbs sampler with *random* sweep (with equal probabilities), each index j is picked according to the uniform distribution on $1, \ldots, k$, independently at successive time steps. The transition distribution of the corresponding Markov chain at each time is

$$Q_{(r)}(x, dy) = \frac{1}{k}\sum_{j=1}^{k} Q_j(x, dy) = \frac{1}{k}\sum_{j=1}^{k} p_j(x_{-j}, dy_j)\varepsilon_{x_{-j}}(dy_{-j}).$$

The subscript (r) stands for *random.* The chain is reversible.

Sections 2 and 3 recall probabilistic and statistical results for general Markov chains. The asymptotic variance of the empirical estimator $E_n f$ is described in Section 2. Section 3 determines a lower bound for the asymptotic variance of estimators for πf when the observations come from a Markov chain model. Section 4 shows, for reversible

Markov chains, that for arbitrary f the asymptotic variance of $E_n f$ is reduced if f is replaced by an appropriate conditional expectation, a form of Rao–Blackwellization of the empirical estimator. Section 5 applies Section 3 to Gibbs samplers with deterministic and random sweep and compares the asymptotic variances of the empirical estimator obtained using these samplers. The asymptotic variance under deterministic sweep is about half that under random sweep. Section 6 applies Section 4 to Gibbs samplers and gives lower bounds for the asymptotic variance of estimators for πf. The information bounds coincide for continuous π. If the components of π are not strongly dependent, the empirical estimator is close to efficient under any deterministic sweep. In Sections 7 and 8 we consider Gibbs samplers for random fields on a square lattice and exploit local interactions and symmetries of the random field to improve empirical estimators.

For an introduction to MCMC methods, with emphasis on convergence diagnostics, we refer to the monographs by Robert (1996), Gamerman (1997) and Robert and Casella (1999). For a review with applications to probabilistic inference see Neal (1993). MCMC methods for Gibbs fields are described in Brémaud (1999). Applications to Bayesian statistics are discussed in Besag, Green, Higdon and Mengersen (1996) and in the monograph by Gilks, Richardson and Spiegelhalter (1996).

Preprints on MCMC methods can be downloaded via the MCMC Preprint Service of the Statistical Laboratory at the University of Cambridge. The BUGS software for various samplers is developed jointly by the Biostatistics Unit of the Medical Research Council in Cambridge and by the Imperial College School of Medicine at St Mary's in London; see Spiegelhalter, Thomas and Best (1996). Christian Robert has a web page on convergence diagnostics; see also Mengersen, Robert and Guihenneuc-Jouyaux (1999). The homepages are:

http://www.statslab.cam.ac.uk/∼mcmc/,
http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml,
http://www.crest.fr/pageperso/ls/robert/robert.htm.

2 The asymptotic variance of empirical estimators for Markov chains

In this section we consider observations $X^0, \dots, X^n$ from an Markov chain on an arbitrary state space E, with transition distribution $Q(x, dy)$. See Tierney (1996) for an introduction to Markov chains on general state spaces. We give conditions under which the empirical estimator

$$E_n f = \frac{1}{n} \sum_{i=1}^{n} f(X^i)$$

is asymptotically normal, and describe its asymptotic variance in various ways. Applications to MCMC methods will be given in later sections.

As usual, we write

$$\pi \otimes Q(dx, dy) = \pi(dx) Q(x, dy).$$

For functions $f(x)$ and $h(x,y)$, we write

$$\begin{aligned} Q_x f = Q(x,f) &= \int Q(x,dy)f(y), \\ Q_x h = Q(x,h) &= \int Q(x,dy)h(x,y), \\ \pi \otimes Qh &= \int \pi(dx) \int Q(x,dy)h(x,y). \end{aligned}$$

Then $(Qf)(x) = Q_x f$ defines an operator on $L_2(\pi)$. We assume that the chain is positive Harris recurrent, with invariant distribution $\pi(dx)$, and that it is *V-uniformly ergodic*, i.e., $V : E \to [1,\infty)$ and

$$\sup_x \sup_{|v|\le V} \frac{|Q_x^r v - \pi v|}{V(x)} \to 0 \quad \text{for } r \to \infty.$$

We refer to Meyn and Tweedie (1993) for these concepts. Under these assumptions, if $f^2 \le V$, the empirical estimator $E_n f$ is asymptotically normal. The asymptotic variance is described as follows. Introduce the *potential* U by

$$U_x f = \sum_{r=0}^{\infty} Q_x^r f \quad \text{if } \pi f = 0. \tag{2.1}$$

Define the operator A by centering U conditionally given x,

$$Af(x,y) = U_y(f - \pi f) - Q_x U(f - \pi f) = \sum_{r=0}^{\infty} (Q_y^r f - Q_x^{r+1} f). \tag{2.2}$$

The empirical estimator admits a *martingale approximation*

$$n^{1/2}(E_n f - \pi f) = n^{-1/2} \sum_{i=1}^{n} Af(X^{i-1}, X^i) + o_P(1). \tag{2.3}$$

The approximation is due to Gordin (1969). Write

$$H = \{h(x,y) : h \in L_2(\pi \otimes Q), \quad Q_x h = 0 \quad \text{for } x \in E\}. \tag{2.4}$$

Under the stationary distribution of the chain, $h(X^{i-1}, X^i)$ is a martingale increment. For $h \in H$ we have a martingale central limit theorem,

$$n^{-1/2} \sum_{i=1}^{n} h(X^{i-1}, X^i) \Rightarrow (\pi \otimes Q\, h^2)^{1/2} \cdot N, \tag{2.5}$$

where N is a standard normal random variable, and convergence is in distribution. See e.g. Meyn and Tweedie (1993, Chapter 17). Note that Af is in H. Hence the term $Af(X^{i-1}, X^i)$ is a martingale increment. From the martingale approximation (2.3) and the central limit theorem (2.5) it follows that the empirical estimator $E_n f$ is asymptotically normal with variance

$$\begin{aligned}\pi \otimes Q(Af)^2 &= \pi\Big(U(f-\pi f)\Big)^2 - \pi\Big(QU(f-\pi f)\Big)^2 \\ &= \pi(f-\pi f)^2 + 2\sum_{r=1}^{\infty} \pi\Big((f-\pi f)\cdot Q^r(f-\pi f)\Big). \qquad (2.6)\end{aligned}$$

Suppose that the Markov chain is reversible. This means that Q is in detailed balance (1.1) with π. Then the asymptotic variance of the empirical estimator can be written in the following way; compare Mira and Geyer (1999). By Theorem 2.1 of Roberts and Rosenthal (1997) and the V-uniform ergodicity of Q, the transition distribution Q is $L_2(\pi)$-*geometrically ergodic*: There are constants $\rho < 1$ and $C < \infty$ such that

$$\sup_{\pi f^2 \le 1} \pi(Q^r f - \pi f)^2 \le C\rho^r.$$

Further, detailed balance is equivalent to selfadjointness of Q as an operator on $L_2(\pi)$,

$$\pi(f \cdot Qg) = \pi(Qf \cdot g) \quad \text{for } f, g \in L_2(\pi). \qquad (2.7)$$

Write $I(x, dy) = \varepsilon_x(dy)$ for the identity kernel. The *spectrum* σ of Q is the set of λ such that $\lambda I - Q$ is not invertible. It is a nonempty closed subset of $[-1, 1]$. Let

$$L_{2,0}(\pi) = \{f \in L_2(\pi) : \pi f = 0\}.$$

By the spectral theorem, there is a unique *spectral measure* M on Borel sets of σ such that

$$Qf = \int \lambda M(d\lambda) f \quad \text{for } f \in L_{2,0}(\pi).$$

See e.g. Conway (1990, Theorem IX.2.2). Introduce

$$M_f(d\lambda) = \pi(f \cdot M(d\lambda) f) \quad \text{for } f \in L_{2,0}(\pi). \qquad (2.8)$$

For $f \in L_2(\pi)$, the asymptotic variance of the empirical estimator $E_n f$ can be written as

$$\int_{-1}^{1} \frac{1+\lambda}{1-\lambda} M_{f-\pi f}(d\lambda); \qquad (2.9)$$

see Kipnis and Varadhan (1986). We refer to Mira and Geyer (1999) for an exposition of these results.

As noted in the Introduction, it is reasonable to judge an MCMC sampler by the asymptotic variance of empirical estimators. Let $P(x, dy)$ and $Q(x, dy)$ be transition distributions of reversible Markov chains with common invariant distribution $\pi(dx)$. Write $v(f, P)$ for the asymptotic variance of the empirical estimator $E_n f$ if the Markov chain is generated by P. Mira and Geyer (1999) say that P is *at least as efficient as* Q if

$$v(f,P) \leq v(f,Q) \quad \text{for all } f \in L_2(\pi).$$

They show that this *efficiency ordering* is equivalent to *covariance ordering*,

$$\pi(f \cdot Pf) \leq \pi(f \cdot Qf) \quad \text{for all } f \in L_{2,0}(\pi).$$

Tierney (1998) says that P *dominates* Q *off the diagonal* if

$$P(x,B) \geq Q(x,B) \quad \text{for } B \text{ not containing } x, \text{ and for } \pi\text{-a.a. } x.$$

He proves that in this case P is at least as efficient as Q, and $Q-P$ is positive definite on $L_2(\pi)$. Domination off the diagonal is a strong ordering. If the probability of staying in x is zero under Q, then we must have $P(x,\cdot) = Q(x,\cdot)$. Domination off the diagonal was introduced by Peskun (1973) for discrete state space; he also proved that the property implies efficiency ordering. Mira and Tierney (1999) show that, given any independence Metropolis algorithm, it is possible to construct a slice sampler that dominates it off the diagonal.

Suppose that the state space E is finite, with N elements. Let $1 = \lambda_1$ and $1 > \lambda_2 \geq \cdots \geq \lambda_N$ be the eigenvalues of Q, and $e_1 = 1, e_2, \ldots, e_N$ the corresponding eigenvectors, with $\pi e_r = 1$ for all r. For all f,

$$\pi((f - \pi f) \cdot e_1) = \pi(f - \pi f) = 0$$

and

$$f - \pi f = \sum_{r=2}^{N} \pi((f - \pi f) \cdot e_r) e_r.$$

If the chain is reversible, we have

$$M_f(d\lambda) = \sum_{r=2}^{N} \varepsilon_{\lambda_r}(d\lambda)(\pi(f \cdot e_r))^2 \quad \text{for } f \text{ with } \pi f = 0, \tag{2.10}$$

and the asymptotic variance (2.9) of $E_n f$ is

$$\sum_{r=2}^{N} \frac{1+\lambda_r}{1-\lambda_r} \Big(\pi\big((f - \pi f) \cdot e_r\big)\Big)^2. \tag{2.11}$$

We refer to Frigessi, Hwang and Younes (1992) and Green and Han (1992).

A large spectral gap $1 - \lambda_2$ entails a fast rate of convergence of the Markov chain to stationarity. On the other hand, the asymptotic variance (2.11) of the empirical estimator involves all eigenvalues and is small if $\lambda_2, \ldots, \lambda_N$ are small and negative.

Suppose P and Q are reversible transition matrices with common invariant probability vector π and eigenvalues $1 = \lambda_{1P} > \lambda_{2P} \geq \cdots \geq \lambda_{NP}$ and $1 = \lambda_{1Q} > \lambda_{2Q} \geq \cdots \geq$

λ_{NQ}. Mira and Geyer (1999) note that if $Q - P$ is positive, then $\lambda_{rP} \leq \lambda_{rQ}$ for all r. This follows from the Courant–Fisher minimax representation

$$\lambda_{r+1,Q} = \min_{g_1,\ldots,g_r} \max_{\substack{f:\pi(f\cdot g_s)=0\\ s=1,\ldots,r}} \frac{\pi(f\cdot Qf)}{\pi(f\cdot f)},$$

where the minimum is taken over all vectors $g_1, \ldots, g_r$. For this representation see e.g. Horn and Johnson (1985, Theorem 4.2.11).

For Monte Carlo methods based on i.i.d. realizations $X^0, \ldots, X^n$ from π, a well-known variance reduction method consists in generating *antithetic variables* $Y^0, \ldots, Y^n$ with the same distribution as $X^0, \ldots, X^n$ but negative correlation between $f(X^i)$ and $f(Y^i)$. Then the empirical estimator $\frac{1}{2n}\sum_{i=1}^n (f(X^i) + f(Y^i))$ has smaller asymptotic variance than the usual empirical estimator based on $2n$ realizations. Similar results hold for MCMC data; see in particular Frigessi, Gåsemyr and Rue (2000) for Gibbs samplers.

3 Efficient estimation for Markov chain models

In this section we determine a lower bound for the asymptotic variance of estimators for πf when the observations $X^0, \ldots, X^n$ come from a Markov chain on an arbitrary state space E. For a review of efficient estimation of functionals on Markov chain models we refer to Wefelmeyer (1999). The variance bound is based on a nonparametric version of Hájek's (1970) convolution theorem. It requires the model to be locally asymptotically normal in the following sense.

Let Θ be a possibly infinite-dimensional set, the *parameter space*. A Markov chain model is described by a family Q_ϑ, $\vartheta \in \Theta$, of transition distributions on the state space E. Fix $\vartheta \in \Theta$ such that the Markov chain corresponding to $Q = Q_\vartheta$ is positive Harris recurrent with invariant distribution $\pi = \pi_\vartheta$. Assume that Θ is smooth in the following sense. There are a linear space M, the *tangent space* of Θ at ϑ, and a linear map $D : M \to H$, with H defined in (2.4), and for each $m \in M$ there is a sequence ϑ_{nm} such that $Q_{nm} = Q_{\vartheta_{nm}}$ is *Hellinger differentiable* with *derivative* Dm,

$$\int Q(x, dy) \left(\left(\frac{dQ_{nm}}{dQ}(x, y) \right)^{1/2} - 1 - \frac{1}{2} n^{-1/2} Dm(x, y) \right)^2 \leq n^{-1} r_n(x), \tag{3.1}$$

where r_n decreases to 0 pointwise and is π-integrable for large n. This version of Hellinger differentiability is due to Höpfner, Jacod and Ladelli (1990).

Write P_n and P_{nm} for the joint distribution of $X^0, \ldots, X^n$ under Q and Q_{nm}, respectively. As in Höpfner (1993) we have a nonparametric version of *local asymptotic normality* for the likelihood ratio. For $m \in M$,

$$\log \frac{dP_{nm}}{dP_n} = n^{-1/2} \sum_{i=1}^n Dm(X^{i-1}, X^i) - \frac{1}{2} \pi \otimes Q(Dm)^2 + o_{P_n}(1),$$

$$n^{-1/2}\sum_{i=1}^{n} Dm(X^{i-1},X^i) \Rightarrow \big(\pi\otimes Q(Dm)^2\big)^{1/2}\cdot N \quad \text{under } P_n, \tag{3.2}$$

where N is a standard normal random variable. The last result is just the central limit theorem (2.5). Local asymptotic normality for Markov chains was first proved by Roussas (1965) for parametric models, and by Penev (1991) for nonparametric models.

The norm $\pi\otimes Q(Dm)^2$ induces an inner product $\pi\otimes Q(Dm\cdot Dm')$ on M. Consider $\pi_\vartheta f$ as a functional of ϑ. The functional is *differentiable* at ϑ with *gradient* g if $g\in H$ and

$$n^{1/2}(\pi_{nm}f-\pi f)\to \pi\otimes Q(Dm\cdot g) \quad \text{for } m\in M. \tag{3.3}$$

The *canonical gradient* $g_0 = Dm_0$ is the projection of g onto DM. The function m_0 is uniquely determined by

$$n^{1/2}(\pi_{nm}f-\pi f)\to \pi\otimes Q(Dm\cdot Dm_0) \quad \text{for } m\in M. \tag{3.4}$$

The canonical gradient is not always easy to calculate. Sometimes it is easier to find another gradient first (which may, in turn, be canonical in a larger model). One can then try to project that gradient into the tangent space. One such gradient is the function Af defined in (2.2), as follows from the perturbation expansion of Kartashov (1985a, b) and (1996, Section 4.2), and using $QDm=0$: For $m\in M$,

$$n^{1/2}(\pi_{nm}f-\pi f)\to \int \pi(dx)Q(x,dy)Dm(x,y)U_y f = \pi\otimes Q(Dm\cdot Af). \tag{3.5}$$

Another approach to calculating the canonical gradient is possible when the tangent space M comes equipped with some inner product, as is usually the case. This approach is used in Section 6 for the Gibbs sampler. We denote the inner product by $(m,m')_M$. It may then be possible to find the gradient m_M with respect to this inner product,

$$n^{1/2}(\pi_{nm}f-\pi f)\to (m,m_M)_M \quad \text{for } m\in M.$$

Comparing with (3.4), we see that the canonical gradient Dm_0 is now determined by

$$(m,m_M)_M = \pi\otimes Q(Dm\cdot Dm_0) \quad \text{for } m\in M.$$

If D has an adjoint $D^*: H\to M$, we have

$$\pi\otimes Q(Dm\cdot Dm') = (m, D^*Dm')_M \quad \text{for } m,m'\in M.$$

Hence, if D^*D has an inverse, the canonical gradient is Dm_0 with

$$m_0 = (D^*D)^{-1}m_M.$$

In fact, one can avoid calculating D^*. It suffices to find an operator C such that

$$\pi\otimes Q(Dm\cdot Dm') = (m, Cm')_M \quad \text{for } m,m'\in M.$$

Then the canonical gradient is Dm_0 with $m_0 = C^{-1}m_M$. It may happen that C^{-1} is difficult to determine but that C can be written as a perturbation of the identity operator, say $C = I - B$. If B is not too large, C^{-1} may then be written as the von Neumann series $C^{-1} = \sum_{r=0}^{\infty} B^r$, and $m_0 = \sum_{r=0}^{\infty} B^r m_M$.

Efficient estimators for πf are characterized as follows. Call an estimator T_n *regular* for πf with *limit* L if

$$n^{1/2}(T_n - \pi_{nm} f) \Rightarrow L \quad \text{under } P_{nh} \text{ for } m \in M.$$

Call T_n *asymptotically linear* with *influence function* h if $h \in H$ and

$$n^{1/2}(T_n - \pi f) = n^{-1/2} \sum_{i=1}^{n} h(X^{i-1}, X^i) + o_{P_n}(1).$$

By a result of LeCam, see Bickel, Klaassen, Ritov and Wellner (1998, Section A.9), an asymptotically linear estimator is regular if and only if its influence function is a gradient. The martingale approximation (2.3) says that the empirical estimator $E_n f$ is asymptotically linear with influence function Af. Since Af is a gradient by (3.5), the empirical estimator is regular.

The convolution theorem of Hájek (1970) in the version of Pfanzagl and Wefelmeyer (1982, Theorem 9.3.1), or see Bickel, Klaassen, Ritov and Wellner (1998, p. 63, Theorem 2), applied now for Markov chains, says that if T_n is regular with limit L, then

$$\begin{aligned} &\Big(n^{-1/2} \sum_{i=1}^{n} g_0(X^{i-1}, X^i),\ n^{1/2}(T_n - \pi f) - n^{-1/2} \sum_{i=1}^{n} g_0(X^{i-1}, X^i)\Big) \\ &\Rightarrow \Big((\pi \otimes Q\, g_0^2)^{1/2} \cdot N, M\Big) \quad \text{under } P_n, \end{aligned} \tag{3.6}$$

with M independent of N, and g_0 the canonical gradient. In particular,

$$L = (\pi \otimes Q\, g_0^2)^{1/2} \cdot N + M \quad \text{in distribution.}$$

For every $a > 0$ we have $P(-a < cN < a) \geq P(-a < cN + M < a)$. This justifies calling T_n *efficient* if

$$L = (\pi \otimes Q\, g_0^2)^{1/2} \cdot N \quad \text{in distribution.}$$

It follows from (3.6) that T_n is efficient if and only if it is asymptotically linear with influence function equal to the canonical gradient,

$$n^{1/2}(T_n - \pi f) = n^{-1/2} \sum_{i=1}^{n} g_0(X^{i-1}, X^i) + o_{P_n}(1).$$

An efficient estimator is asymptotically normal with variance $\pi \otimes Q\, g_0^2$. We call this variance the *asymptotic variance bound.* In Section 6 we calculate the asymptotic variance bound for Gibbs samplers with random and deterministic sweep.

4 Improving empirical estimators by conditioning

To begin let $X^1, \dots, X^n$ be *independent* and identically distributed as π, and let f be a π-square-integrable function. The empirical estimator $E_n f = \frac{1}{n}\sum_{i=1}^n f(X^i)$ for the expectation πf is asymptotically normal with variance $\pi(f-\pi f)^2$. Now replace $f(X^i)$ by a conditional expectation $E_\pi(f(X^i)|h(X^i))$, where h is some function. Then the *Rao–Blackwellized* empirical estimator

$$E_n E_\pi(f|h) = \frac{1}{n}\sum_{i=1}^n E_\pi(f(X^i)|h(X^i)) \tag{4.1}$$

has asymptotic variance $\pi\Big(E_\pi(f|h)-\pi f\Big)^2$. The Rao–Blackwell theorem says that it is smaller than the asymptotic variance of $E_n f$,

$$\pi\Big(E_\pi(f|h)-\pi f\Big)^2 \le \pi(f-\pi f)^2. \tag{4.2}$$

Of course, the "estimator" $E_n E_\pi(f|h)$ can be used only if $E_\pi(f|h)$ does not depend on π, e.g., when h is sufficient.

Recently, there has been considerable interest in developing versions of the Rao–Blackwell theorem in the context of stochastic simulation, and for Markov chain Monte Carlo (MCMC) in particular. See Casella and Robert (1996) and the references cited therein. Early references are Kalos and Whitlock (1986, Section 4.2) and Pearl (1987). See also Neal (1993, Section 6.3). Gelfand and Smith (1990, 1991) consider i.i.d. runs of a Gibbs sampler. In the empirical estimator based on the final value of each run, they replace f by a conditional expectation under π. For long runs, the final values are distributed approximately according to π, so the classical Rao–Blackwell theorem (4.2) implies that the variance is reduced. Single runs of Markov chains are studied in the following references. Liu, Wong and Kong (1994) consider auxiliary variable algorithms of the form $Q(x_1, dy_1) = \int p_2(x_1, x_2) p_1(x_2, dy_1)$ and the Rao–Blackwellized empirical estimator

$$E_n p_1 f = \frac{1}{n}\sum_{i=1}^n \int p_1(X_2^i, dy_1) f(y_1).$$

They prove that the variance is always reduced. Casella and Robert (1996) propose some types of Rao–Blackwellization for the Metropolis–Hastings algorithm. Their approach is to integrate out some or all of the uniform random variables involved.

Let $X^0, \dots, X^n$ be realizations of an arbitrary Markov chain with transition distribution $Q(x, dy)$ and invariant distribution $\pi(dx)$. Assume that the Markov chain is positive Harris recurrent and V-uniformly ergodic, and that $f^2 \le V$. Then the Rao–Blackwellized empirical estimator $E_n E_\pi(f|h)$ is asymptotically normal, and by (2.6) its asymptotic variance is

$$\pi\Big(E_\pi(f|h)-\pi f\Big)^2 + 2\sum_{r=1}^\infty \pi\Big((E_\pi(f|h)-\pi f)\cdot Q^r(E_\pi(f|h)-\pi f)\Big).$$

Geyer (1995) gives necessary and sufficient conditions for $E_n E_\pi(f|h)$ to have smaller asymptotic variance than $E_n f$ for all f simultaneously, and points out that these conditions are unlikely to be satisfied in practice.

McKeague and Wefelmeyer (2000) suggest a different version of Rao–Blackwellization. Rather than conditioning $f(X^i)$ on a function $h(X^i)$, they condition on the previous value of the chain. The function $f(x)$ in the empirical estimator $E_n f$ is replaced by $Q(x,f) = E(f(X^i)|X^{i-1} = x)$. Their Rao–Blackwellized empirical estimator is therefore

$$E_n Qf = \frac{1}{n}\sum_{i=1}^{n} Q(X^i, f).$$

Rao–Blackwellization can be repeated, leading to the estimator

$$E_n Q^k f = \frac{1}{n}\sum_{i=1}^{n} Q^k(X^i, f).$$

For *reversible* chains, this Rao–Blackwellization reduces the asymptotic variance simultaneously for all f. Schmeiser and Chen (1991) prove this result for the Hit-and-Run algorithm proposed by Belisle, Romeijn and Smith (1993). The result does not hold, in general, for non-reversible chains.

Theorem 1. *Let $X^0, \dots, X^n$ be realizations of a Markov chain which is positive Harris recurrent, V-uniformly ergodic and reversible. For $f \in L_2(\pi)$, the asymptotic variance of $E_n Q^k f$ is less than that of $E_n f$, and the variance reduction is*

$$\sum_{j=0}^{k-1} \pi\Big((I+Q)Q^j(f-\pi f)\Big)^2.$$

The asymptotic variance of $E_n Q^k f$ tends to zero as k goes to infinity.

The proof is simple. By (2.6) and selfadjointness (2.7), the asymptotic variance of $E_n Q^k f$ is

$$\begin{aligned}
&\pi\Big(Q^k(f-\pi f)\cdot\Big(I + 2\sum_{r=1}^{\infty} Q^r\Big)Q^k(f-\pi f)\Big)\\
&= \pi\Big((f-\pi f)\cdot Q^{2k}(f-\pi f)\Big) + 2\sum_{r=2k+1}^{\infty} \pi\Big((f-\pi f)\cdot Q^r(f-\pi f)\Big). \qquad (4.3)
\end{aligned}$$

The asymptotic variance (4.3) of $E_n Q^k f$ is obtained from the asymptotic variance (2.6) of $E_n f$ by omitting the term $\pi(f-\pi f)^2$, the terms of order $r = 1, \dots, 2k-1$, and half the term of order $2k$. This implies the second part of Theorem 1.

The difference between (2.6) and (4.3) can be written as

$$\pi\Big((f-\pi f)\cdot\sum_{j=0}^{k-1}(I+Q)^2 Q^{2j}(f-\pi f)\Big).$$

This implies the first part of Theorem 1.

Theorem 1 can also be expressed in terms of the spectral measure. By (2.9) and (2.8), the asymptotic variance of $E_n Q^k f$ is

$$\int_{-1}^{1} \frac{1+\lambda}{1-\lambda} \lambda^{2k} \pi\Big((f-\pi f)\cdot E(d\lambda)(f-\pi f)\Big),$$

and the variance reduction over $E_n f$ is

$$\int_{-1}^{1} \frac{1+\lambda}{1-\lambda} (1-\lambda^{2k}) \pi\Big((f-\pi f)\cdot E(d\lambda)(f-\pi f)\Big).$$

Suppose that the state space E is finite, with N elements. Let $1=\lambda_1>\lambda_2\geq\cdots\geq\lambda_N$ be the eigenvalues of Q, and $e_1=1, e_2, \dots, e_N$ the corresponding eigenvectors, with $\pi e_r = 1$. Then the variance reduction is

$$\sum_{r=2}^{N} \frac{1+\lambda_r}{1-\lambda_r} (1-\lambda_r^{2k}) \Big(\pi\Big((f-\pi f)\cdot e_r\Big)\Big)^2.$$

McKeague and Wefelmeyer (2000) illustrate Theorem 1 with simulations for the Ising model and the Gibbs sampler, and for f the r-th nearest neighbor correlation.

5 Asymptotic variance of empirical estimators for Gibbs samplers

Let $E = E_1 \times \ldots \times E_k$ be a product of measurable spaces, with product σ-field, and let $\pi(dx)$ be a distribution on E. Gibbs samplers successively use the transition distributions $Q_j(x,dy) = p_j(x_{-j}, dy_j)\varepsilon_{x_{-j}}(dy_{-j})$, with $p_j(x_{-j}, dx_j)$ the one-dimensional conditional distribution under π of x_j given x_{-j}, introduced in (1.3). For a function $f(x)$ we have by definition of Q_j,

$$Q_j(x,f) = \int Q_j(x,dy) f(y) = \int p_j(x_{-j}, dx_j) f(x_{-j}, x_j) = p_j(x_{-j}, f).$$

In particular, $Q_j(x,dy)$ does not depend on x_j. Hence Q_j is *idempotent*,

$$Q_j^2 = Q_j. \tag{5.1}$$

This means that Q_j is a *projection operator* on $L_2(\pi)$. Indeed, we can write $L_2(\pi)$ as the orthogonal sum of two subspaces, one consisting of functions f with $Q_j f = 0$, the other of functions $f(x)$ not depending on x_j, and Q_j is the projection on the second subspace along the first. Therefore,

$$\pi(f\cdot Q_j f') = \pi(Q_j f \cdot Q_j f') \quad \text{for } f, f' \in L_2(\pi). \tag{5.2}$$

Relation (5.2) implies that Q_j, as an operator on $L_2(\pi)$, is *positive*,

$$\pi(f \cdot Q_j f) = \pi(Q_j f \cdot Q_j f) \geq 0 \quad \text{for } f \in L_2(\pi), \tag{5.3}$$

and *selfadjoint*,

$$\pi(f \cdot Q_j f') = \pi(Q_j f \cdot Q_j f') = \pi(Q_j f \cdot f') \quad \text{for } f, f' \in L_2(\pi). \tag{5.4}$$

The last relation is seen to be equivalent to detailed balance,

$$\pi(dx) Q_j(x, dy) = \pi(dy) Q_j(y, dx). \tag{5.5}$$

This, in turn, implies again that Q_j has invariant law π.

The Gibbs sampler for π with deterministic (and cyclic) sweep has transition distribution Q_j at time $i = (q-1)k + j$; the subchain of full sweeps has transition distribution $Q_1 \cdots Q_k$. Let $X^0, \dots, X^n$ be realizations from the Gibbs sampler with deterministic sweep, with n a multiple of k, say $n = pk$. We want to estimate the expectation πf of a function f. The most common estimator for πf is the empirical estimator based on the *subchain* $X^k, \dots, X^{pk}$,

$$E_n^k f = \frac{1}{p} \sum_{q=1}^{p} f(X^{qk}).$$

The empirical estimator based on the *full* chain $X^1, \dots, X^n$ is

$$E_n f = \frac{1}{n} \sum_{i=1}^{n} f(X^i) = \frac{1}{k} \sum_{j=1}^{k} E_n^j f$$

with

$$E_n^j f = \frac{1}{p} \sum_{q=1}^{p} f_{\leq j}(X^{(q-1)k}, X^{qk})$$

and

$$f_{\leq j}(x, y) = f(y_{\leq j}, x_{>j}).$$

The estimator $E_n^j f$ is based on the subchain $X^j, X^{k+j}, \dots, X^{(p-1)k+j}$.

To fix things, by *asymptotic distribution* of an estimator T_n we will mean the asymptotic distribution of $n^{1/2}(T_n - \pi f)$, even though standardizing by $p^{1/2}$ rather than $n^{1/2}$ is more common for the empirical estimator $E_n^j f$.

Greenwood, McKeague and Wefelmeyer (1998) calculate the asymptotic variance of $E_n^j f$ and $E_n f$, using the form (2.6) of the asymptotic variance in the central limit theorem for Markov chains.

Theorem 2. *Assume that the Gibbs sampler for π with deterministic sweep is positive Harris recurrent and the subchains are V-uniformly ergodic, and that $f^2 \leq V$. Then the empirical estimator $E_n^j f$ is asymptotically normal with variance*

$$\sigma_j^2 = k\pi(f - \pi f)^2 + 2k \sum_{r=1}^{\infty} \pi\Big((f - \pi f) \cdot p_j^{\mathrm{cycl}\, rk}(f - \pi f)\Big),$$

where $p_j^{\mathrm{cycl}\,r} = p_j p_{j+1} \cdots p_k p_1 p_2 \cdots$ *with* r *terms.*

Theorem 3. *Under the assumptions of Theorem 2, the empirical estimator* $E_n f$ *is asymptotically normal with variance*

$$\sigma_d^2 = \pi(f-\pi f)^2 + 2\sum_{r=1}^{\infty} \frac{1}{k} \sum_{j=1}^{k} \pi\Big((f-\pi f)\cdot p_j^{\mathrm{cycl}\,r}(f-\pi f)\Big).$$

Because the empirical estimator $E_n^k f$ based on the subchain of full sweeps is often used in practice, we have included the description of its asymptotic variance in Theorem 2. However, we do not recommend this estimator; the simulations in Figure 1 below show that $E_n^k f$ can be considerably worse than $E_n f$. This is true even when π has only two components; see Greenwood, McKeague and Wefelmeyer (1996).

Theorem 4. *Assume that the Gibbs sampler for* π *with random sweep is positive Harris recurrent and* V*-uniformly ergodic, and that* $f \in L_2(\pi)$*. Then the empirical estimator* $E_n f$ *is asymptotically normal with variance*

$$\begin{aligned}\sigma_{(r)}^2 &= \pi(f-\pi f)^2 + 2\sum_{r=1}^{\infty} \pi\Big((f-\pi f)\cdot Q_{(r)}^r(f-\pi f)\Big)\\ &= \pi(f-\pi f)^2 + 2\sum_{r=1}^{\infty} \frac{1}{(k-1)^r} \sum_{\substack{j_1,\dots,j_r=1\\ j_i\neq j_{i+1}}}^{k} \pi\Big((f-\pi f)\cdot p_{j_1}\cdots p_{j_r}(f-\pi f)\Big).\end{aligned}$$

The second summation in $\sigma_{(r)}^2$ contains $k(k-1)^{r-1}$ terms, each being an r-order autocovariance of the form $\pi\Big((f-\pi f)\cdot p_{j_1}\cdots p_{j_r}(f-\pi f)\Big)$. From Theorem 3, the r-order term in $\sigma_{(d)}^2$ is an average of k of these r-order autocovariances, those of the form $\pi\Big((f-\pi f)\cdot p_j^{\mathrm{cycl}\,rk}(f-\pi f)\Big)$. One might expect that $\sigma_{(r)}^2 \approx (k/(k-1))\sigma_{(d)}^2$, or that $\sigma_{(r)}^2$ is slightly larger than $\sigma_{(d)}^2$. Such a result holds if one considers a random sweep without repetition, see Fishman (1996, Theorem 8). Greenwood, McKeague and Wefelmeyer (1996) argue, however, that $\sigma_{(r)}^2$ can be up to twice as large as $\sigma_{(d)}^2$, even if k is large. This is also seen in simulations; see Figure 1. The reason is that the higher-order terms in $\sigma_{(r)}^2$ can decay more slowly than those in $\sigma_{(d)}^2$. This is easily seen in the special case of *independent* components. Then $\pi\Big((f-\pi f)\cdot p_j^{\mathrm{cycl}\,rk}(f-\pi f)\Big)$ vanishes for $s \geq k$, because integration of $f(x)$ cyclically over k components gives πf. However, $\pi\Big((f-\pi f)\cdot p_{j_1}\cdots p_{j_r}(f-\pi f)\Big)$ vanishes only if all k components are present among $j_1 \dots, j_s$. Also, if some of the j_i are equal, fewer than r components are integrated out, so $\pi\Big((f-\pi f)\cdot p_{j_1}\cdots p_{j_r}(f-\pi f)\Big)$ is larger than any $\pi\Big((f-\pi f)\cdot p_j^{\mathrm{cycl}\,rk}(f-\pi f)\Big)$ "covering" $j_1 \dots, j_r$.

6 Asymptotic variance bounds for Gibbs samplers

In Section 5 we have determined the asymptotic variance of the empirical estimator for an expectation πf under the Gibbs sampler with random and deterministic sweep. In this section we consider another criterion by which MCMC methods can be judged: How much information about πf is contained in the simulated values $X^0, \dots, X^n$, given the knowledge that a particular sampler was used to generate them? In particular: What fraction of the information is exploited by the empirical estimator? It is assumed that no information about π itself is made available to the statistician, apart from the link between π and the transition distribution of the observed Markov chain. Of course, π is known in principle, and part of that knowledge can sometimes be exploited to improve upon the empirical estimator. An example is Rao–Blackwellization, see Section 4. If π is a random field on a lattice and the interactions between the sites are known to be local, improved estimators are described in Section 7. Symmetries of π are exploited Section 8.

We keep the setting of Section 5. Consider first the Gibbs sampler with *deterministic* sweep. The subchain of full sweeps has transition distribution $Q_{(d)} = Q_1 \cdots Q_k$, see (1.4). It is parametrized by π. To determine the information bound of (regular) estimators of πf, we must prove that the model is locally asymptotically normal (3.2). A perturbation of π is of the form

$$\pi_{nk}(dx) = \pi(dx)(1 + n^{-1/2}k(x)),$$

with k in

$$M = \{m(x) : m \text{ measurable, bounded, } \ \pi m = 0\}. \tag{6.1}$$

Write $p_{j,nm}(x_{-j}, dx_j)$ for the one-dimensional conditional distribution under $\pi_{nm}(dx)$ of x_j given x_{-j}. The effect of the perturbation of π on p_j is easily obtained as follows (Greenwood, McKeague and Wefelmeyer, 1998, Lemma 1): Uniformly in x,

$$p_{j,nm}(x_{-j}, dx_j) = p_j(x_{-j}, dx_j)\Big(1 + n^{-1/2} m_j(x) + O(n^{-1})\Big)$$

with $m_j(x) = m(x) - Q_j(x, m) = m(x) - p_j(x_{-j}, m)$.

Write $Q_{(d)nm}$ for the transition distribution (1.4) of the Gibbs sampler for π_{nm} with deterministic sweep,

$$Q_{(d)nm}(x, dy) = (Q_{1,nm} \cdots Q_{k,nm})(x, dy) = \prod_{j=1}^{k} p_{j,nm}(y_{<j}, x_{>j}, dy_j).$$

It follows easily that $Q_{(d)nm}$ is obtained by perturbing $Q_{(d)}$ as follows: Uniformly in x and y,

$$Q_{(d)nm}(x, dy) = Q_{(d)}(x, dy)\Big(1 + n^{-1/2}(D_{(d)}m(x, y) + O(n^{-1})\Big) \tag{6.2}$$

with

$$D_{(d)}m(x,y) = \sum_{j=1}^{k} m_j(y_{\leq j}, x_{>j}). \tag{6.3}$$

Relation (6.2) implies that $Q_{(d)nm}$ is Hellinger differentiable (3.1) with derivative $D_{(d)}m$.

Write $P_{(d)n}$ for the joint distribution of $X^0, X^k, \dots, X^{pk}$ if π is true, and $P_{(d)nm}$ if π_{nm} is true. If the Gibbs sampler for π with deterministic sweep is positive Harris recurrent, we obtain local asymptotic normality (3.2) of the form

$$\log dP_{(d)nm}/dP_{(d)n} = n^{-1/2} \sum_{q=1}^{p} D_{(d)}m(X^{(q-1)k}, X^{qk}) - \frac{1}{2}\pi \otimes Q(D_{(d)}m)^2 + o_{P_{(d)n}}(1). \tag{6.4}$$

The desired minimal asymptotic variance of regular estimators of πf is the squared length of the gradient of πf. To determine this gradient, we note that by definition of π_{nm}, and since $\pi m = 0$,

$$n^{1/2}(\pi_{nm} f - \pi f) = \pi m f = \pi\Big(m \cdot (f - \pi f)\Big) \quad \text{for } m \in M.$$

Comparing this relation with definition (3.4) of the gradient, we see that the canonical gradient is $g_{(d)} = D_{(d)}m_{(d)}$ with $m_{(d)}$ fulfilling

$$\pi \otimes Q(D_{(d)}m \cdot D_{(d)}m_{(d)}) = \pi\Big(m \cdot (f - \pi f)\Big) \quad \text{for } m \in M.$$

The left side can be written

$$\pi\Big(m \cdot (I - Q_{(r)})m_{(d)}\Big);$$

see Greenwood, McKeague and Wefelmeyer (1998, Lemma 2). Surprisingly, this inner product for *deterministic* sweep involves the transition distribution for *random* sweep. Let $\| \, \|_2$ be the operator norm on $L_2(\pi)$, defined by $\|Q\|_2^2 = \sup_{\pi f^2 \leq 1} \pi(Qf)^2$ for an operator Q on $L_2(\pi)$. If $\|Q_{(r)}^t\|_2 < 1$ for some t, we obtain

$$m_{(d)} = (I - Q_{(r)})^{-1}(f - \pi f).$$

Hence the asymptotic variance bound is

$$\pi\Big((f - \pi f) \cdot (I - Q_{(r)})^{-1}(f - \pi f)\Big).$$

After some calculation we arrive at the following result; see Greenwood, McKeague and Wefelmeyer (1998, Theorem 1).

Theorem 5. *Let $f \in L_2(\pi)$, and assume that $\|Q_{(r)}^t\|_2 < 1$ for some t. For the Gibbs sampler with deterministic sweep, the asymptotic variance bound is*

$$\begin{aligned} B_{(d)} &= \pi\Big((f - \pi f) \cdot (I - Q_{(r)})^{-1}(f - \pi f)\Big) \\ &= \pi(f - \pi f)^2 + \sum_{r=1}^{\infty} \frac{1}{(k-1)^r} \sum_{\substack{j_1, \dots, j_r = 1 \\ j_i \neq j_{i+1}}}^{k} \pi\Big((f - \pi f) \cdot p_{j_1} \cdots p_{j_r}(f - \pi f)\Big). \end{aligned}$$

Note that $B_{(d)}$ does not depend on the order of the deterministic sweep; it only depends on π and f.

The result for *random* sweep is more involved. Write $Q_{(r)nm}$ for the transition distribution of the Gibbs sampler with random sweep for π_{nm},

$$Q_{(r)nm}(x, dy) = \frac{1}{k}\sum_{j=1}^{k} Q_{j,nm}(x, dy) = \frac{1}{k}\sum_{j=1}^{k} p_{j,nm}(x_{-j}, dy_j)\varepsilon_{x_{-j}}(dy_{-j}).$$

The perturbation of $Q_{(r)}$ now involves the probabilities of not changing the value when updating a component. The reason is that the transition distribution $Q_j(x, dy)$ is supported by the line through x parallel to the j-th coordinate axis, $\{y : y_{-j} = x_{-j}\}$. Hence the support of $Q_{(r)}(x, dy)$ is contained in the union of the k lines. The supports of the $Q_j(x, dy)$ are disjoint except for the point x, which may be charged by some or all of them. Therefore, to calculate the $Q_{(r)}(x, dy)$-density of $Q_{(r)nm}(x, dy)$, we must treat x separately. We assume that the σ-field on each E_j contains the one-point sets, which will be the case in all applications.

Greenwood, McKeague and Wefelmeyer (1998, Lemma 4) show that for $m \in M$, and uniformly in x and y,

$$Q_{(r)nm}(x, dy) = Q_{(r)}(x, dy)\Big(1 + n^{-1/2}D_{(r)}m(x, y) + O(n^{-1})\Big) \tag{6.5}$$

with

$$\begin{aligned} D_{(r)}m(x, y) &= \sum_{j=1}^{k}(D_j m)(x, y), \\ (D_j m)(x, y) &= \Big(1(y_{-j} = x_{-j}) - \Big(1 - \frac{r_j(x)}{r(x)}\Big)1(y = x)\Big)m_j(x_{-j}, y_j), \\ r_j(x) &= p_j(x_{-j}, \{x_j\}), \quad r(x) = \sum_{j=1}^{k} r_j(x). \end{aligned}$$

Relation (6.5) implies that $Q_{(r)nm}$ is Hellinger differentiable (3.1) with derivative $D_{(r)}m$. Write $P_{(r)n}$ for the joint distribution of $X^0, X^k, \dots, X^{pk}$ if π is true, and $P_{(r)nm}$ if π_{nm} is true. If the Gibbs sampler for π with random sweep is positive Harris recurrent, we obtain local asymptotic normality (3.2) of the form

$$\log dP_{(r)nm}/dP_{(r)n} = n^{-1/2}\sum_{i=1}^{n}(K_{(r)}m)(X^{i-1}, X^i) - \frac{1}{2}\pi \otimes Q(D_{(r)}m)^2 + o_{P_{(r)n}}(1). \tag{6.6}$$

Similarly as for deterministic sweep, the canonical gradient is $g_{(r)} = D_{(r)}m_{(r)}$ with $m_{(r)}$ fulfilling

$$\pi \otimes Q(D_{(r)}m \cdot D_{(r)}m_{(r)}) = \pi\Big(m \cdot (f - \pi f)\Big) \quad \text{for } m \in M.$$

By Greenwood, McKeague and Wefelmeyer (1998, Lemma 5), the left side can be written

$$\pi\Big(m \cdot (I - Q_{(r)} + S)m_{(r)}\Big)$$

with

$$\begin{aligned} Sm &= \frac{1}{k}\sum_{i,j=1}^{k} Q_i(R_{ij}Q_j m), \\ R_{ij}(x) &= \delta_{ij}r_j(x) - r_i(x)r_j(x)/r(x). \end{aligned}$$

If $\|(Q_{(r)} - S)^t\|_2 < 1$ for some t, we obtain

$$m_{(r)} = (I - Q_{(r)} + S)^{-1}(f - \pi f).$$

After some calculation, we arrive at the following result, Greenwood, McKeague and Wefelmeyer (1998, Theorem 2).

Theorem 6. *Let $f \in L_2(\pi)$, and assume that $\|(Q_{(r)} - S)^t\|_2 < 1$ for some t. For the Gibbs sampler with random sweep, the asymptotic variance bound is*

$$\begin{aligned} B_{(r)} &= \pi\Big((f-\pi f)\cdot(I - Q_{(r)} + S)^{-1}(f-\pi f)\Big) \\ &= \pi(f-\pi f)^2 + \sum_{r=1}^{\infty}\pi\Big((f-\pi f)\cdot(Q_{(r)} - S)^r(f-\pi f)\Big). \end{aligned}$$

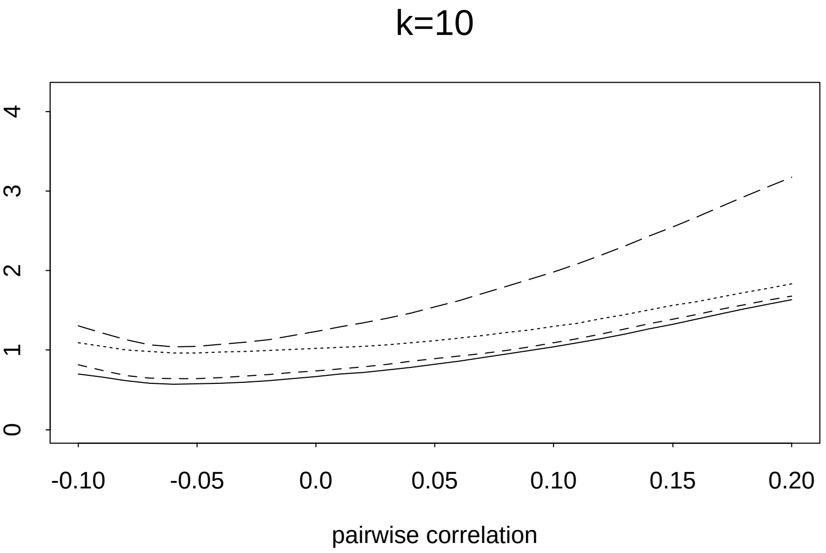
Both S and $I - Q_{(r)}$ are positive operators on $L_2(\pi)$. Write $K \geq L$ if $K - L$ is positive. Then $I - Q_{(r)} + S \geq I - Q_{(r)}$ and therefore $(I - Q_{(r)} + S)^{-1} \leq (I - Q_{(r)})^{-1}$. Thus the variance bound is no larger for random sweep than for deterministic sweep: $B_{(r)} \leq B_{(d)}$.

Suppose that π is continuous in the sense that it is absolutely continuous with respect to the product of its marginals, and the marginals have no atoms. Then $p_j(x_{-j}, dx_j)$ has no atoms for $\pi_{-j}(dx_{-j})$-a.s. x_{-j}. Hence $r_j(x) = p_j(x_{-j}, \{x_j\}) = 0$ for π-a.a. x, and therefore $R_{ij}(x) = 0$ for π-a.a. x, and the operator S reduces to 0. This implies that information bound for random sweep coincides with the information bound for deterministic sweep: $B_{(r)} = B_{(d)}$.

The last term in the asymptotic variance bound $B_{(d)}$ for deterministic sweep appears with a factor 2 in the asymptotic variance $\sigma^2_{(r)}$ of the empirical estimator for random sweep. In most applications, the leading term $\pi(f - \pi f)^2$ of the variances is relatively small. Then $\sigma^2_{(r)}$ is nearly twice as large as $B_{(d)}$. When π is continuous, we have $B_{(d)} = B_{(r)}$. Hence the efficiency of the empirical estimator for random sweep is close to 50%.

As mentioned in Section 5, the asymptotic variance of the empirical estimator is about twice as large for random sweep as for deterministic sweep. This implies that the empirical estimator for deterministic sweep is close to efficient.

We illustrate the results with an exchangeable k-dimensional multivariate normal distribution π in which each component has zero mean and unit variance, and all the pairwise correlations are identical. This example has been widely used in the literature



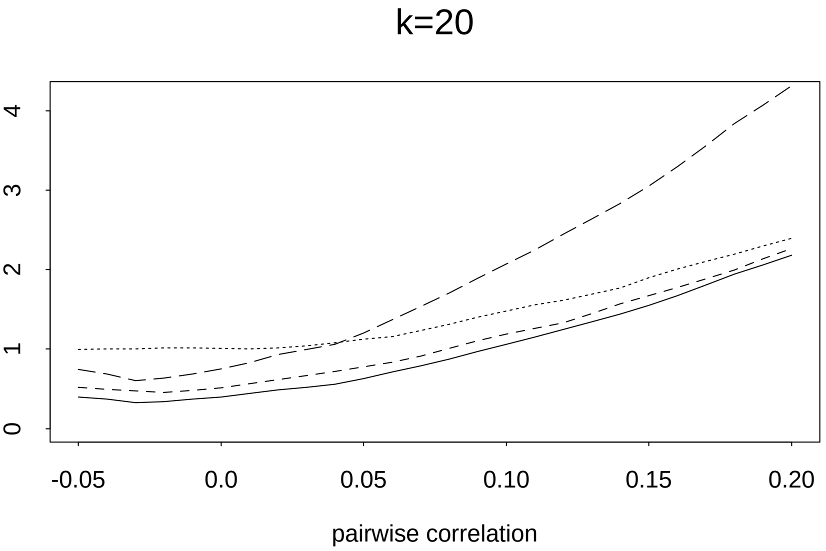


Figure 1: Exchangeable k-dimensional normal example. The information bounds for both random and deterministic sweep (solid line) and the asymptotic variances (in units of $k\pi(f-\pi f)^2$) of the usual empirical estimator $E_n^k f$ under deterministic sweep (dotted line), and the full chain estimator $E_n f$ under deterministic sweep (short dashed line) and random sweep (long dashed line).

for studying convergence rates of Gibbs samplers, see, e.g., Raftery and Lewis (1992, Example 3) and Roberts and Sahu (1997). The function f is taken to be the indicator that the random field exceeds a unit threshold: $f(x) = 1(\max_j x_j > 1)$. The results for 10 and 20 dimensions are shown in Figure 1.

7 Improving empirical estimators for random fields with local interactions

To begin let π be a distribution on a K-dimensional space $E = E_1 \times \cdots \times E_K$, and let $X^1, \ldots, X^n$ be *independent* and identically distributed as π. Let f be a π-square-integrable function. In the nonparametric setting, with nothing known about π, the empirical estimator $E_n f = \frac{1}{n}\sum_{i=1}^n f(X^i)$ is efficient; see Bickel, Klaassen, Ritov and Wellner (1998, Section 3.3). If the components of π are known to be independent, $\pi = \pi_1 \otimes \cdots \otimes \pi_K$, then $E_n f$ is no longer efficient, and a better estimator of πf is the generalized von Mises statistic

$$M_n f = \frac{1}{n^K} \sum_{i_1, \ldots, i_K = 1}^{n} f(X_1^{i_1}, \ldots, X_K^{i_K}).$$

Since it is the expectation of f under the product of the marginal empiricals, $M_n f$ is again efficient if nothing is known about the components $\pi_1, \ldots, \pi_K$; see Levit (1974) and Koshevnik and Levit (1976).

Note that the terms $(X_1^{i_1}, \ldots, X_K^{i_K})$ have law π: They are obtained by mixing the components from the different i.i.d. copies $X^i = (X_1^i, \ldots, X_K^i)$. In other words, the von Mises statistic is obtained by replacing values of the components by values with different time indices. This works because there are no interactions either among the K components or among values with different time indices.

Greenwood, McKeague and Wefelmeyer (1999) extend the idea behind the von Mises statistic to samplers on random fields with local interactions. For simplicity, we restrict attention to nearest neighbor random fields and the Gibbs sampler with a specific sweep. Let $S = \{0, \ldots, k-1\}^d$ be a square lattice of dimension d. For simplicity, take k to be even. The lattice has $K = k^d$ sites. Let π be the law of a random field on E^S. As in (1.3), factor π in K different ways,

$$\pi(dx) = \pi_{-s}(dx_{-s}) p_s(x_{-s}, dx_s), \quad s \in S,$$

where x_{-s} is obtained from x by omitting x_s. The one-dimensional conditional distributions $p_s(x_{-s}, dx_s)$ are called the *local characteristics* of the random field.

A Gibbs sampler with deterministic sweep is based on some ordering $s_1, \ldots, s_K$ of the sites. Let $n = qK$. The subchain $X^0, X^K, \ldots, X^{qK}$ of full sweeps, see (1.4), has transition distribution

$$Q(x, dy) = \prod_s p_s(y_{<s}, x_{>s}, dy_s),$$

where $x_{<s}$ is the subconfiguration of all sites that come before site s. As in Section 5, the usual estimator for πf is the empirical estimator based on the subchain,

$$E_n^K f = \frac{1}{p} \sum_{q=1}^{p} f(X^{qK}).$$

For each s, the partially updated configuration $(X^{qK}_{\leq s}, X^{(q-1)K}_{>s})$ also has stationary law π, and further empirical estimators are

$$E^s_n f = \frac{1}{p} \sum_{q=1}^{p} f(X^{qK}_{\leq s}, X^{(q-1)K}_{>s}).$$

The empirical estimator based on the full chain $X^1, \dots, X^n$ is

$$E_n f = \frac{1}{n} \sum_{i=1}^{n} f(X^i) = \frac{1}{K} \sum_{s} E^s_n f.$$

The set of nearest neighbors of a site s is $\partial s = \{t : |t - s| = 1\}$, with $|t - s| = \sum_j |t_j - s_j|$. We use a free boundary, in which case the boundary sites have fewer than $2d$ neighbors. We assume *nearest neighbor interactions*,

$$p_s(x_{-s}, dx_s) = p_s(x_{\partial s}, dx_s), \tag{7.1}$$

i.e., the local characteristics at site s depend only on the nearest neighbors of s.

A widely used updating scheme for nearest neighbor models *respects the checkerboard pattern* of the lattice in the sense that it updates first the sites with, say, even parity and then those with odd parity. See, e.g., Heermann and Burkitt (1992). The corresponding Gibbs sampler updates a single site s using the local characteristic $p_s(x_{\partial s}, dx_s)$. Therefore, all even, or all odd, sites can be updated simultaneously, and the sampler can be written as a two-step Gibbs sampler. Write a configuration $x = (y_e, y_o)$, where y_e and y_o are the subconfigurations of x on the even and odd sites, respectively. The subchain of full sweeps has transition distribution

$$Q(y_e, y_o, d(z_e, z_o)) = Q_e(y_o, dz_e) Q_o(z_e, dz_o) \tag{7.2}$$

with

$$\begin{aligned} Q_e(y_o, dy_e) &= \prod_{s \text{ even}} p_s(y_{o,\partial s}, dy_s), \\ Q_o(y_e, dy_s) &= \prod_{s \text{ odd}} p_s(y_{e,\partial s}, dy_s). \end{aligned}$$

Let $X^0 = (Y^0, Y^1)$ be an initial configuration. The Gibbs sampler based on this updating scheme first creates a subconfiguration Y^2 on the even sites, then a subconfiguration Y^3 on the odd sites, and so on. Here, rather than counting the update of a complete configuration as a time step, we define a *full time step* to be the update of an even *or* an odd subconfiguration. This means that the output of the Gibbs sampler is $Y^0, Y^1, Y^2, \dots$, and the sequence of complete configurations, or full sweeps, is given by $X^0 = (Y^0, Y^1)$, $X^K = (Y^2, Y^3), \dots$.

To motivate the construction of our estimators, we assume for now that the initial configuration X^0 is distributed according to the stationary law π. Then the Gibbs sampler Markov chain $X^0, X^1, \dots$ is stationary. Now suppose that we replace a component

X_s^{qK} of the configuration X^{qK} by a future value $X_s^{(q+j)K}$. Which replacements leave the joint law of the configuration unchanged? We have already seen in $E_n^s f$ an example of such replacements for the general case with possibly non-local interactions — we replaced the values X_t^{qK} by $X_t^{(q+1)K}$ for $t \leq s$. We will see that for nearest neighbor models more general replacements are possible.

It is convenient to describe such replacements by an *update function* $I : S \to \{0, 1, \ldots\}$, with $I(s)$ even for s even, and odd otherwise. An update function I describes a new configuration $Z^I = (Y_s^{I(s)})_{s \in S}$ in terms of the observed chain $Y^0, Y^1, Y^2, \ldots$ by specifying, for each site s, the time index $I(s)$ of the value $Y_s^{I(s)}$ going into this configuration. For example, the initial configuration is $X^0 = (Y^0, Y^1) = Z^{I^0}$, where

$$I^0(s) = \begin{cases} 0, & s \text{ even}, \\ 1, & s \text{ odd}, \end{cases} \tag{7.3}$$

and X^{qK} is obtained from X^0 by shifting I^0 to yield $X^{qK} = Z^{I^0+2q}$ for $q = 1, 2, \ldots$. We say that an update function is *admissible* if its values at any two neighboring sites differ by 1. A *move* picks a site s, then replaces $I(s)$ by $I(s) + 2$, leaving I unchanged otherwise. A move is *admissible* if it preserves admissibility of the update function, see Figure 2. Note that an admissible move can be made at site s if and only if

$$I(t) = I(s) + 1 \quad \text{for all } t \in \partial s.$$

Note also that I^0 is an admissible update function, and that all admissible update functions are built up by applying finitely many admissible moves to I^0.

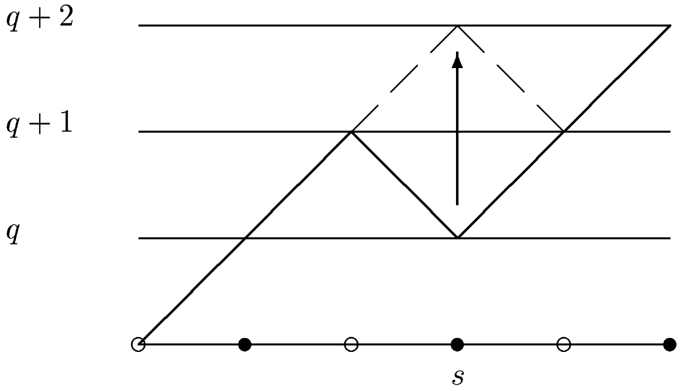


Figure 2: An admissible move at site s.

The following theorem of Greenwood, McKeague and Wefelmeyer (1999) shows that for an admissible update function I, the process Z^{I+2q}, $q = 0, 1, \ldots$ is distributed as output from another Gibbs sampler for π. The sweep of this new Gibbs sampler is *ordered* by I in the sense that it first updates the sites on the lowest level of I, then proceeds upwards layer by layer. In general, the new sweep does not respect the checkerboard pattern.

Theorem 7. *Suppose π has nearest neighbor interactions and X^i, $i = 0, 1, \ldots$, is generated by a Gibbs sampler for π whose updating respects the checkerboard pattern of the lattice. If I is an admissible update function, then Z^{I+2q}, $q = 0, 1, \ldots$, is distributed as a full sweep Gibbs sampler for π having sweep ordered by I.*

The idea of the proof is simple: Note that an update at a site s is obtained by adding 2 to the current value of the update function at that site. Thus, the configuration Z^{I+2q} is obtained from Z by applying Gibbs sampler updates site by site in the order of the sweep associated with I. A more formal version of this argument is as follows. Let I_s denote the update function obtained from I by applying the moves at the sites before s in the order of the new sweep. If $I_s(s) = q$, then since I_s is admissible, $I_s(t) = q + 1$ for $t \in \partial s$. The move at s replaces $I_s(s) = q$ by $q + 2$. Recall that $Y_s^{q+2} = Z_s^{I+2}$ was generated using the conditional law $p_s(Y_{\partial s}^{q+1}, dx_s)$ which equals $p_s(Z_{-s}^{I_s}, dx_s)$. Hence Z^{I+2} is obtained from Z^I using the Gibbs sampler with the new sweep.

Call an admissible update function I a *update function* if it uses part of the initial configuration X^0, i.e., if $\min I$ equals 0 or 1. Each update function I gives an estimator for πf,

$$E_n^I f = \frac{1}{n-h+2} \sum_{q=0}^{n-h+1} f(Z^{I+2q}).$$

Here h is the *height* of I, i.e., the number of full time steps, or half sweeps, it straddles. This means that $\max I$ equals $2h - 2$ or $2h - 1$.

The asymptotic variance of $E_n^I f$ can be substantially different from that of the usual empirical estimator $E_n f$. Estimators with reduced variance might be obtained by averaging over some family $\mathcal{I}$ of update functions:

$$E_n^{\mathcal{I}} f = \frac{1}{|\mathcal{I}|} \sum_{I \in \mathcal{I}} E_n^I f,$$

where $|\mathcal{I}|$ denotes the cardinality of $\mathcal{I}$.

Averaging over update functions can be interpreted as symmetrizing $E_n f$, as in a generalized von Mises statistic. In general we expect such estimators to have smaller variance for larger families of update functions. However, there is a trade-off in terms of computational cost: For high update functions we would need to store more configurations, and for large families we would need to evaluate f more frequently, which could be critical in large lattices or when $f(x)$ is expensive to compute.

If the random field is arbitrary, with not necessarily local interactions, we can only use the update functions involved in the empirical estimators $E_n^s f$. For s even the update function I^s is

$$I^s(t) = \begin{cases} 0, & t > s, t \text{ even}, \\ 2, & t \le s, t \text{ even}, \\ 1, & t \text{ odd}. \end{cases}$$

The update function for s odd is similar:

$$I^s(t) = \begin{cases} 1, & t > s, t \text{ odd}, \\ 3, & t \le s, t \text{ odd}, \\ 2, & t \text{ even}. \end{cases}$$

To make use of the nearest neighbor assumption, we must go beyond the update functions just described. For large lattices it may not be computationally feasible to use all update functions. If one uses only a few update functions, they should be well spaced to reduce correlation between different $E_n^I f$. The higher the update functions we allow, the better we can space them. However, high update functions require more storage: To calculate $E_n^I f$ for a update function I of height h, we must store h configurations at a time.

Simulations in Greenwood, McKeague and Wefelmeyer (1999) for the Ising model and the Gibbs and Metropolis samplers, and for f the r-th nearest neighbor correlation, show that the variance reduction can be considerable, even if only a few of the update functions are used.

8 Exploiting symmetries of random fields

Let $X^0, \dots, X^n$ be observations from an arbitrary Markov chain with transition distribution $Q(x, dy)$. Assume that the chain is positive Harris recurrent, with invariant distribution $\pi(dx)$, and V-uniformly ergodic. Let T be a measurable transformation on E which leaves π invariant and has a measurable inverse. For $f^2 \le V$, we obtain a new consistent and asymptotically normal empirical estimator for πf,

$$E_n(f \circ T) = \frac{1}{n} \sum_{i=1}^{n} f(TX^i).$$

The same is true for any power T^j of T.

Suppose now that the transition distribution Q is invariant under T in the sense that $Q(x, dy) = Q(Tx, Tdy)$. This forces π to be invariant under T. Also, for the stationary chain, $(X^0, X^1, \dots, X^n)$ is distributed as $(TX^0, \dots, TX^n)$. Hence $E_n(f \circ T^j)$ has the same asymptotic variance as $E_n f$. Better estimators, in the sense of asymptotic variance, can be obtained by linear combinations of $E_n(f \circ T^j)$. The following result of Greenwood, McKeague and Wefelmeyer (1996) shows how best to use linear combinations of such estimators if the powers form a finite cyclic group.

Proposition 1. *Let Q be invariant under a transformation T with $T^m = T^0$ for some $m \geq 2$. Then the best linear combination of $E_n(f \circ T^j)$, $j = 0, \dots, m-1$, is the average,*

$$\bar{E}_n f = \frac{1}{m} \sum_{j=0}^{m-1} E_n(f \circ T^j).$$

The proof is simple. Observe that the pair $E_n(f \circ T^k)$, $E_n(f \circ T^j)$ is the pair $E_n f$, $E_n(f \circ T^{(j-k) \bmod m})$ evaluated with the chain $X^0, X^1, \dots$ replaced by the chain $T^k X^0, T^k X^1, \dots$. Hence the asymptotic covariances of the two pairs agree. Therefore, the asymptotic covariance matrix of $E_n(f \circ T^j)$, $j = 0, \dots, m-1$, is circulant. In particular, it has equal row sums. If the covariance matrix is nonsingular, Proposition 1 follows from Greenwood, McKeague and Wefelmeyer (1996, Lemma 2). Proposition 1 also follows from the observation that if Σ is positive semidefinite with equal row sums, then $b'\Sigma b$ is minimized over vectors b with $\sum_{j=1}^m b_j = 1$ by $b_j = 1/m$ for all j. To see this, let $e_1, \dots, e_m$ be an orthonormal basis of eigenvectors of Σ with non-negative eigenvalues $\mu_1, \dots, \mu_m$. Assume w.l.g. that $e_1 = (m^{-1/2}, \dots, m^{-1/2})$. Write $b = \sum_{j=1}^m \lambda_j e_j$. Then $\lambda_1 = m^{-1/2}$ and

$$b'\Sigma b = \sum_{j=1}^m \lambda_j^2 \mu_j = \mu_1/m + \sum_{j=2}^m \lambda_j^2 \mu_j,$$

which is minimized by $\lambda_j = 0$ for $j = 2, \dots, m$.

Greenwood, McKeague and Wefelmeyer (1996) apply Proposition 1 to the two-step Gibbs sampler. Let π be a distribution on a two-dimensional space $E = E_1 \times E_2$. Let $p_1(x_2, dx_1)$ be the conditional distribution of x_1 given x_2, and $p_2(x_1, dx_2)$ the conditional distribution of x_2 given x_1. Let $n = 2p$, and let $X^0, X^1, \dots, X^{2p}$ be observations from the two-step Gibbs sampler with deterministic sweep. The subchain $X^0, X^2, \dots, X^{2p}$ of full sweeps has transition distribution

$$Q(x, dy) = p_1(x_2, dy_1) p_2(y_1, dy_2).$$

Call a transformation T on $E_1 \times E_2$ *parallel* if it is a direct product $T(x_1, x_2) = (T_{11}x_1, T_{22}x_2)$, and *transverse* if $T(x_1, x_2) = (T_{21}x_2, T_{12}x_1)$. Note that the composition of two transverse transformations is parallel, and the composition of a parallel with a transverse is transverse. If T is parallel and leaves π invariant, then

$$p_1(x_2, dx_1) = p_1(T_{22}x_2, T_{11}dx_1), \quad p_2(x_1, dx_2) = p_2(T_{11}x_1, T_{22}dx_2). \tag{8.1}$$

It follows that under the stationary law of the sampler,

$$(X^0, X^1, \dots, X^{2p}) = (TX^0, TX^1, \dots, TX^{2p}) \quad \text{in distribution.}$$

If T is transverse and leaves π invariant, then

$$p_1(x_2, dx_1) = p_2(T_{21}x_2, T_{12}dx_1), \quad p_2(x_1, dx_2) = p_1(T_{12}x_1, T_{21}dx_2). \tag{8.2}$$

It is easy to see that the transformed time-reversed chain has the same law as the original chain:

$$(X^0, X^1, \ldots, X^{2p}) = (TX^{2p}, TX^{2p-1}, \ldots, TX^0) \quad \text{in distribution.}$$

As in Section 5, the empirical estimator $E_n f = \frac{1}{n}\sum_{i=1}^n f(X^i)$ can be written as the average $\frac{1}{2}(E_n^1 f + E_n^2 f)$ of the empirical estimators based on the two subchains,

$$E_n^1 f = \frac{1}{p}\sum_{q=1}^p f(X^{2q-1}), \quad E_n^2 f = \frac{1}{p}\sum_{q=1}^p f(X^{2q}).$$

Greenwood, McKeague and Wefelmeyer (1996, Theorem 2) use Proposition 1 to show the following result.

Theorem 8. *Assume that the two-step Gibbs sampler for π is positive Harris recurrent, and that the two subchains are V-uniformly ergodic. Let π be invariant under a parallel or transverse transformation T with $T^m = T^0$, and let $f^2 \le V$. Then the empirical estimators $E_n^1(f \circ T^j), E_n^2(f \circ T^j)$, $j = 0, \ldots, m-1$, have equal asymptotic variances, and the best linear combination is the average,*

$$\bar{E}_n f = \frac{1}{m}\sum_{j=0}^{m-1} E_n(f \circ T^j).$$

Theorem 8 can be generalized to more than one transformation, as long as the transformations commute; see Greenwood, McKeague and Wefelmeyer (1996, Theorem 3).

Theorem 8 can be applied to nearest neighbor random fields. For simplicity, let

$$S = \{0, \ldots, k_1 - 1\} \times \{0, \ldots, k_2 - 1\},$$

where k_1 and k_2 are even. Let $\pi(dx)$ be the law of a random field on E^S with nearest neighbor interactions (7.1). As in Section 7, number first the even and then the odd sites, respecting the checkerboard pattern of the lattice. Then the Gibbs sampler with deterministic sweep can be written as a two-step Gibbs sampler, with full sweep transition distribution (7.2),

$$Q(y_e, y_o, d(z_e, z_o)) = Q_e(y_o, dz_e)Q_o(z_e, dz_o),$$

where y_e and y_o are the subconfigurations of x on the even and odd sites, respectively.

Define addition on S by $(s+t)_1 = s_1 + t_1 \bmod k_1$, $(s+t)_2 = s_2 + t_2 \bmod k_2$. For $t \in S$, the translation of S by t is defined as $T_t s = s - t$. This induces a translation on E^S by $(T_t x)_s = x_{T_t^{-1}s} = x_{s+t}$. Translations on E^S by an even or odd number of sites are parallel or transverse transformations, respectively. For a *horizontal* translation, think of the lattice as wrapped around a cylinder so that the vertical boundaries meet. The neighbors of each site $s = (s_1, s_2)$ along the vertical boundary now include $(s_1 \pm 1, s_2)$ with addition mod k_1.

A horizontal translation by an *even* number of sites is $T = T_{(p,0)}$, with p even. This translation takes even into even sites and odd into odd and is a parallel transformation in the sense of Section 4. Suppose that k_1 is a multiple of p, say $k_1 = mp$. Suppose that π is invariant under T. Then it is also invariant under powers $T^j = T_{(jp,0)}, j = 0, \dots, m-1$. These transformations form a cyclic group. Theorem 8 implies that the empirical estimators

$$E_n^1(f \circ T_{(jp,0)}), \ E_n^2(f \circ T_{(jp,0)}), \quad j = 1, \dots, m-1,$$

have equal asymptotic variances, and the best linear combination is the average.

A horizontal translation by an *odd* number of sites is $T = T_{(p,0)}$, with p odd. This translation takes even into odd sites and odd into even and is a transverse transformation. Even powers of T are parallel. Suppose that k_1 is a multiple of p, say $k_1 = mp$. Suppose that π is invariant under T. As above, the best linear combination of the corresponding empirical estimators is the average.

Horizontal and vertical translations commute. Hence the best linear combination of the corresponding empirical estimators is again the average. Greenwood, McKeague and Wefelmeyer (1996) present simulations for the Ising model without and with external field, and for f the nearest neighbor correlation.

Acknowledgment. Work supported by NSERC, Canada. We thank Antonietta Mira, Radford Neal and Gareth Roberts for their comments, and the referee for his careful reading of the manuscript and his many corrections and suggestions.

References

Athreya, K. B., H. Doss and J. Sethuraman (1996). On the convergence of the Markov chain simulation method. *Ann. Statist.* **24**, 69–100.

Belisle, C. J., P. Romeijn and R. L. Smith (1993). Hit-and-run algorithms for generating multivariate distributions. *Math. Oper. Res.* **18**, 255-266.

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

Besag, J., P. Green, D. Higdon and K. Mengersen (1995). Bayesian computation and stochastic systems (with discussion). *Statist. Sci.* **10**, 3–66.

Bickel, P. J., C. A. J. Klaassen, Y. Ritov and J. A. Wellner (1998). *Efficient and Adaptive Estimation for Semiparametric Models.* Springer, New York.

Brémaud, P. (1999). *Markov Chains. Gibbs Fields, Monte Carlo Simulation, and Queues.* Texts in Applied Mathematics 31, Springer, New York.

Casella, G. and C. P. Robert (1996). Rao–Blackwellization of sampling schemes. *Biometrika* **83**, 81–94.

Chan, K. S. (1993). Asymptotic behavior of the Gibbs sampler. *J. Amer. Statist. Assoc.* **88**, 320–326.

Clifford, P. and G. Nicholls (1995). A Metropolis sampler for polygonal image reconstruction. Technical Report, Department of Statistics, Oxford University.

Conway, J. B. (1990). *A Course in Functional Analysis*, 2nd ed. Graduate Texts in Mathematics 96, Springer, New York.

Damien, P., J. Wakefield and S. Walker (1999). Gibbs sampling for Bayesian nonconjugate and hierarchical models by using auxiliary variables. *J. Roy. Statist. Soc. Ser. B* **61**, 331–344.

Diaconis, P. and L. Saloff-Coste (1998). What do we know about the Metropolis algorithm? *J. Comput. Syst. Sci.* **57**, 20–36.

Edwards, R. G. and A. D. Sokal (1988). Generalization of the Fortuin–Kasteleyn–Swendsen–Wang representation and Monte Carlo algorithm. *Phys. Rev. Lett.* **38**, 2009–2012.

Fishman, G. (1996). Coordinate selection rules for Gibbs sampling. *Ann. Appl. Probab.* **6**, 444–465.

Fishman, G. (1999). An analysis of Swendsen–Wang and related sampling methods. *J. Roy. Statist. Soc. Ser. B* **61**, 623–641.

Frigessi, A., J. Gåsemyr and H. Rue (2000). Antithetic coupling of two Gibbs sampler chains. *Ann. Statist.* **28**.

Frigessi, A., C.-R. Hwang, S. J. Sheu and P. Di Stefano (1993). Convergence rates of the Gibbs sampler, the Metropolis algorithm, and other single-site updating dynamics. *J. Roy. Statist. Soc. Ser. B* **55**, 205–220.

Frigessi, A., C.-R. Hwang and L. Younes (1992). Optimal spectral structure of reversible stochastic matrices, Monte Carlo methods and the simulation of Markov random fields. *Ann. Appl. Probab.* **2**, 610–628.

Gamerman, D. (1997). *Markov Chain Monte Carlo. Stochastic Simulation for Bayesian Inference.* Chapman and Hall, London.

Gelfand, A. E., S. E. Hills, A. Racine-Poon and A. F. M. Smith (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *J. Amer. Statist. Assoc.* **85**, 972–985.

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

Gelfand, A. E. and A. F. M. Smith (1991). Gibbs sampling for marginal posterior expectations. *Comm. Statist. Theory Methods* **20**, 1747–1766.

Gelman, A., G. O. Roberts, and W. R. Gilks (1996). Efficient Metropolis jumping rules. In: *Bayesian Statistics 5* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M.

Smith, eds.), 599–608, Oxford University Press.

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

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

Geyer, C. J. (1995). Conditioning in Markov chain Monte Carlo. *J. Comput. Graph. Statist.* **4**, 148–154.

Gilks, W. R., S. Richardson and D. J. Spiegelhalter (eds.) (1996). *Introducing Markov Chain Monte Carlo.* Chapman and Hall, London.

Gordin, M.I. (1969). The central limit theorem for stationary processes. *Soviet Math. Dokl.* **10**, 1174-1176.

Graham, J. (1994). Monte Carlo Markov chain likelihood ratio test and Wald test for binary spatial lattice data. Technical Report, Department of Statistics, North Carolina State University.

Green, P. J. and X.-l. Han (1992). Metropolis methods, Gaussian proposals and antithetic variables. In: *Stochastic Models, Statistical Methods, and Algorithms in Image Analysis* (P. Barone, A. Frigessi and M. Piccioni, eds.), 142–164, Lecture Notes in Statistics 74, Springer, Berlin.

Green, P. J. and A. Mira (1999). Delaying rejection in Metropolis-Hastings algorithms with reversible jumps. Technical Report, Department of Mathematics, University of Bristol.
`http://aim.unipv.it/∼anto/`

Greenwood, P. E., I. W. McKeague and W. Wefelmeyer (1996). Outperforming the Gibbs sampler empirical estimator for nearest neighbor random fields. *Ann. Statist.* **24**, 1433–1456.

Greenwood, P. E., I. W. McKeague and W. Wefelmeyer (1998). Information bounds for Gibbs samplers. *Ann. Statist.* **26**, 2128–2156.

Greenwood, P. E., I. W. McKeague and W. Wefelmeyer (1999). Von Mises type statistics for single site updated local interaction random fields. *Statistica Sinica* **9**, 699–712.

Grenander, U. (1983). *Tutorial in Pattern Theory.* Lecture Notes, Division of Applied Mathematics, Brown University.

Hájek, J. (1970). A characterization of limiting distributions of regular estimates. *Z. Wahrsch. Verw. Gebiete* **14**, 323–330.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109. Reprinted with introduction in: Kotz and Johnson (1997).

Heermann, D. W. and A. N. Burkitt (1992). Parallel algorithms for statistical physics problems. In: *The Monte Carlo Method in Condensed Matter Physics* (K. Binder, ed.), 53–74, Springer, Berlin.

Higdon, D. M. (1998). Auxiliary variable methods for Markov chain Monte Carlo with applications. *J. Amer. Statist. Assoc.* **93**, 585–595.

Höpfner, R. (1993). On statistics of Markov step processes: representation of log-likelihood ratio processes in filtered local models. *Probab. Theory Related Fields* **94**, 375–398.

Höpfner, R., J. Jacod and L. Ladelli (1990). Local asymptotic normality and mixed normality for Markov statistical models. *Probab. Theory Related Fields* **86**, 105–129.

Horn, R. A. and C. R. Johnson (1985). *Matrix Analysis.* Cambridge University Press.

Ingrassia, S. (1994). On the rate of convergence of the Metropolis algorithm and Gibbs sampler by geometric bounds. *Ann. Appl. Probab.* **4**, 347–389.

Jarner, S. F. and G. Roberts (2000). Polynomial convergence rates of Markov chains. Technical Report, Department of Mathematics and Statistics, Lancaster University. `http://www.maths.lancs.ac.uk/~jarner/`

Jerrum, M. (1998). Mathematical foundations of the Markov chain Monte Carlo method. In: *Probabilistic Methods for Algorithmic Discrete Mathematics* (M. Habib, C. McDiarmid, J. Ramirez-Alfonsin and B. Reed, eds.), 116–165, Algorithms and Combinatorics 16, Springer, New York.

Johnson, V. E. (1996). Studying convergence of Markov chain Monte Carlo algorithms using coupled sample paths. *J. Amer. Statist. Assoc.* **91**, 154–166.

Kartashov, N. V. (1985a). Criteria for uniform ergodicity and strong stability of Markov chains with a common phase space. *Theory Probab. Math. Statist.* **30**, 71–89.

Kartashov, N. V. (1985b). Inequalities in theorems of ergodicity and stability for Markov chains with common phase space. I. *Theory Probab. Appl.* **30**, 247–259.

Kartashov, N. V. (1996). *Strong Stable Markov Chains.* VSP, Utrecht.

Kalos, M. H. and P. A. Whitlock (1986). *Monte Carlo Methods. Volume 1. Basics.* Wiley, New York.

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

Kira, E. and C. Ji (1997). Rates of convergence for the Gibbs sampler. *Markov Process. Related Fields* **3**, 89–102.

Koshevnik, Y. A. and B. Y. Levit (1976). On a non-parametric analogue of the information matrix. *Theory Probab. Appl.* **21**, 738-753.

Kotz, S. and N. L. Johnson (eds.) (1997). *Breakthroughs in Statistics. Volume III.* Springer, New York.

Levit, B. Y. (1974). On optimality of some statistical estimates. In: *Proceedings of the Prague Symposium on Asymptotic Statistics* (J. Hájek, eds.), **2**, 215–238, Charles University, Prague.

Liu, J. S. (1996). Peskun's theorem and a modified discrete-state Gibbs sampler. *Biometrika* **83**, 681–682.

Liu, J. S., W. H. Wong and A. Kong (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika* **81**, 27–40.

Liu, J. S., W. H. Wong and A. Kong (1995). Covariance structure and convergence rate of the Gibbs sampler with various scans. *J. Roy. Statist. Soc. Ser. B* **57**, 157–169.

McEachern, S. N. and L. M. Berliner (1994). Subsampling the Gibbs sampler. *Amer. Statist.* **48**, 188–190.

McKeague, I. W. and W. Wefelmeyer (2000). Markov chain Monte Carlo and Rao–Blackwellization. *J. Statist. Plann. Inference* **85**, 171–182.

Mengersen, K. L., C. P. Robert and C. Guihenneuc-Jouyaux (1999). MCMC convergence diagnostics: a "reviewww". In: *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.), 415–440, Oxford University Press.

Mengersen, K. L. and R. L. Tweedie (1996). Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Statist.* **24**, 101–121.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–1092. Reprinted with introduction in: Kotz and Johnson (1997).

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

Meyn, S. P. and R. L. Tweedie (1994). Computable bounds for geometric convergence rates of Markov chains. *Ann. Appl. Probab.* **4**, 981–1011.

Mira, A. and C. J. Geyer (1999). Ordering Monte Carlo Markov Chains. Technical Report, School of Statistics, University of Minnesota.
`http://aim.unipv.it/~anto/`

Mira, A. and L. Tierney (1999). On the use of auxiliary variables in Markov chain Monte Carlo sampling. Technical Report, Medical Informatics Laboratory, University of Pavia. To appear in: *Scand. J. Statist.*
`http://aim.unipv.it/~anto/`

Neal, R. M. (1993). *Probabilistic Inference Using Markov Chain Monte Carlo Methods.* Technical Report, Department of Statistics, University of Toronto.
`http://www.cs.toronto.edu/~radford/`

Neal, R. M. (2000). Slice sampling. Technical Report, Department of Statistics, Uni-

versity of Toronto.
`http://www.cs.toronto.edu/~radford/`

Pearl, J. (1987). Evidential reasoning using stochastic simulation. *Artificial Intelligence* **32**, 245–257.

Penev, S. (1991). Efficient estimation of the stationary distribution for exponentially ergodic Markov chains. *J. Statist. Plann. Inference* **27**, 105–123.

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

Pfanzagl, J. and W. Wefelmeyer (1982). *Contributions to a General Asymptotic Statistical Theory.* Lecture Note in Statistics 13, Springer, New York.

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

Robert, C. P. (1996). *Méthodes de Monte Carlo par Chaînes des Markov.* Economica, Paris.

Robert, C. P. (ed.) (1998). *Discretization and MCMC. Convergence Assessment.* Lecture Notes in Statistics 135, Springer, New York.

Robert, C. P. and G. Casella (1999). *Monte Carlo Statistical Methods.* Springer, New York.

Roberts, G. O., A. Gelman and W. R. Gilks (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.* **7**, 110–120.

Roberts, G. O. and N. G. Polson (1994). On the geometric convergence of the Gibbs sampler. *J. Roy. Statist. Soc. Ser. B* **56**, 377–384.

Roberts, G. O. and J. S. Rosenthal (1997). Geometric ergodicity and hybrid Markov chains. *Electron. Comm. Probab.* **2**, 13–25.
`http://www.math.washington.edu/~ejpecp/`

Roberts, G. O. and J. S. Rosenthal (1998). Markov-chain Monte Carlo: Some practical implications of theoretical results (with discussion). *Canad. J. Statist.* **26**, 5–31.

Roberts, G. O. and J. S. Rosenthal (1999). Convergence of slice sampler Markov chains. *J. Roy. Statist. Soc. Ser. B* **61**, 643–660.

Roberts, G. O. and S. K. Sahu (1997). Updating schemes, correlation structure, blocking and parameterisation for the Gibbs sampler. *J. Royal. Statist. Soc. Ser. B* **59**, 291–317.

Roberts, G. O. and R. L. Tweedie (1996). Geometric convergence and central limit theorems for multivariate Hastings and Metropolis algorithms. *Biometrika* **83**, 95–110.

Roberts, G. O. and R. L. Tweedie (1999). Bounds on regeneration times and convergence rates for Markov chains. *Stochastic Process. Appl.* **80**, 211–229.

Roberts, G. O. and R. L. Tweedie (2000). Rates of convergence of stochastically monotone and continuous time Markov models. *J. Appl. Probab.* **37**, 359–373.

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

Roussas, G. G. (1965). Asymptotic inference in Markov processes. *Ann. Math. Statist.* **36**, 987–992.

Schervish, M. J. and B. P. Carlin (1992). On the convergence of successive substitution sampling. *J. Comput. Graph. Statist.* **1**, 111–127.

Schmeiser, B. and M. H. Chen (1991). On random-direction Monte Carlo sampling for evaluating multidimensional integrals. Technical Report, Department of Statistics, Purdue University.

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

Spiegelhalter, D. J., A. P. Dawid, S. L. Lauritzen and R. G. Cowell (1993). Bayesian analysis in expert systems (with discussion). *Statist. Science* **8**, 219–283.

Spiegelhalter, D. J., A. Thomas and N. G. Best (1996). Computation on Bayesian graphical models. In: *Bayesian Statistics 5* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.), 407–426, Oxford University Press.

Swendsen, R. H. and J.-S. Wang (1987). Nonuniversal critical dynamics in Monte Carlo simulations. *Phys. Rev. Lett.* **58**, 86–88.

Tanner, M. A. (1996). *Tools for Statistical Inference. Methods for the Exploration of Posterior Distributions and Likelihood Functions*, 3rd ed. Springer, New York.

Tanner, M. A. and W. H. Wong (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. (1996). Introduction to general state-space Markov chain theory. In: Gilks, Richardson and Spiegelhalter (1996), 59–74.

Tierney, L. (1998). A note on Metropolis–Hastings kernels for general state spaces. *Ann. Appl. Probab.* **8**, 1–9.

Tierney, L. and A. Mira (1999). Some adaptive Monte Carlo methods for Bayesian inference. *Statistics in Medicine* **18**, 2507–2515.

Wefelmeyer, W. (1998). Judging MCMC estimators by their asymptotic variance. In: *Prague Stochastics '98*, Vol. 2 (M. Hušková, P. Lachout and J. A. Víšek, eds.), 591–596, Union of Czech Mathematicians and Physicists, Prague.

Wefelmeyer, W. (1999). Efficient estimation in Markov chain models: an introduction. In: *Asymptotics, Nonparametrics, and Time Series* (S. Ghosh, ed.), 427–459, Dekker, New York.