Statistical Science
2001, Vol. 16, No. 4, 340–350

Ordering and Improving the Performance of Monte Carlo Markov Chains

Antonietta Mira

Abstract. An overview of orderings defined on the space of Markov chains having a prespecified unique stationary distribution is given. The intuition gained by studying these orderings is used to improve existing Markov chain Monte Carlo algorithms.

Key words and phrases: Asymptotic variance, convergence ordering, covariance ordering, efficiency ordering, Metropolis–Hastings algorithm, Peskun ordering, reversible jumps.

1. MOTIVATION

Suppose we are given a probability distribution π, possibly known only up to a normalizing constant, on a finite set $\mathbf{X}$, and we are interested in estimating the expectation of some function f, $E_\pi f(x) = \mu$. When the dimension of $\mathbf{X}$ is large, Markov chain Monte Carlo (MCMC) methods could be used to estimate μ by the ergodic average $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} f(x_i)$, where x_i's are dependent samples collected along the path of length n of a Markov chain that has π as its unique stationary (and limiting) distribution. The Markov chain is identified with its transition matrix (kernel in general state spaces): $P(x, A) = \Pr(x_n \in A | x_{n-1} = x)$ for every set A.

One of the key observations that motivates this paper is the fact that, given a distribution π, there are many Markov chains that are stationary with respect to it and could therefore be used for MCMC purposes. A choice is thus necessary and criteria to guide this choice are needed. Orderings defined on the space of Markov chains that have a specified stationary distribution allow a comparison and can aid the selection of one chain over another. The second motivation is the fact that often the intuition behind an ordering leads to the definition of new MCMC algorithms or to the improvement of existing ones. This twofold motivation drives the structure of this paper which is divided in two parts. In the first part, we define some of the orderings available in the MCMC literature; in the second part, we provide strategies for improving existing MCMC algorithms relative to the given orderings.

In this paper, we will mainly refer to finite state spaces and reversible transition matrices but will point out when extensions of the results stated are available (to our knowledge) in more general settings.

2. ORDERING MCMC ESTIMATORS RELATIVE TO THEIR PERFORMANCE

The main criteria used to evaluate the performance of a transition matrix used for MCMC simulation are the *asymptotic variance* (AV) of the resulting estimates and the *speed of convergence* (SC) to stationarity. In this paper, we will mainly focus on the AV measured by $v(f, P)$, the limit, as n tends to ∞, of n times the variance of the estimator, $\hat{\mu}_n$, computed on a P-chain, that is, a π-stationary chain updated using the transition matrix P. On the other hand, SC is measured by how fast the chain approaches its stationary distribution where the distance considered is typically total variation: $\|P^n(x, \cdot) - \pi(\cdot)\| = \sup_{y \in \mathbf{X}} |P^n(x, y) - \pi(y)|$, where $P^n(x, y) = P(x_n = y | x_0 = x)$ is the n-step transition matrix.

As observed by Besag and Green (1993), the two goals lead to different notions of optimality since the AV depends on the eigenvalues while convergence to stationarity depends on the absolute value of the eigenvalues of the transition matrix, as appears from Theorem 1. We state and prove the theorem for sufficiently regular problems with transition matrices described by a sequence of eigenvalues $\{\lambda_{0P} \geq \lambda_{1P} \geq \cdots\}$ and corresponding eigenvectors $\{e_{0P}, e_{1P}, \ldots\}$, but the result holds in general:

Antonietta Mira is Associate Professor, Department of Economics, University of Insubria, Via Ravasi 2, 21100 Varese, Italy. *This work has been supported by EU TMR network ERB-FMRX-CT96-0095 on "Computational and Statistical methods for the analysis of spatial data."*

THEOREM 1. *The AV of MCMC estimates obtained by averaging along the chain sample path is an increasing function of the eigenvalues of the transition matrix used to update the chain. The speed at which convergence to stationarity is achieved depends on the second largest eigenvalue in absolute value.*

PROOF. Let A be the matrix whose rows are all equal to π. This is the transition matrix we would use if we could sample i.i.d. observations from π (MCMC would then reduce to Monte Carlo simulation). Then $v(f, A)$ is the theoretical independence sampling variance of the estimate $\hat{\mu}_n$. The first statement in the theorem follows from this representation of the AV:

$$v(f, P) = \sum_j \frac{1 + \lambda_{jP}}{1 - \lambda_{jP}} k_j v(f, A), \tag{2.1}$$

where k_j are some nonnegative constants such that $\sum_j k_j = 1$. The result is then a consequence of the fact that $(1+\lambda)/(1-\lambda)$ is an increasing function of λ.

The second statement follows from this representation of the n-step transition matrix:

$$P^n(x, y) = \sum_j e_{jP}(x) e_{jP}(y) \lambda_{jP}^n.$$

For large n, the dominant term in the above representation is $e_{0P}(\cdot) = \pi(\cdot)$ and since $\lambda_{0P} = 1$, the speed at which convergence, in total variation distance, to stationarity is achieved, depends on the second largest eigenvalue in absolute value which we will indicate by λ_P^*. □

If the transition matrix is positive definite, that is, its eigenvalues are positive, the optimal Markov chains relative to the two goals of AV and SC coincide (see Theorem 6). Thus, while the optimal samplers may be different, we could get good performance with respect to both criteria simultaneously. An alternative and appealing practice is to use a sampler with good convergence properties for the first part of the simulation (until convergence is presumably achieved) and then switch to a sampler with better properties in terms of AV.

We conclude by further stressing that AV and SC are strongly related concepts and spectral gap bounds lead to bounds on the asymptotic variance of MCMC estimators: expression (2.1) leads to the following inequality:

$$v(f, P) \leq \frac{1 + \lambda_P^*}{1 - \lambda_P^*} v(f, A) \leq \frac{2}{1 - \lambda_P^*} v(f, A),$$

where $1 - \lambda_P^*$ is known as the spectral gap.

2.1 Orderings Relative to the Asymptotic Variance of the Estimators

To justify this criterion, consider that, in classical asymptotic statistics, estimates are compared in terms of their asymptotic relative efficiency. If a Markov chain is irreducible and the state space is finite, the corresponding MCMC estimates are strongly consistent and asymptotically normally distributed; therefore, efficiency can be measured by the asymptotic variance of estimates. Under slightly stronger conditions, asymptotic normality and consistency are also guaranteed on general state spaces [Tierney (1994)].

Sometimes we run an MCMC simulation having in mind a specific function of interest whose expectation relative to π we want to estimate. In other situations, either we have a range of functions or we do not have any specific function in mind and are instead interested in studying π per se. In the first setting, we want a sampler that has good efficiency properties *relative* to the specific function of interest. In the second setting, we will look for a chain that is more efficient *uniformly* over all sensible f's. This distinction leads to the following definitions. Indicate with $\mathscr{S}$ the class of Markov chains stationary with respect to π [i.e., $\sum_x \pi(x)P(x, y) = \pi(y), \forall\, y \in \mathbf{X}$], $\mathscr{R}$ the subset of the reversible ones ($\pi(x)P(x, y) = \pi(y)P(y, x), \forall\, x, y \in \mathbf{X}$) and $L^2(\pi)$ the space of all f that have a finite variance with respect to π.

DEFINITION 2.1. Let $P, Q \in \mathscr{S}$. P is at least as efficient as Q, relative to a particular function f, $P \succeq_f Q$, if $v(f, P) \leq v(f, Q)$.

DEFINITION 2.2. Let $P, Q \in \mathscr{S}$. P is at least as uniformly efficient as Q, $P \succeq_E Q$, if $v(f, P) \leq v(f, Q)$ for all $f \in L^2(\pi)$.

2.1.1 *Uniform efficiency.* The first ordering of MCMC samplers introduced in the literature refers to uniform efficiency and is due to Peskun (1973):

DEFINITION 2.3. Let $P, Q \in \mathscr{R}$. P dominates Q in the Peskun sense, $P \succeq_P Q$, if each of the off-diagonal elements of P is greater than or equal to the corresponding off-diagonal element of Q.

Tierney (1998) extended Peskun ordering from finite to general state spaces. These orderings provide a sufficient condition for uniform efficiency as the following theorem [Peskun (1973) and Tierney (1998)] states:

THEOREM 2. *If P dominates Q in the Peskun sense, then $P \succeq_E Q$. Furthermore, $P \succeq_E Q$ if and only if $\lambda_{iP} \leq \lambda_{iQ}$ for all i.*

COMMENTS ON THE PROOF. Peskun (1973) proved the first part of Theorem 2 for finite state spaces by writing the asymptotic variance as a function of the off-diagonal elements of the transition matrix and showing that this function is decreasing. Tierney (1998) gave a more elegant and general proof by first showing that Peskun ordering implies that $Q - P$ is a positive semidefinite matrix and then proving that this implies an ordering on the asymptotic variances for all functions of interest (uniform efficiency). We will sketch the proof of this result as part of the proof of Theorem 3.

Assuming Tierney's result, the ordering on the eigenvalues is then well known for symmetric matrices. In our setting, neither P nor Q need be symmetric but if we consider them as operators on $L^2(\pi)$, that is, if we take the inner product to be $(f, g) = E_\pi[f(x)g(x)]$, then the transition matrices are indeed self-adjoint operators, provided that the reversibility condition holds. Sufficiency follows from the Courant–Fisher min–max representation of the ith largest eigenvalue for self–adjoint operators [Bellman (1972)]:

$$\lambda_{iP} = \min_{\substack{(g_j, g_j)=1 \\ j=1,\ldots,i-1}} \left\{ \max_{\substack{(f, g_j)=0 \\ j=1,\ldots,i-1}} \frac{(f, Pf)}{(f, f)} \right\},$$

where g_j are arbitrary vectors. The reverse implication follows from the spectral theorem of self-adjoint operators.

Theorem 2 readily extends to general state spaces [Mira and Geyer (1999)] except that the eigenvalues cannot be ordered anymore (since it does not even make sense to talk about eigenvalues). What can be ordered are the respective suprema of the spectra of the operators defined by the transition kernels (refer to Section 2.2 for a formal definition of the spectrum).

The first use of Peskun ordering appears in Peskun (1973) and states that the Metropolis–Hastings (MH) algorithm dominates a class of competitors reversible with respect to some π, all with the same propose/accept updating structure [see also Billera and Diaconis (2001)]. We remind the reader how the MH updating mechanism works: suppose the chain, at time t, is in position x: $X_t = x$. Propose a candidate move y by generating it from a distribution, $q(x, \cdot)$, that is allowed to depend on the current position. With probability $\alpha(x, y) = \min\{1, \pi(y)q(y, x)/\pi(x)q(x, y)\}$, accept the move and thus set $X_{t+1} = y$; otherwise, retain the same position: $X_{t+1} = x$.

Another interesting use of Peskun ordering appears in Tierney (1998) where the author compared the benefit, in terms of uniform efficiency, of two approaches to using mixtures of MH kernels. In Frigessi, Hwang and Younes (1992), the optimal transition matrix relative to the Peskun ordering is constructed. The construction only applies to finite state spaces and requires the exact knowledge of π while, in typical MCMC applications, π is known only up to a normalizing constant (indeed this is one of the strengths of MCMC methods).

Unfortunately, there are many Markov chains that are not comparable relative to the Peskun partial ordering. Just to give some examples, consider two transition matrices with zeros along the main diagonal. Since the row sums have to be equal to one, Peskun will never be able to order them. For a similar reason, Gibbs samplers on a continuous state space are not comparable in the Peskun sense: the Gibbs sampler is a special MH algorithm where the proposal for each coordinate is the conditional target distribution of that coordinate given everything else (the full-conditional distribution). This results in an acceptance probability equal to one and thus the probability of staying put is zero. Likewise, Peskun does not help in finding a good variance for the proposal in random-walk MH algorithms: here the proposal is constructed by adding a random increment to the current position. The choice of the spread of the increment is crucial in designing samplers with good performance. In conclusion, Peskun ordering is nice when it works but it is far from being the natural way to compare MCMC algorithms. This motivates the introduction, in Mira and Geyer (1999), of a weaker ordering (implied by the Peskun ordering), the covariance ordering. The definition and the interest in this ordering are given in the next theorem where each one of the equivalent conditions stated can be taken as defining the covariance ordering. We prefer to take (2) as the defining condition (thus the name of this ordering), where $\mathrm{Cov}_\pi(f, Pf) = E_\pi[f(x_0)f(x_1)]$ is the lag-one autocovariance along a P-chain and f belongs to $L_0^2(\pi)$, the functions of $L^2(\pi)$ with zero mean relative to π.

THEOREM 3. *Given two reversible Markov chains $P, Q \in \mathscr{R}$, the following statements are equivalent:*

(1) *$Q - P$ is a positive semidefinite matrix;*
(2) *$\mathrm{Cov}_\pi(f, Qf) \geq \mathrm{Cov}_\pi(f, Pf)$ for all $f \in L_0^2(\pi)$;*
(3) *P is uniformly more efficient than Q.*

Mira and Geyer (2000) extended the previous theorem from finite to general state spaces. We sketch

the proof for finite state spaces. The equivalence of (1) and (2) follows from the fact that

$$(f, Qf) \geq (f, Pf) \quad \forall f \in L_0^2(\pi)$$

is equivalent to

$$(f, Qf) \geq (f, Pf) \quad \forall f \in L^2(\pi).$$

One implication is obvious. For the other, let f in $L^2(\pi)$. Then $f_0 = f - \mu$ with $f_0 \in L_0^2(\pi)$ and $(f, Pf) = (f_0, Pf_0) + \mu^2$. Similarly, we have $(f, Qf) = (f_0, Qf_0) + \mu^2$ and this gives what we want. The equivalence of (1) and (3) is a consequence of the following representation of the asymptotic variance [Peskun (1973)]: $v(f, P) = 2(f, l_P^{-1}f) - (f, (I+A)f)$, where $l_P^{-1} = (I - P + A)^{-1}$ is the inverse Laplacian and I denotes the identity matrix. Let $A \geq 0$ mean $(f, Af) \geq 0$, $\forall f \in L^2(\pi)$. Then P is at least as efficient as Q if and only if $l_P^{-1} \leq l_Q^{-1}$ which, in turn, is equivalent to $I - P + A \geq I - Q + A$ [by Löwner's theorem; Löwner (1934) and Bendat and Sherman (1955)], which is equivalent to (1).

At first sight, Theorem 3 may seem a little counterintuitive: by imposing a condition on the relative ordering of the lag-one autocovariances, we can order the asymptotic variance of $\hat{\mu}$ which, by definition, equals the doubly infinite sum of autocovariances at all lags. The fact is that the condition imposed on the lag-one autocovariances is stronger than it seems since it is required to hold over a large set of functions [namely $L_0^2(\pi)$].

A necessary and sufficient condition for uniform efficiency in a *nonreversible* setting is given in Mira and Geyer (2000).

2.1.2 *Relative efficiency.* Suppose now we have a specific function f in mind and, without loss of generality, assume it is monotone (a finite state space can always be reordered to make any function monotone).

Let $\mathbf{X}$ be a finite set with N elements. Define the summation matrix T to be an $N \times N$ upper triangular matrix with zeros below the main diagonal and ones elsewhere. Define the southwest submatrix of a matrix M to be the submatrix of M obtained by deleting the first row and the last column. Let Π be the $N \times N$ diagonal matrix with π_i as the ith element on the diagonal. Indicate by f' the transpose of f.

An ordering related to relative efficiency appears in Mira (2000a):

DEFINITION 2.4. Let $P, Q \in \mathscr{S}$. P dominates Q in the southwest ordering, $P \succeq_{\mathrm{SW}} Q$, if all the elements of the southwest submatrix of $T\Pi(P - Q)T$ are nonnegative.

This ordering is interesting because of the following theorem [Mira (2000a)]:

THEOREM 4. *Let $P, Q \in \mathscr{S}$. If $l_P^{-1} \succeq_{\mathrm{SW}} l_Q^{-1}$, then $P \succeq_f Q$ for every f monotone on the state space.*

Because of the Peskun representation of the asymptotic variance (Section 2.1.1), to prove the above result, it is sufficient to show that, for any monotone function f, $f'\Pi(l_P^{-1} - l_Q^{-1})f \leq 0$. Consider the identity $f'\Pi(l_P^{-1} - l_Q^{-1})f = f'T^{-1}T\Pi(l_P^{-1} - l_Q^{-1})TT^{-1}f = f'T^{-1}BT^{-1}f$. The last column and the first row of B contain only zeros as a consequence of the fact that P and Q are stochastic matrices stationary with respect to π. This, together with the definition of southwest ordering, implies that the entries of the matrix B are nonnegative. The monotonicity of f implies that the first $(N-1)$ elements of $T^{-1}f$ and the last $(N-1)$ elements of $f'T^{-1}$ have opposite signs. The result follows since $f'T^{-1}BT^{-1}f$ is then the sum of nonpositive terms.

The southwest ordering can be used to also compare transition matrices that are not reversible, unlike most of the other orderings introduced so far. On the other hand, a limitation of Theorem 4 is that it requires inverse Laplacians: computing the inverse can be quite computationally intensive. Realizing that $(I + P - A)$ provides a first-degree approximation to the inverse Laplacian [Peskun (1973)] allows us to work with transition matrices directly. This realization leads to the following:

DEFINITION 2.5. Let $P, Q \in \mathscr{S}$. P is at least as first degree efficient as Q, for a particular function f, $P \succeq_{1f} Q$, if $f'\Pi(P - Q)f \leq 0$.

The condition $f'\Pi Pf \leq f'\Pi Qf$ is equivalent to requiring that the lag-one covariance of f along a P-chain is less than or equal to the one obtained along a Q-chain, assuming f has zero mean under π. The difference between this and the covariance ordering is that here the ordering on the lag-one covariances has to hold only for the specific function of interest and not for all functions in $L_0^2(\pi)$. Thus we require here a much weaker and easier to verify condition, but, of course, paying less we get less. We get the following theorem that buys us first-degree efficiency (as opposed to uniform efficiency):

THEOREM 5. *If $P \succeq_{\mathrm{SW}} Q$, then $P \succeq_{1f} Q$.*

The proof is similar to that of Theorem 4.

2.2 Orderings Relative to the Speed of Convergence to Stationarity

To justify this criterion as a measure of MCMC performance, consider the fact that π, the distribution of interest, is both the unique stationary and the limiting distribution of the Markov chain we simulate. This means that, as the simulation goes on ($n \to \infty$), the distribution of the chain at time n, $P^n(x, \cdot)$, becomes closer and closer to π in total variation distance: $\|P^n(x, \cdot) - \pi(\cdot)\| \to 0$ regardless of the starting point x. As stated in Theorem 1, the rate of convergence is governed by the spectral radius, which, in finite state spaces, is the second largest eigenvalue (of the transition matrix) in absolute value [Besag and Green (1993) and Roberts (1996), page 48]. This leads to one definition of a natural ordering. Let $\sigma(P)$ be the spectrum of P: the set of λ's such that $\lambda I - P$ is not invertible. The spectrum of a Markov chain transition matrix is a nonempty closed subset of the interval $[-1, +1]$ and always contains an eigenvalue equal to one associated to constant functions. For this reason, we also define $\sigma_1(P)$ which equals $\sigma(P)$ except that the eigenvalue associated with constant functions is removed. It is known that the eigenvalues are real valued only if the Markov chain is reversible. This is why, in the following definition, the class $\mathscr{R}$ is considered.

DEFINITION 2.6. Let $P, Q \in \mathscr{R}$. We say that P dominates Q in the convergence ordering and write $P \succeq_{\mathrm{SC}} Q$, if

$$\sup\{|\lambda| : \lambda \in \sigma_1(P)\} \leq \sup\{|\lambda| : \lambda \in \sigma_1(Q)\},$$

that is, on finite state spaces, if the second largest eigenvalues in absolute value are ordered.

The above definition readily extends to general state spaces. As noted before, for positive definite transition matrices, AV and SC criteria lead to the same optimal Markov chain and we can therefore state the following:

THEOREM 6. *For positive definite transition matrices $P, Q \in \mathscr{R}$, the following implications hold:*

$$P \succeq_P Q \to P \succeq_{\mathrm{SC}} Q \leftrightarrow P \succeq_E Q \leftrightarrow P \succeq_C Q \leftrightarrow \lambda_{iP} \leq \lambda_{iQ} \quad \forall i.$$

The proof follows from Theorem 2 and can be extended to general state spaces.

There are a few Markov chains with positive definite transition matrices used for MCMC purposes. Among them are the independence MH algorithm [Liu (1996a)], the random-scan Gibbs sampler [Liu, Wong and Kong (1995)] and the slice sampler [Mira and Tierney (2001)].

The ordering defined in Definition 2.6 also appears in Frigessi, Di Stefano, Hwang and Sheu (1993) where the random-scan, single-site update, MH algorithm and Gibbs sampler are compared in terms of SC both in general and for simulating the Ising model at different temperatures. Beside Frigessi, Di Stefano, Hwang and Sheu (1993), there are only a few other papers available in the literature that refer to the convergence ordering. This is possibly due to the fact that studying the spectral structure of Markov chains used for practical applications is typically not an easy task. As a consequence, the articles we reference in the rest of the paper only consider special Markov chains such as the Gibbs sampler, the independence MH and the slice sampler, or study particular distributions of interest that are easier to analyze such as Gaussian distributions or the Ising model. We recall the following results by Mira and Tierney (2001):

THEOREM 7. *Given any independence MH sampler, it is always possible to construct a slice sampler that dominates it in the convergence ordering and thus in the uniform efficiency ordering.*

The *slice sampler* is an auxiliary variable construction. Suppose a factorization of the target distribution, possibly up to a constant of proportionality, is available: $\pi(x) \propto q(x)l(x)$, where $l(x)$ is a nonnegative function. The auxiliary variable u is introduced by specifying its conditional distribution given x to be uniform on the interval $(0, l(x))$. A Gibbs sampler, stationary with respect to the joint distribution of x and u, is then constructed on the enlarged state space. This amounts to generating u given x from a uniform distribution on the interval $(0, l(x))$ and x given u from $q(x)$ restricted to the set $A_{u,l} = \{x : l(x) > u\}$. On the negative side, note that the cost of sampling from this latter distribution may be high. On the positive side, note that if l is a bounded function the slice sampler is uniformly ergodic [Mira and Tierney (2001)].

Different factorizations of π give rise to different slice samplers and it is not clear which factorizations result in samplers with good properties. Unfortunately, slice samplers on continuous state spaces cannot be Peskun ordered since Peskun cannot compare continuous state space Gibbs samplers.

The *independence MH* is characterized by having a proposal that does not depend on the current position $X_t = x$ of that chain: $q(\cdot, x) = q(\cdot)$.

Suppose $q(\cdot)$ is the proposal used for an independence MH sampler and let $l(\cdot) = \pi(\cdot)/q(\cdot)$. This provides a possible factorization of π and the resulting slice sampler outperforms the original independence MH in that the MCMC estimates have uniformly smaller asymptotic variance.

Theoretical results on convergence properties of the independence MH sampler compared to various Monte Carlo algorithms can be found in Liu (1996a). In Roberts and Sahu (1997), many different types of Gibbs sampler on Gaussian distributions are compared in terms of SC. The work covers parameterization, blocking and random and deterministic scan. The comparisons are partially extended to the non-Gaussian case via a weak convergence result in Roberts and Sahu (2001). The relationship in terms of SC between the EM algorithm and the Gibbs sampler is investigated in Sahu and Roberts (1999). The authors showed that the rate of convergence of the Gibbs sampler, obtained by Gaussian approximation, is equal to that of the corresponding EM-type algorithm.

The lack of many theoretical results is more than compensated by numerous papers where different MCMC samplers are compared in terms of speed of convergence to stationarity via simulation studies.

Distances different from total variation could be used to measure how far $P^n(x, \cdot)$ is from $\pi(\cdot)$. One example is the χ^2 distance [Diaconis, Holmes and Neal (2000)]. Furthermore, other forms of convergence could be of interest in the evaluation of the performance of a Markov chain for MCMC purposes. We can, for example, consider convergence of moments or pathwise convergence, that is, $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f(x_i) = E_\pi f(x)$ a.s. for all starting points x.

Depending on our aims, different orderings concerning convergence of MCMC samplers can be defined. We do not pursue this further.

3. IMPROVING THE EFFICIENCY OF MCMC SAMPLERS

3.1 Improving with Respect to Absolute Efficiency

What can we learn from Peskun ordering? The intuition behind Theorem 2 is clear: a Markov chain that has smaller probability of remaining in the same position over time (i.e., smaller diagonal elements in the transition matrix) explores the state space more efficiently and thus produces better estimates.

This is the intuition used in the *Metropolized Gibbs sampler* [Liu (1996b)] which improves, in terms of uniform efficiency, the regular Gibbs sampler on discrete state spaces (or when at least one of the state space components is discrete). The Metropolized Gibbs sampler modifies the random-scan Gibbs sampler by excluding, when updating each coordinate x_i, an immediate draw of x_i itself, thus preventing the sampler from retaining the same position over successive points in time. The proposal distribution is not the full-conditional, $\pi(x_i|x_{-i})$, anymore, where $x_{-i} = (x_j, j \neq i)$, but a value y_i, different from x_i, is drawn with probability $\pi(y_i|x_{-i})/[1-\pi(y_i|x_{-i})]$. By doing so, one needs to insert a reject–accept step to correct for the bias induced by not using the full-conditional distribution as the proposal. An acceptance probability that preserves reversibility relative to π is given in Liu (1996b).

The Metropolized Gibbs sampler idea relies heavily on the discreteness of the state space. On general state spaces, the same position is retained over time when a proposal is rejected. Thus, in principle, we could improve the MH and the reversible-jump [Green (1995)] algorithms in the efficiency ordering by decreasing the rejection frequency of proposed moves. We remind the reader that reversible jumps allow the extension of the MH strategy to settings where there is no elementary dominating measure for the target distribution. Examples include variable-dimension problems such as mixture models with an unknown number of components or change point problems with an unknown number of change points. The idea of decreasing the rejection frequency is used in Tierney and Mira (1999) and Green and Mira (2001) where a delaying rejection mechanism is introduced to improve the MH and the reversible-jump algorithms (respectively) on a sweep-by-sweep basis.

The *delaying rejection strategy* works as follows: suppose that, at some point in time, the chain is at position x. Generate a candidate move y from some proposal distribution that may depend on the current position of the chain, $q_1(x, \cdot)$. With the usual MH probability $\alpha_1(x, y)$, accept the move. If the move is rejected, instead of staying put and advancing time, according to standard Metropolis–Hastings (which, we know, leads to a loss in terms of efficiency of the resulting MCMC estimates: Peskun told us!), propose a second-stage candidate move z from a new proposal distribution $q_2(x, y, \cdot)$ and with some probability $\alpha_2(x, z)$ accept the move. If z is also rejected, we could either interrupt the delaying rejection process and remain in the current position (advancing time) or continue with higher stage proposals. The second-stage (or higher) proposal distribution can be different

from the first-stage one and is allowed to depend on the previously rejected candidate. This means that a form of within-steps adaptation of the proposal is allowed: we can learn from our previous "mistakes" (without losing the Markovian property of the resulting sampler). In Tierney and Mira (1999) and Green and Mira (1999), formulas are provided for second-stage acceptance probabilities that preserve reversibility with respect to π separately at each stage of the delaying process. In Mira (2001b), iterative formulas for higher stage acceptance probabilities are given.

The delaying rejection strategy certainly reduces the overall probability of remaining in the current state and thus leads to improved samplers in the Peskun ordering. The price paid for this gain is that it takes longer (in terms of simulation time) for the chain to perform a step (sweep). Thus, whether delaying rejection is useful in practice depends on whether the reduction in variance obtained more than compensates for the additional computational cost. The experimental results reported in Green and Mira (1999), where simulation time is taken into account, indicate that the gain often more than compensates for the price.

Finally, we note that there are various alternatives to improving efficiency of an existing MH sampler by adding one or more delaying rejection steps. Reparameterization, auxiliary variables, data augmentation, simulated tempering, Langevin diffusions or other samplers [Gilks, Roberts and Sahu (1998)] should also be considered as valid if not preferred alternatives, depending on the problem at hand and depending on how poor the performance of the original MH sampler is: a really bad algorithm should be replaced completely by a different method instead of trying to fix it via delaying rejection, unless there is really no option.

3.2 Improving with Respect to Relative Efficiency

In Mira (2000a), we introduced the idea of stationarity-preserving and efficiency-increasing probability mass transfers performed on a transition matrix P having the proper stationary distribution. Transfers of probability mass are motivated by Definition 2.5 of first-degree efficiency and by Theorem 5 and are performed in the following way: given integers $1 \le i < j \le N$ and $1 \le k < l \le N$ and a quantity $h > 0$, increase the entries of the P matrix, $p_{i,l}$ and $p_{j,k}$, by h and $h\pi_i/\pi_j$, respectively, and decrease $p_{i,k}$ and $p_{j,l}$ by h and $h\pi_i/\pi_j$, respectively. The quantity h must be chosen so that, after the mass transfer, in the resulting matrix all the entries are nonnegative and less than one. Notice that the knowledge of π up to a normalizing constant is sufficient to perform these transfers of probability mass.

If P is derived from Q via a stationarity-preserving and efficiency-increasing probability mass transfer (probability mass transfer in short), then $P \succeq_{\mathrm{SW}} Q$. We can thus increase the relative efficiency of a transition matrix (provided the first-degree approximation to the asymptotic variance is good), while preserving its stationary distribution, via a sequence of probability mass transfers. Indeed, until there exist two indices $i < j$ such that $p_{i,k}$ and $p_{j,l} > 0$ for some $k < l$, we can keep moving probability mass around increasing first-degree efficiency.

The extreme transition matrix obtained by applying a sequence of probability mass transfers, that is, a matrix that cannot be further improved in terms of first-degree efficiency, has at most one nonzero element along the main diagonal and presents a path of positive entries connecting the northeast to the southwest corner of the matrix. Which specific pattern is optimal depends on π and f. Of course, once the extreme matrix is obtained, one has to check that it is irreducible. The intuition behind the structure of the extreme matrix is the following: when we perform probability mass transfers, we try to induce first-order negative correlation along the chain path so that the variance of the resulting MCMC estimates will be reduced, possibly even to values smaller than the variance we would get with i.i.d. sampling from π.

As proved in Mira (2001a), if f does not assume the same value on any two points of the state space (so that it is strictly increasing after having reordered the state space to make f monotone), the final result of a sequence of probability mass transfers is *unique* and corresponds to the matrix that minimizes the linear function $f'\Pi(P - A)f$ under the set of linear constraints that ensure that the resulting matrix is stochastic and has the proper stationary distribution (these constraints define a convex region; thus the minimum is unique). We will refer to such a matrix as the first-degree optimal matrix. As proved in Mira, Omtzigt and Roberts (2001), this matrix is always *reversible* with respect to the stationary distribution π even if the starting point is only stationary with respect to π.

Obtaining, from an initial π-stationary matrix, the first-degree optimal matrix via a sequence of probability mass transfers, as presented above, can be quite computationally expensive, especially for large state spaces, to the point that exact computation of the normalizing constant of π and of the mean of the function of interest via brute-force evaluation might become a competitive strategy. In this

regard, we observe the following. Even if we stop the process of performing probability mass transfers at an intermediate stage (before the extreme optimal matrix is reached), we obtain an improvement over the original π-stationary matrix (this intermediate matrix might not be reversible). Second, in Mira, Omtzigt and Roberts (2001), an algorithm to derive the first-degree optimal matrix, optimized from a computational point of view, is presented. The algorithm requires, as its inputs, only the function of interest f and π up to a normalizing constant. The construction of the first-degree optimal matrix becomes thus of actual practical interest (and competitive relative to brute-force evaluation) since it does not even require the knowledge of an initial π-stationary transition matrix.

As pointed out by Billera and Diaconis (2001), there is an insightful geometrical interpretation of how the MH algorithm transforms a generic stochastic matrix K, the proposal distribution in an MH sampler, into a reversible Markov chain. Along the same lines, we can interpret the final result of a sequence of probability mass transfers. In the 2×2 case considered in Figure 2 of Billera and Diaconis (2001), the first-degree optimal matrix is the one at the intersection of the stationarity line of equation $b = a\pi_1/\pi_2$ with the boundary of the unit square containing all the admissible stochastic matrices. The second largest eigenvalue λ_2 of a 2×2 matrix parameterized as in Figure 2 of Billera and Diaconis (2001) (i.e., with $p_{1,2} = a$ and $p_{2,1} = b$) is $\lambda_2 = 1-a-b$. The figure thus clearly shows that the first-degree optimal transition matrix has a smaller second largest eigenvalue than the MH matrix and any other matrix stationary with respect to π. This is an indication of the optimal performance of the corresponding Markov chain in terms of the convergence ordering. But note that this is an artifact of the fact that we are considering a state space with only two degrees of freedom. Furthermore, the first-degree optimal matrix is also the optimal matrix with respect to uniform efficiency in this special case because every function is monotone on the state space considered.

The natural question that occurs is: how does the construction of the first-degree optimal matrix extend to *general state spaces*? Intuition tells us that there might exist a function, ϕ: $\mathbf{X} \to \mathbf{X}$, that associates, to every point in the state space x, the point which has the highest negative correlation with x. Indeed, such a function exists, preserves stationarity and is related to the minimum of the Fréchet class and to the overrelaxation ideas proposed in Neal (1998). The use of the function ϕ for MCMC purposes is further studied in Mira, Omtzigt and Roberts (2001).

4. EXAMPLES

Following Besag (2000), consider the hidden Markov model for a noisy channel outlined in the sequel. Let $x_1, \ldots, x_k$ be the output sequence from a process with $x_i \in \{0, 1, \ldots, s\}$. Suppose that the signal $x = (x_1, \ldots, x_k)$ is unobservable but that each x_i generates an observation y_i that takes values on the same state space. To keep things simple (but everything extends readily to the general setting), we restrict our attention to a binary channel; that is, we let $s = 1$ and assume that only two observations are available ($k = 2$). Let the log-odds of correct to incorrect transmission of x_i to y_i be a known value α. Now suppose that the x_i's form a stationary Markov chain with symmetric transition probability matrix and with log-odds β in favor of $x_{i+1} = x_i$. The object of interest (target distribution) is the posterior probability of a true signal x given data y:

$$\pi(x|y) \propto \exp\left\{ \alpha \sum_{i=1}^{k} \mathbf{1}[x_i = y_i] + \beta \sum_{i=1}^{k-1} \mathbf{1}[x_i = x_{i+1}] \right\}. \tag{4.1}$$

A Markov chain having (4.1) as its unique stationary distribution is the Gibbs sampler. The full conditional distributions needed to implement it are

$$\pi(x_i|x_{-i}, y) \propto \exp\{\alpha \mathbf{1}[\mathbf{x_i} = \mathbf{y_i}] + \beta(\mathbf{1}[\mathbf{x_i} = \mathbf{x_{i-1}}] + \mathbf{1}[\mathbf{x_i} = \mathbf{x_{i+1}}])\},$$

where $x_0 = x_{n+1} = -1$ to accommodate the end points. Thus, interior sites have two neighbors whereas sites 1 and n have a single neighbor. For the $n = 2$ case, we have

$$\pi(x_1|x_2, y) \propto \exp\{\alpha \mathbf{1}[\mathbf{x_1} = \mathbf{y_1}] + \beta \mathbf{1}[\mathbf{x_1} = \mathbf{x_2}]\}$$

and

$$\pi(x_2|x_1, y) \propto \exp\{\alpha \mathbf{1}[\mathbf{x_2} = \mathbf{y_2}] + \beta \mathbf{1}[\mathbf{x_1} = \mathbf{x_2}]\}.$$

Suppose now that, for our specific example ($s = 2$, $n = 2$), the observation is $y = (0, 0)$ and let $f(x) = x_1 + 2x_2$ be the function of interest. The ordering of the state space $\{(0, 0); (1, 0); (0, 1); (1, 1)\}$ makes f monotone. Note that there is a unique ordering of the state space that makes this function strictly monotone. The possibly more interesting function $g(x) = x_1 + x_2$ is monotone relative to more than one ordering of the space and this causes an uninteresting proliferation of cases. Let $c_1 = 1 + \exp(\alpha + \beta)$ and $c_2 = \exp(\alpha) + \exp(\beta)$. Then the Gibbs sampler transition matrix is

$$G = \begin{bmatrix} \frac{\exp[2(\alpha+\beta)]}{c_1^2} & \frac{\exp[\alpha]}{c_1 c_2} & \frac{\exp[\alpha+\beta]}{c_1^2} & \frac{\exp[\beta]}{c_1 c_2} \\ \frac{\exp[2(\alpha+\beta)]}{c_1^2} & \frac{\exp[\alpha]}{c_1 c_2} & \frac{\exp[\alpha+\beta]}{c_1^2} & \frac{\exp[\beta]}{c_1 c_2} \\ \frac{\exp[2\alpha+\beta]}{c_1 c_2} & \frac{\exp[\alpha+\beta]}{c_2^2} & \frac{\exp[\alpha]}{c_1 c_2} & \frac{\exp[2\beta]}{c_2^2} \\ \frac{\exp[2\alpha+\beta]}{c_1 c_2} & \frac{\exp[\alpha+\beta]}{c_2^2} & \frac{\exp[\alpha]}{c_1 c_2} & \frac{\exp[2\beta]}{c_2^2} \end{bmatrix}$$

and its stationary distribution is the vector

$$\pi(x|y) \propto \big[\exp(2\alpha+\beta); \exp(\alpha); \exp(\alpha); \exp(\beta)\big].$$

Let $\alpha = \log(4)$ and $\beta = \log(3)$. The following matrix, obtained by making the second entry on the main diagonal equal to zero, moving that probability mass to entry $(2, 1)$ and correspondingly adjusting the entries $(1, 1)$ and $(1, 2)$ to preserve stationarity, dominates G in the Peskun and in the covariance sense:

$$P = \begin{bmatrix} 0.848 & 0.048 & 0.071 & 0.033 \\ 0.896 & 0 & 0.071 & 0.033 \\ 0.527 & 0.245 & 0.044 & 0.184 \\ 0.527 & 0.245 & 0.044 & 0.184 \end{bmatrix}.$$

Peskun dominance can be verified by inspection (and this is indeed an advantage of Peskun versus covariance ordering). The fact that the eigenvalues of $G - P$ are all zero but one, which is negative, ensures the P dominates G in the covariance sense.

For the specific f of interest and for every other function monotone on the given ordering of the state space, the first-degree optimal matrix is

$$\text{1-opt} = \begin{bmatrix} 0.771 & 0.083 & 0.083 & 0.063 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

whose eigenvalues are $(1, -0.229, 0, 0)$ which, compared to the eigenvalues of the Gibbs matrix, $(1, 0.124, 0, 0)$, show that the first-degree optimal matrix also has a smaller asymptotic variance (not only up to the first-degree approximation).

To find out the gain we get in terms of the asymptotic variance by switching from G to 1-opt, we can compute the ratio of the corresponding asymptotic variances, which turns out to be 2.17; that is, the asymptotic variance is cut by more than half. For a different function, say the mean, which is also monotone on the state space, the ratio is 2.28. In general, different functions lead to different degrees of improvement.

One might wonder how good the first-degree approximation to the asymptotic variance is. Or, equivalently but more practically, if the sampler is run for a finite number of steps (as opposed to infinity), what is the gain we get in terms of variance reduction? To answer this question, we again need to look at the eigenvalue structure of our transition matrices. The smaller the eigenvalues of P are in absolute value, the better this first-degree approximation to the asymptotic variance is. This is related to the fact that the smaller the eigenvalues of P are in absolute value, the faster convergence to stationarity is achieved. In particular, P and G have the same spectral radius, while the first-degree optimal matrix converges to stationarity more slowly.

5. IMPROVING THE SPEED OF CONVERGENCE OF MCMC SAMPLERS

To conclude, we would like to say something about strategies to improve MCMC algorithms in terms of the SC ordering. Intuition suggests that the closer the proposal distribution is to the target, the faster convergence to stationarity is achieved. As an extreme, consider the case in which the proposal is the target itself. If we could sample from the target, we would then run a Monte Carlo simulation (as opposed to an MCMC simulation) and we would have instantaneous convergence to stationarity. A result by Holden (1998) substantiates this intuition and demonstrates the link between the convergence rate to stationarity and the closeness of the proposal to the target. This suggests the idea of allowing the proposal distribution to depend on points previously sampled along the chain trajectory. Doing so speeds up convergence since, as the Markov chain itself converges to π, previous sampled points should help in designing more sensible proposal distributions.

Several strategies exist for altering the proposal distribution based on the chain's history. One approach is to perform a separate pilot run, and from the resulting sample path, gain insight about the distribution of interest and tune the proposal for the successive run accordingly. In this setting, tuning is done once at the beginning of the simulation (though the procedure can be applied iteratively). This strategy has been labeled *pilot adaptation* [Gilks, Roberts and Sahu (1998)] and approaches range from simple to complex [see, e.g., Haario, Saksman and Tamminen (1999)]. However, a requirement of such a strategy is that a single (possibly imperfect) proposal distribution be fixed in order to generate postconvergence samples that may be summarized for inference.

Gilks, Roberts and Sahu (1998) discussed an alternative adaptive strategy, which allows for

updating of the candidate distribution at the algorithm's *regeneration times*. Regeneration times are time points in the algorithm which divide the Markov chain into sections whose sample paths are independent. Thus updating of the candidate distribution can occur repeatedly without disturbing the chain's stationary distribution. However, in cases other than independence MH, the identification of regeneration times, in a way that makes them sufficiently frequent, is hard in high dimensions, and it is unknown, at the beginning of the simulation, how often the algorithm will regenerate. An interesting way of inducing and detecting regeneration times in a Markov chain via enlargment of the original state space with an artificial atom is presented in Brockwell and Kadane (2001).

In other approaches to adaptive algorithms, the Markovian property and/or time homogeneity of the transition kernel is lost and ergodicity of the resulting samplers is proven from first principles. Typically, if this last strategy is used, either the resulting stationary distribution is not the original one but an approximation of it [as in Holden (1998) and Haario, Saksman and Tamminen (1999)], or restrictive conditions on the original target distributions are assumed [as in Chauveau et al. (1999) and Haario et al. (2001)]. For an overview of adaptive methods, the reader can refer to Tierney and Mira (1999).

6. DISCUSSION

A criterion that we have not yet mentioned, but that can be of practical use when choosing a Markov chain for MCMC purposes, is the *ease of implementation* of the underlying stochastic process. Once an algorithm is chosen, we have to write the computer code to run the simulation. This can be intensive work mostly due to the debugging process of the code. Thus a non-expert programmer might choose a chain easier to implement versus one that has slightly better performance. In this regard, we note that the extra computational effort needed to insert a delaying rejection step into an existing MH or reversible jumps code is minimal.

All the orderings presented (except the convergence and the first-degree efficiency ordering) are partial orderings. This means that they do not allow the comparison of all Markov chains that are either reversible or stationary with respect to the specific distribution of interest. We note that, on the space of nonreversible Markov chains, the covariance ordering, which can still be defined, is not a proper ordering anymore since the antisymmetry property fails to hold.

Finally, we stress that the AV of MCMC estimators can be reduced not only by generating a "better" Markov chain on which to compute the ergodic average (as suggested in Sections 3.1 and 3.2), but also by using a "better" estimator than the ergodic average. In this regard, we recall that the Rao–Blackwellization principle used to reduce variance in i.i.d. sampling can also be exploited in MCMC simulations. The idea is to replace $f(x_i)$ in $\hat{\mu}_n$ by a conditional expectation, $E_\pi[f(x_i)|h(x_i)]$, for some function h or to condition on the previous value of the chain thus using $E[f(x_i)|x_{i-1} = x]$ instead [Gelfand and Smith (1990), Liu, Wong and Kong (1995), Casella and Robert (1996) and McKeague and Wefelmeyer (2000)].

Given a Markov chain sampling scheme, does the ergodic average make the best use of the sample? This question, related to Rao–Blackwellization and complementary to the one we have considered in Section 3, is well analyzed in Greenwood and Wefelmeyer (1995, 1999) and Greenwood, McKeague and Wefelmeyer (1996). In these papers, the authors exploit the specific structure of MCMC samplers to construct new estimators that can be combined with the ergodic average to considerably reduce asymptotic variance.

ACKNOWLEDGMENTS

I thank all my co-authors for helpful discussions and for sharing the intuition and the fun. I am extremely grateful to Richard Tweedie for encouraging me to write the present paper and for insightful discussion on the topic treated here: this paper is dedicated to his memory.

REFERENCES

BELLMAN, R. (1972). *Introduction to Matrix Analysis*. McGraw–Hill, New York.

BENDAT, J. and SHERMAN, S. (1955). Monotone and convex operator functions. *Trans. Amer. Math. Soc.* **79** 58–71.

BESAG, J. (2000). Markov chain Monte Carlo for statistical inference. Technical Report 9, Center for Statistics and the Social Sciences. Available at www.csss.washington.edu/papers/wp9.ps.

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

BILLERA, L. J. and DIACONIS, P. (2001). A geometric interpretation of the Metropolis algorithm. *Statist. Sci.* **16** 335–339.

BROCKWELL, A. E. and KADANE, J. B. (2001). Practical regeneration for Markov chain Monte Carlo simulation. Technical Report 757, Dept. Statist., Carnegie Mellon Univ. Available at www.stat.cmu.edu/cmu-stats/tr/.

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

CHAUVEAU, D. and VANDEKERKHOVE, P. (2001). Improving convergence of the Hastings–Metropolis algorithm with a learning proposal. *Scand. J. Statist.* To appear.

DIACONIS, P., HOLMES, S. and NEAL, R. M. (2000). Analysis of a nonreversible Markov chain sampler. *Ann. Appl. Probab.* **10** 726–752.

FRIGESSI, A., DI STEFANO, P., HWANG, A. and SHEU, A. (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.

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

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

GILKS, W. R., ROBERTS, G. O. and SAHU, S. K. (1998). Adaptive Markov chain Monte Carlo through regeneration. *J. Amer. Statist. Assoc.* **93** 1045–1054.

GREEN, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82** 711–732.

GREEN, P. J. and MIRA, A. (2001). Delayed rejection in reversible jump Metropolis–Hastings. *Biometrika* **88** 1035–1053.

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

GREENWOOD, P. E. and WEFELMEYER, W. (1995). Efficiency of empirical estimators for Markov chains. *Ann. Statist.* **23** 132–143.

GREENWOOD, P. E. and WEFELMEYER, W. (1999). Reversible Markov chains and optimality of symmetrized empirical estimators. *Bernoulli* **5** 109–123.

HAARIO, H., SAKSMAN, E. and TAMMINEN, J. (1999). Adaptive proposal distribution for random walk Metropolis algorithm. *Comput. Statist.* **14** 375–395.

HAARIO, H., SAKSMAN, E. and TAMMINEN, J. (2001). An adaptive Metropolis algorithm. *Bernoulli* **7** 223–242.

HOLDEN, L. (1998). Adaptive chains. Technical Report SAND/11/98, Norwegian Computing Center. Available at www.maths.surrey.ac.uk/personal/st/S.Brooks/MCMC/.

LIU, J. S. (1996a). Metropolized independent sampling. *Statist. Comput.* **6** 113–119.

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

LIU, J. S., WONG, W. H. and KONG, A. (1995). Correlation structure and convergence rate of the Gibbs sampler with various scans. *J. Roy. Statist. Soc. Ser. B* **57** 157–169.

LÖWNER, K. (1934). Über monotone Matrixfunktionen. *Math. Z.* **38** 177–216.

MCKEAGUE and WEFELMEYER (2000). Markov chain Monte Carlo and Rao–Blackwellization. *J. Statist. Plann. Inf.* **85** 171–182.

MIRA, A. (2001a). Efficiency increasing and stationarity preserving probability mass transfers for MCMC. *Statist. Probab. Lett.* To appear.

MIRA, A. (2001b). On Metropolis–Hastings algorithms with delayed rejection. *Metron*. To appear.

MIRA, A. and GEYER, C. J. (1999). Ordering Monte Carlo Markov chains. Technical Report 632, School of Statistics, Univ. Minnesota. Available at aim.unipv.it/~anto/order.ps.

MIRA, A. and GEYER, C. J. (2000). On non-reversible Markov chains. *Fields Inst. Comm.* **26** 93–108.

MIRA, A., OMTZIGT, P. and ROBERTS, G. (2001). Stationary preserving and efficiency increasing probability mass transfers made possible. Technical Report 14, Dept. Economics, Univ. Insubria. Available at aim.unipv.it/~anto/possible.ps.

MIRA, A. and TIERNEY, L. (2001). Efficiency and convergence properties of slice samplers. *Scand. J. Statist.* **29** 1035–1053.

NEAL, R. M. (1998). Suppressing random walks in Markov chain Monte Carlo using ordered overrelaxation. In *Learning in Graphical Models* (M. I. Jordan, ed.) 205–225. Kluwer Academic, Dordrecht.

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

ROBERTS, G. O. (1996). Markov chain concepts related to sampling algorithms. In *Markov Chain Monte Carlo in Practice* (W. R. Gilks, S. Richardson and D. J. Spiegelhalter, eds.). Chapman and Hall, New York.

ROBERTS, G. O. and SAHU, S. K. (1997). Updating schemes, covariance structure, blocking and parametrisation for the Gibbs sampler. *J. Roy. Statist. Soc. Ser. B* **59** 291–317.

ROBERTS, G. O. and SAHU, S. K. (2001). Approximate predetermined convergence properties of the Gibbs sampler. *J. Comput. Graph. Statist.* **10** 216–229.

SAHU, S. K. and ROBERTS, G. O. (1999). On convergence of the EM algorithm and the Gibbs sampler. *Statist. Comput.* **9** 55–64.

TIERNEY, L. (1994). Markov chains for exploring posterior distributions. *Ann. Statist.* **22** 1701–1762.

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

TIERNEY, L. and MIRA, A. (1999). Some adaptive Monte Carlo methods for Bayesian inference. *Statist. Med.* **18** 2507–2515.